A Hierarchical Model for Integration of Narrowband Cues in Speech

by

Hau Hwang

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

May 2001

Author . . . .
Department of Electrical Engineering and Computer Science
May 23, 2001

Certified by . . . .
Lawrence K. Saul
VI-A Company Thesis Supervisor

Certified by . . . .
Tommi S. Jaakkola
M.I.T. Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

A Hierarchical Model for Integration of Narrowband Cues in Speech

by

Hau Hwang

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2001, in partial fulfillment of the
requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis we develop techniques to emulate the robust characteristics of the human auditory system. In particular, we simulate the multiband processing of the peripheral auditory system and investigate various probabilistic graphical models for integrating cues derived from narrow frequency bands. We apply our models to the task of detecting speech in noise and detecting the phonetic feature [+/− sonorant]. The primary contribution of this thesis is a hierarchical network with good performance characteristics for both tasks. Ideas explored in this study include multiband processing, probabilistic graphical networks, and learning from examples.

Thesis Supervisor: Lawrence K. Saul
Title: Principal Technical Staff Member, AT&T Labs-Research

Thesis Supervisor: Tommi S. Jaakkola
Title: Assistant Professor

Acknowledgments

I would especially like to thank my VI-A thesis advisor Lawrence Saul for enthusiastically helping me with this thesis. He introduced me to many of the ideas outlined in this work and helped me extend the concepts discussed in some of his earlier studies [21] into a large part of this thesis project. His concern, patience, and caring are greatly appreciated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is arguably the most important mode of human communication. It is a highly efficient and convenient means for sharing ideas. To a large extent, speech made possible the development of human civilization because it gave mankind the ability to proliferate and exchange knowledge easily.

One of the many reasons why speech is such a desirable medium for communication is its resistance to variation. The ability to understand speech is largely unaffected by differences across speakers, accents, and dialects. When we speak to a stranger, chances are we will have no trouble communicating as long as a common language is used. Speech is also quite resistant to corrupting influences like interference, distortion, and noise. We generally have no problems communicating even in a loud sports stadium.

Scientists are immensely interested in speech because of its interesting characteristics and importance to human communication. Many researchers explore models for speech production and auditory processing to better understand the human speech generation and recognition processes. Much work has also been invested in uncovering the perceptual aspects of hearing. All these speech studies have tremendous impact in our everyday lives through technologies like stereo systems, telephones, and various multimedia formats and devices. In the last 50 years, scientists have also begun to focus on the challenge of developing automatic speech recognition (ASR), one of the most ambitious and exciting technologies to come from speech research.

1.1 Automatic Speech Recognition

ASR is concerned with developing machinery to match the capabilities of human listeners. By working on ASR problems, scientists have gained insights into the workings of speech production and auditory processing. Researchers also hope to realize the potential of ASR to revolutionize user interfaces by making devices like computers much easier to interact with. Instead of using keyboards and mice, people could communicate naturally with machines directly through speech.

Traditionally, researchers have taken two separate approaches towards the design of speech recognition machinery. One method, commonly known as the knowledge-based approach, stresses the use of expert knowledge to guide the design of ASR systems. ARPA's speech understanding project is an early example of a knowledge-based approach that relied on expert phonetic, lexical, and syntactic knowledge [12]. These techniques for the most part have been unsuccessful due to the inability to quantify and exploit expert knowledge effectively. For example, we do not know how to reliably extract phonemes, the smallest units of speech, even though we have expert knowledge about cues that indicate various phonetic distinctions. Our limited knowledge of auditory processing prevents us from building recognizers based purely on expert knowledge.

A second approach to designing ASR machinery is to view speech as a signal that can be analyzed using statistical signal analysis techniques. Recognizers that follow this approach often use methods involving time-frequency analysis techniques [19] and pattern-matching and language modeling [10]. Such methods are appealing because of their ability to model uncertainties in our understanding of the speech process. Their success is evidenced by modern commercial recognizers which use sophisticated statistical techniques for constrained speech tasks like digit recognition and simple dictation. Usually, these systems employ hidden Markov models to process a mel-frequency cepstra representation of the acoustic signal [18, 20].

Although modern statistical ASR systems perform reasonably when used in ideal quiet conditions, their performance deteriorates rapidly in the presence of noise. A recent paper [13] mentions a study on speaker-independent digit recognition which found the error rate for a carefully tuned ASR system to be less than 2% in quiet conditions but over 40% in a 0 dB speech-to-noise ratio (SNR) environment. For comparison, human listeners perform speaker-independent digit recognition with less than 1% error under both quiet and 0 dB SNR conditions. Unlike today's ASR machinery, the human auditory system is able to exploit the characteristics of speech that make it highly resistant to variation and degradation. The sensitivity of modern ASR systems to noise has motivated great interest in signal representations and recognition techniques that are more resistant to variabilities in speech and the environment.

Achieving good machine speech recognition under poor listening conditions is much more than just an interesting academic problem in ASR research. Because it is often impossible to control the acoustic environment, it is important for ASR systems to function effectively in quiet as well as noisy environments. Many believe that ASR systems will not find widespread use until techniques are developed to improve their robustness to noise.

1.2 Human Speech Recognition

We believe ASR systems stand to gain by emulating the auditory processing that enables human listeners to recognize corrupted speech. Many researchers agree that the success of modern statistical recognizers is limited and that future improvements in ASR will require the greater use of expert speech knowledge [23]. There is great potential in emulating auditory processing because human listeners are much more resistant to noise than modern ASR machinery. Emulating auditory processing is also exciting because it provides the opportunity to develop new models and algorithms for speech processing.

Auditory physiology and psychoacoustic studies provide valuable insights into human auditory processing. Early studies by Georg von Békésy [22] demonstrated that the human cochlea analyzes acoustic signals by frequency. His experiments showed that the place of maximum excitation along the cochlea's basilar membrane varies according to the frequency of the acoustic stimulus. This observation led to the "place theory" of hearing which postulates that excitation along the basilar membrane encodes the frequency content of the acoustic signal. Figure 1-1 illustrates this idea. The figure shows a vibration envelope along the basilar membrane for an arbitrary sinusoid. Along the horizontal axis is the tonotopic axis which corresponds auditory nerve activity with the physical place of excitation along the basilar membrane. This axis is distributed logarithmically. Higher frequency sinusoids cause greater vibrations towards the stapes (one of the three small bones in the ear) while lower frequencies generate excitations further away from it. This observation suggests that the cochlea separates incoming signals into different frequencies for processing.

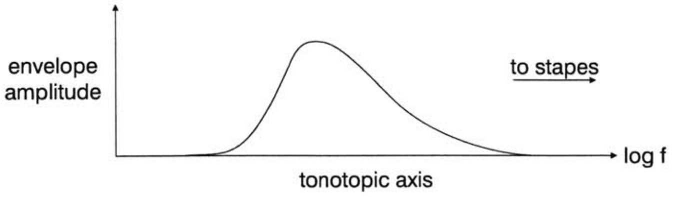

Figure 1-1: Tonotopic axis.

Psychoacoustic experiments support the conjecture that the ear resolves sounds into different frequencies. Harvey Fletcher [7] established the important concept of the critical band, or auditory filtering, in masking experiments where he showed that the detection of a pure tone in noise depends only on the noise within a certain bandwidth of the tone. Figure 1-2 demonstrates this idea. The figure on the left depicts an arbitrary pure tone masked by bandlimited noise centered at the frequency of the

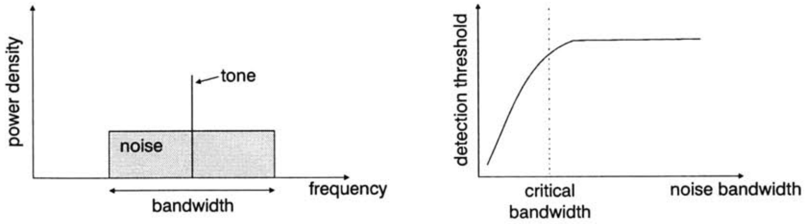

Figure 1-2: Critical band concept.

tone. By keeping the power density of the noise constant, the detection threshold of the signal can be measured as a function of the noise bandwidth as shown on the right. Initially, as the noise bandwidth is increased, the detection threshold increases. Once the noise bandwidth exceeds the critical bandwidth, the detection threshold levels off. Apparently, noise outside the critical bandwidth has no effect on the detection of the tone. Critical bands are generally narrower for lower frequencies and wider at higher ones [17]. This effect suggests that the auditory system behaves like a bank of filters decomposing acoustic signals into different frequency bands for separate processing.

Harvey Fletcher's articulation experiments give further insight into the human speech recognition process [1, 7]. Fletcher determined the error of articulation, or recognition ability in the absence of context, to be modeled by the product of the individual articulation errors in different frequency bands. In other words, good articulation can be achieved if the SNR within just a single frequency band is satisfactory. From this we can hypothesize that the auditory system processes information from each frequency band independently of the others. Degraded speech is recognized by integrating evidence from the cleaner portions of the spectrum and ignoring bands corrupted by noise. Fletcher's articulation experiments point to a layered model for human speech recognition [1]. In this model, processing begins with critical band filtering of the acoustic waveform. Phonetic features are extracted locally across the spectrum from bands with high SNR and integrated in the next layer to realize discrete phonemes. These phonemes are then used by additional layers in the recognition chain to piece together syllables and words.

The work of Miller and Nicely [15] also provides important clues into auditory processing. In this study, human subjects were asked to identify 16 consonants from nonsense syllables in noisy bandlimited speech. Miller and Nicely found that their human subjects were able to recognize basic phonetic distinctions even though entire phonemes could not be identified. Table 1.2 reproduces one of the confusion matrices from their experiments. The numbers in the table indicate the frequency of identifying the consonants in the first column as consonants listed in the first row on the top. As seen from this example, there are five groups of consonants that are difficult to confuse

	p	t	k	f	θ	s	š	b	d	g	v	ð	z	ž	m	n
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
š	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ž								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

Table 1.1: Confusion matrix for SNR = -6 dB and frequency response 200-6500 Hz (Table III from Miller and Nicely, 1955 [15])

with one another even under poor listening conditions. Miller and Nicely grouped these five classes based on articulatory properties, or phonetic features. At higher SNR levels, the separation among the consonants based on these phonetic features is even more apparent. The ability to make these distinctions indicates that certain phonetic features are detected quite early by the auditory system prior to recognizing entire phonemes, syllables, and words. This result along with Fletcher's human speech recognition model suggests that the robust detection of certain phonetic distinctions is fundamental to robust human speech recognition.

1.3 Probabilistic Graphical Models

The facts and properties of the human auditory system that we have discussed are all potentially useful for emulating auditory processing. A natural question to ask is how to incorporate this knowledge in speech systems. We certainly cannot build a system based purely on this knowledge since our understanding of auditory processing is incomplete. Uncertainties about auditory processing suggest a statistical approach to emulating the processing in the auditory system.

We believe probabilistic graphical models [11] are a natural choice for combining expert knowledge about human speech processing into statistical models. Probabilistic graphical models, also known as Bayesian networks, provide a formal way for making probabilistic inferences based on the prior knowledge of the problem. Structurally, these models are graphs composed of nodes and edges. Nodes in the model represent random variables while edges assert dependencies among the nodes. Since the laws of probability govern inferences in these models, the output is easy to understand and interpret.

To specify a probabilistic graphical model, we need to indicate its graphical structure along with the probability densities for the nodes. For our purposes, we use the structure of the probabilistic graphical model to formalize hypotheses of auditory processing. As an example, consider the simple graphical model in Figure 1-3. In this example, random variables $X_1, \ldots, X_Q$ are combined using an OR gate. The basic processing in this network is to set random variable Z to true if any of the random variables X_q are true. We use this type of structure in a union model where the X_q indicate the presence of speech cues from different parts of the spectrum. Once we select a network structure like the one in this example, we train the model to learn the probability densities for the nodes in the network. The trained network can then be used to make inferences on unseen data.

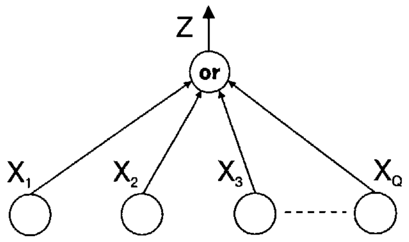

Figure 1-3: Probabilistic graphical network example.

1.4 Goals

The work in this thesis combines two ideas from the ASR and machine-learning communities. Among ASR researchers, there is a growing consensus that recognition systems require greater use of expert speech knowledge to improve robustness. In the machine-learning community, scientists are becoming more appreciative of the ability of probabilistic graphical models to incorporate expert knowledge into statistical models. These two realizations motivate our design of automatic methods for the robust detection of certain phonetic distinctions. We look to include expert knowledge about the auditory system to construct speech systems that better match human performance. To do this, we use the structure of probabilistic graphical models to incorporate knowledge about multiband auditory processing.

We focus on the problem of robustly detecting certain phonetic distinctions because we believe it to be essential to robust speech recognition. Our automatic methods are designed for two basic tasks: detection of speech in noise and detection of the phonetic feature [+/−sonorant]. We first explore models for detecting speech since robust recognition would not be possible without the ability to reliably differentiate speech from noise. We then apply these models to the robust detection of the phonetic feature [+/−sonorant] which distinguishes vowels, semivowels, and nasals (sonorants) from stops, fricatives, and affricates (obstruents). Sonorants and obstruents form the first major division of the phonemes.

The type of noise we are concerned with in this study is additive and bandlimited Gaussian noise. There are however many other types of noises and environmental effects that may corrupt speech. Interference like background music and multiple speakers can confound the recognition process. Plus, distortions from reverberations and reflections may also destroy intelligibility. We do not deal with these cases. Instead we focus on bandlimited Gaussian noise because it is a good approximation to a wide variety of corrupting influences. When a signal is decomposed into narrow frequency bands, the noise within each band is usually well modeled by bandlimited Gaussian noise.

Unlike other studies that explore multiband models for robust ASR [4, 16], we do not build a connected word recognizer. One of the major challenges when working with multiband models for speech systems is determining how best to unite the information from the different frequency bands to make a decision. We investigate the much simpler task of detecting certain phonetic features to focus on this fundamental problem of integrating meaningful speech cues distributed in frequency. We study three statistical models, the union, weighted union, and hierarchical models, for combining narrowband cues from across the spectrum. Issues regarding the architecture and training of these networks are discussed in detail.

1.5 Outline

This thesis is organized as follows. Chapter 2 develops various models for the speech detection task. We also present experimental evidence comparing our models to the performance of human listeners. In Chapter 3, we describe modifying our speech detection models to detect the phonetic feature [+/−sonorant]. After evaluating these models, we discuss overall conclusions and areas for future work in Chapter 4.

Chapter 2

Speech Detection

Perhaps the most basic problem in speech processing is simply to detect the presence of speech in noise. This is the problem addressed in this chapter.

We begin by describing the speech detection problem as presented in an experiment at AT&T Shannon Laboratory designed to determine the speech detection abilities of human listeners. This experiment shows the remarkable ability of human listeners to identify amplitude modulations from speech in noisy bandlimited signals. The data from this study serves as an upper bound on the performance we expect to achieve by emulating the auditory system.

We then develop the union, weighted union, and hierarchical models to match the human speech detection results from the AT&T experiment. These models use a front-end that decomposes inputs into critical bands and a probabilistic graphical network back-end to integrate speech cues from across the spectrum. For each model, we detail its structure and the EM learning algorithm used for training. Evaluations show that these models come close to matching human speech detection behavior.

2.1 Human Speech Detection

The robust speech recognition ability of human listeners is quite impressive. Despite noise and other adverse environmental effects, the auditory system is often able to recognize speech without difficulty. Measuring human speech detection performance is one way of quantifying the resistance of the auditory system to noise.

Miriam Furst and Jont Allen at AT&T Shannon Laboratory measured the speech detection capabilities of 25 human subjects with normal hearing. Recordings of monosyllabic words like "tin" and "pill" articulated by 4 male and 3 female speakers were played to the human subjects. Three parameters were varied during the experiments. The speech was filtered into various bandwidths around two different center frequencies and corrupted with additive bandlimited Gaussian noise. If we let f_{lower} and f_{upper} be the lower and upper passband cutoff frequencies, we may define the first two parameters, the bandwidth (BW) and center frequency (CF), as follows.

$$CF = \sqrt{f_{\text{lower}} f_{\text{upper}}} \qquad BW = \log_2\left(\frac{f_{\text{upper}}}{f_{\text{lower}}}\right) \tag{2.1}$$

Bandlimited signals were used because one of the working hypotheses for the experiment was that the auditory system extracts information from narrow frequency bands. As discussed earlier, psychoacoustic and physiological studies support this assumption. The third parameter was the speech-to-noise ratio (SNR). For this experiment, the SNR was defined as the ratio of the maximum sample variance of the bandlimited speech to the maximum sample variance of the noise calculated over 20 ms sections of the signals.

The experiment proceeded as follows. An arbitrary monosyllabic word was filtered to a particular bandwidth around a selected center frequency. Gaussian noise with the same bandwidth and center frequency was also generated. The listener was then presented with either the bandlimited signal degraded by the additive noise or just the bandlimited noise and asked to determine whether the signal contained speech. By varying the SNR through repeated trials, the speech detection ability of the subject was determined. What the experiment basically measured was the ability of human listeners to detect amplitude modulations from bandlimited speech embedded in additive noise. Figure 2-1 gives an example of one of the speech waveforms used. The signal shown on the left is the word "Bill" spoken by a male speaker. In the middle is the signal after being filtered to 40 Hz bandwidth around 1000 Hz center frequency.

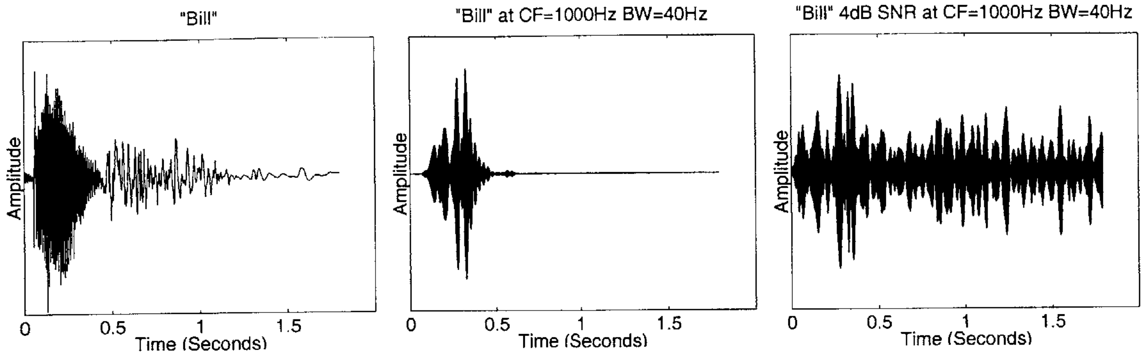

Figure 2-1: The word "Bill" filtered and corrupted by noise.

On the right is the filtered signal corrupted with additive noise to 4 dB SNR. We see that even at 4 dB SNR, the speech signal is evident as larger spikes embedded in the noise.

Given a fixed bandwidth and center frequency, the listener's ability to identify noisy speech can be modeled as a function of the SNR. This function traces out a psychometric curve as shown in Figure 2-2. The vertical axis corresponds to the percentage of correct responses given by the human listener. SNR is measured along the abscissa. With an equal presentation of bandlimited noise and bandlimited noise plus speech, the worst possible performance is 50% correct. This result may be obtained

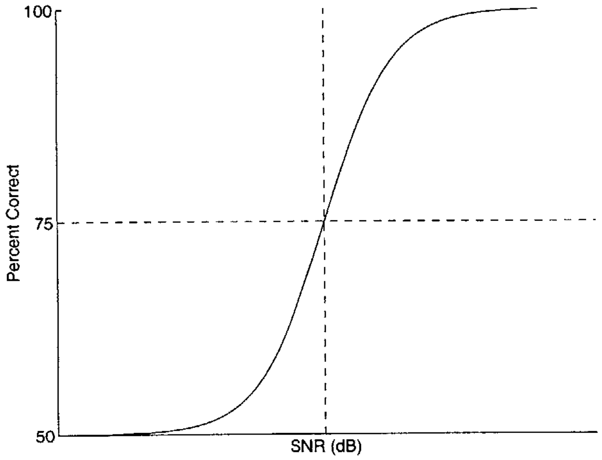

Figure 2-2: Psychometric curve.

by either perpetually declaring the signal to contain speech or always stating that the signal is noise. Perfect performance corresponds to always correctly identifying signals as either speech plus noise or just noise. The 75% correct level is the detection threshold. For a particular bandwidth and center frequency, the detection threshold specifies the SNR at which the human listener is just able to distinguish signals containing speech from those that contain only noise. Although the psychometric curve in Figure 2-2 is highly idealized, it illustrates some important properties. First, we see that at low SNR levels, the percentage of correct responses drops to 50%. This is reasonable since under poor listening conditions, human listeners will have difficulty concluding that the signal is anything other than noise. At high SNR levels, the curve tends towards 100%. This is also expected since it is much easier to detect amplitude modulations from speech as the SNR improves.

The aggregate data from the experiment is illustrated in Figure 2-3. It depicts detection thresholds for various bandwidths around the center frequencies 650 Hz and 1000 Hz. For example, the figure shows that the detection threshold for a speech signal with center frequency 1000 Hz and bandwidth 0.057 octaves (approximately 40 Hz) is roughly 4 to 5 dB SNR. We also see that a signal with center frequency 650 Hz and bandwidth 0.138 octaves (approximately 62 Hz) has a detection threshold of 3 dB SNR. Each data point resulted from a series of experiments with a fixed bandwidth and center frequency where the SNR of the bandlimited speech was altered to determine the detection threshold. Using different words and different speakers did not affect these results.

Two important results stand out from Figure 2-3. The first is that detection threshold improves with increasing speech bandwidth. This is not surprising since more information is conveyed when the speech bandwidth is increased. The second observation is that the detection threshold varies depending on the speech center frequency. We will use the data collected from this experiment as a benchmark for the automatic speech detection methods we develop in this chapter.

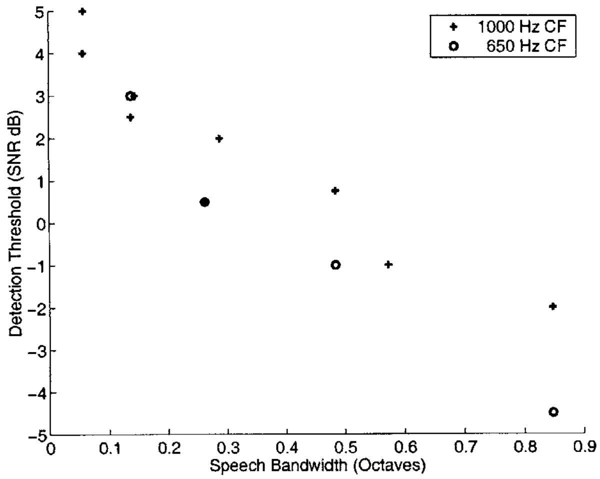

Figure 2-3: Human speech detection performance.

2.2 Speech Detection in a Narrowband

We first try to emulate the human speech detection capabilities for a single narrowband signal with bandwidth 40 Hz around 1000 Hz center frequency. Working on this simpler problem allows us to motivate a multiband approach to the speech detection problem. Figure 2-4 illustrates our narrowband model for speech detection.

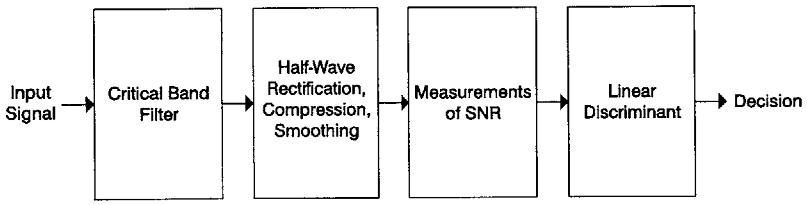

Figure 2-4: Schematic of narrowband model.

The front-end signal processing for the model begins by filtering the acoustic signal into the desired 40 Hz band centered around 1000 Hz. This filter can be thought of as simulating one of the critical bands in the human cochlea. After filtering, nonlinearities are applied. These nonlinearities include half-wave rectification to imitate the unidirectional response of hair cells along the basilar membrane and cube-rooting to simulate the compressive behavior of auditory filters. We then smooth the signal envelope to emphasize peaks and blur minor variations. The samples of the smoothed signal are obtained from the energy of the original waveform within a sliding window. Figure 2-5 illustrates our nonlinearities on a noisy sine wave. From left to right, we see the signal successively processed by half-wave rectification, cube-root compression, and smoothing. The end result of this processing is to remove the minor fluctuations in the signal and emphasize the peaks.

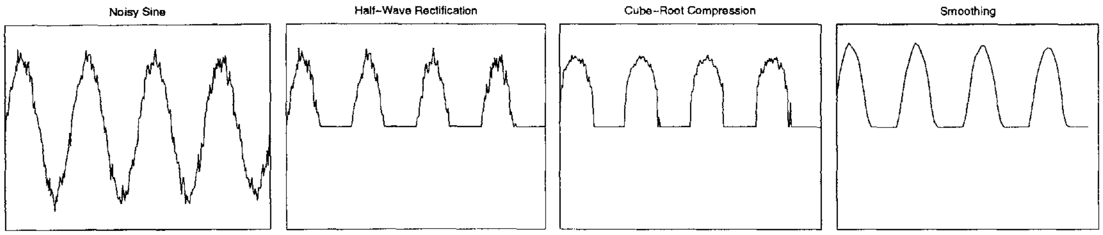

Figure 2-5: Noisy sine wave processed by half-wave rectification, cube-root compression, and smoothing.

After the nonlinearities, three types of SNR measurements are computed to detect amplitude modulations in the signal. These are the crest factor, the coefficient of variation, and the logarithm of the ratio of the nth to $(100 - n)$th percentile values. The crest factor is defined as $\max|x(t)|/\text{rms}[x(t)]$ and is generally used to measure whether a signal has large peaks and deep valleys. We expect signals with higher SNR to exhibit larger crest factors. For example, the additive noise in Figure 2-1 (the difference between the 4 dB SNR signal and the clean bandlimited "Bill" signal) would have a smaller crest factor than the 4 dB SNR signal because its maximum amplitude is smaller. The coefficient of variation is the ratio of the standard deviation of the signal to the mean and is used to measure the dispersion of data about the mean. Again, we expect a signal with high SNR to be positively correlated with the coefficient of variation. Referring to Figure 2-1, the 4 dB SNR signal would have a greater coefficient of variation than the additive noise since the peaks from the speech signal give it greater variance. The logarithm of the ratio of the nth to $(100 - n)$th percentile values also gives an estimate of the SNR for the signal. Values of n used include 95, 90, 80, and 75. The 4 dB signal in Figure 2-1 also has a larger value for this measure than the noise signal because of the large peaks from the speech waveform.

These measurements are then fed to a linear discriminant for processing. We append a value of unity to the measurement vector to include a bias term. The output of the discriminant is the scalar $\sigma(\mathbf{w} \cdot \mathbf{m})$ where $\sigma(x) = 1/[1+\exp(-x)]$ is the sigmoid activation, $\mathbf{w}$ is a column vector of weights, and $\mathbf{m}$ is the column vector of measurements for the input signal. Note that the output ranges between 0 and 1. We interpret this result as our estimation of how likely the input signal contains speech. A value of unity means the system is certain that the input contains speech whereas a value of zero indicates that the input is just noise.

To train the weights for the linear discriminant, we maximize the log-likelihood function. Let binary random variable Y indicate the presence of speech in the input signal. Then $\Pr[\mathrm{Y} = 1|\mathbf{m}] = \sigma(\mathbf{w} \cdot \mathbf{m})$. If we have a collection of measurements $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_N\}$ and the corresponding target labels $\mathcal{Y} = \{\mathrm{Y}_1, \ldots, \mathrm{Y}_N\}$ specifying the desired output of the linear discriminant associated with the measurements, the logarithm of the likelihood function is then:

$$l(\mathcal{Y}|\mathbf{w}, \mathcal{M}) = \sum_i \mathrm{Y}_i \ln(\sigma(\mathbf{w} \cdot \mathbf{m}_i)) + (1 - \mathrm{Y}_i) \ln(1 - \sigma(\mathbf{w} \cdot \mathbf{m}_i)) \tag{2.2}$$

The optimal weight vector for the linear discriminant is the vector $\mathbf{w}$ that maximizes $l(\mathcal{Y}|\mathbf{w}, \mathcal{M})$. We solve for this optimal weight vector using Newton's method [2] by computing the following quantities:

$$\frac{d}{d\mathbf{w}} l(\mathcal{Y}|\mathbf{w}, \mathcal{M}) = \sum_i^N (\mathrm{Y}_i - \sigma(\mathbf{w} \cdot \mathbf{m}_i))\mathbf{m}_i \tag{2.3}$$

$$\frac{d^2}{d\mathbf{w}^2} l(\mathcal{Y}|\mathbf{w}, \mathcal{M}) = \sum_i^N \sigma(\mathbf{w} \cdot \mathbf{m}_i)[\sigma(\mathbf{w} \cdot \mathbf{m}_i) - 1]\mathbf{m}_i \mathbf{m}_i^{\mathrm{T}} \tag{2.4}$$

Then, for the jth iteration of Newton's method, we obtain a new weight vector estimate $\mathbf{w}^{(j)}$ as follows:

$$\mathbf{w}^{(j)} = \mathbf{w}^{(j-1)} - \left[\left(\frac{d^2}{d\mathbf{w}^2} l(\mathcal{Y}|\mathbf{w}^{(j-1)}, \mathcal{M})\right)\right]^{-1} \left(\frac{d}{d\mathbf{w}} l(\mathcal{Y}|\mathbf{w}^{(j-1)}, \mathcal{M})\right) \qquad (2.5)$$

The dimensionality of our measurement vector is small enough such that the matrix inversions required for Newton's method are quite manageable. Newton's method gives fast quadratic convergence to the weight vector that maximizes the log-likelihood function.

The data we used to train the linear discriminant consists of a subset of the words used in the AT&T speech detection experiments. These words are shown in Table 2.2. We computed SNR measurements for each of these words after they were corrupted by various amounts of additive noise and processed by the critical band filter and the nonlinearities. Specifically, we created speech signals at 30 dB, 5 dB, and 3 dB SNR. We also included a Gaussian noise signal. The 30 dB SNR signal and the noise signal serve as prototypical examples of clean speech and noise. For clean speech, we expect the linear discriminant to output unity and for the noise we expect an output of zero. Accordingly, these were the target labels we associated with the measurements for these two signals. The 3 dB and 5 dB SNR signals straddle the detection threshold. By setting the target label for the 3 dB SNR signal to zero and the 5 dB SNR signal to unity we induce the linear discriminant to learn a decision boundary that approximately separates signals with SNR greater than 4 dB from those with lesser SNR. Training a linear discriminant in this fashion causes it to behave similarly to human listeners. Given an equal presentation of noise and speech plus noise at various SNR levels, the performance of the linear discriminant traces out a psychometric curve like the one in Figure 2-2. A total of 14 different words were used to generate 560 examples for training.

We evaluated the capabilities of the narrowband model by testing whether it matches the 4 to 5 dB SNR detection performance of human listeners for 40 Hz bandwidth signals centered at 1000 Hz. The speech we used for this evaluation came from the AT&T speech detection experiment. As shown in Table 2.2, we used several

	Training	Testing
Words	tin, berg, pill, knock, bill jog, tote, kame, pan, perch bah, lace, tonne, do	din, jock, dope, piss, half, mood, dad
Examples	560	1400
(CF, BW) in Hz	(1000, 40)	(1000, 40)

Table 2.1: Training and testing data for narrowband model.

different words for testing to ensure that our models are able to reliably distinguish noise from noisy speech across various words. A total of 1400 examples were used for this test. Half the signals were noise while the other half contained speech at SNR levels of 14, 10, 6, 4, 2, 0, -2, -4, -8, and -12 dB.

Our evaluation revealed the psychometric curve in Figure 2-6 for the narrowband model. Human performance is also plotted for reference. From the figure, we see that

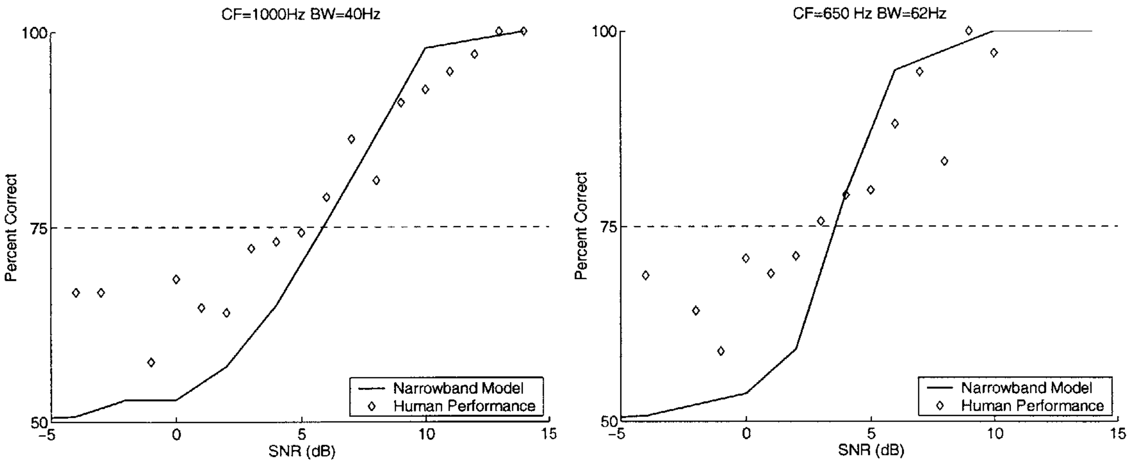

Figure 2-6: Narrowband model and human performance for 40 Hz bandwidth 1000 Hz center frequency and 62 Hz bandwidth 650 Hz center frequency signals.

the narrowband model has an approximate detection threshold of about 5 dB for 1000 Hz center frequency signals with 40 Hz bandwidth. A similar experiment measured the detection threshold of a narrowband model trained to detect speech in 650 Hz center frequency signals with 62 Hz bandwidth. This model achieved a detection threshold of about 3 dB SNR. Both these results match the detection thresholds for human listeners. From this evidence, it appears that this model is a good one for detecting speech in narrowband signals.

We consider whether this narrowband model is sufficient for achieving good speech detection thresholds for wideband signals as well. To test this, we used our 40 Hz bandwidth 1000 Hz center frequency narrowband model to detect speech in wideband signals. The wideband signals we tested were centered at 1000 Hz with bandwidths of 100, 160, 220, and 280 Hz. For each bandwidth, we tested using 1400 examples with an equal presentation of signals with and without speech. The result of this evaluation is shown in Figure 2-7. The figure illustrates that as the bandwidth is increased, the detection threshold actually increases for the narrowband model trained on the 40 Hz bandwidth 1000 Hz center frequency task. This behavior is in direct contrast to the behavior of human listeners since Figure 2-3 shows that the detection threshold for human listeners drops as the bandwidth increases. We conclude that using measurements from a single narrowband signal is insufficient to emulate the speech detection capabilities of humans.

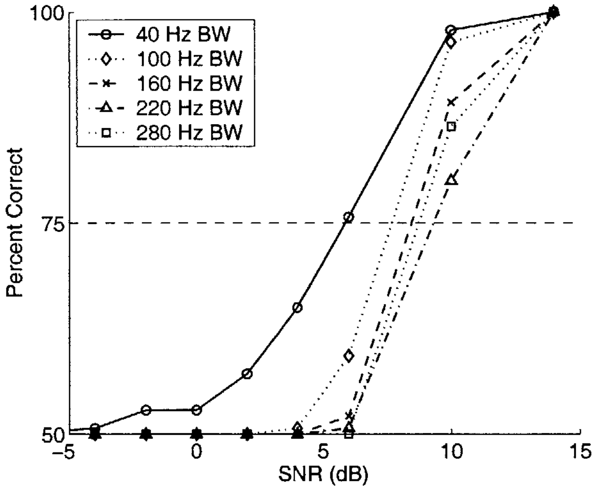

Figure 2-7: Narrowband model performance for various bandwidths around 1000 Hz center frequency.

2.3 Union Model

To better match the speech detection capabilities of human listeners, we will need to analyze more than just a single critical band for the presence of speech. We develop the union model for the purpose of processing and integrating information from several critical bands to detect speech in noise. The union model considered in this section closely mirrors the work by Saul et al [21]. A similar union model was proposed by Ming and Smith [16].

To examine more than just one critical band, we consider a model that is composed of several narrowband models that work in parallel to analyze different narrowband components of the spectrum. A statistical network then integrates the output from each narrowband model to reach a global decision. This in essence is the union model. A schematic of the model is shown in Figure 2-8. The first component of the model is a cochlear filterbank which filters the signal into 32 overlapping bands to simulate the critical band decomposition of signals in the human cochlea. We use critical band filters with center frequencies that are evenly spaced on a logarithmic scale from 300 Hz to 1300 Hz with a constant bandwidth of 0.15 octaves. This range of frequencies adequately covers the spectrum of the signals used in the AT&T speech detection experiments. Next, we half-wave rectify and cube-root each of the 32 narrowband signals. This is followed by computation of SNR measurements for each channel. These measurements are then used by a probabilistic graphical model, the union network, to determine whether the input signal contains speech.

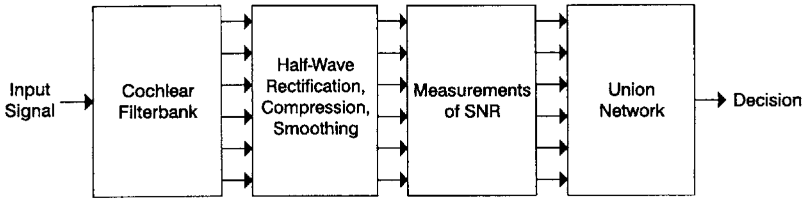

Figure 2-8: Schematic of union model.

Like the narrowband model, the union network first processes the SNR measurements from each subband with a linear discriminant. Each linear discriminant gives a measure of how likely it is for each of the subbands to contain speech. We integrate these individual decisions by taking their union to decide whether the input signal contains speech. In other words, if we treat the outputs of the linear discriminants as binary indications of whether speech is present in a particular subband, then what the network does is to combine these decisions using an OR gate. If any one of the subbands has measurements indicating the presence of high SNR, the union model declares the input to contain speech.

The motivation for the union strategy comes from Fletcher's articulation experiments. Fletcher found that the overall articulation error can be modeled by the product of the individual articulation errors from different frequency bands. As pointed out in Chapter 1, this type of processing suggests the auditory system analyzes each frequency band independently and integrates information from the cleaner bands while ignoring the corrupted ones. Our model duplicates this product-of-errors processing. Since the union network detects speech if any of the bands has high SNR, the error is just the probability that none of the bands detect speech. This is simply the product of the probabilities of not detecting speech in each band. By modeling this processing, we hope to emulate the auditory system's ability to detect degraded speech.

The statistical model for the union network is shown in Figure 2-9. Nodes within the network represent binary random variables while the edges assert dependencies among them. The vector $\mathbf{m}_q$ represents SNR measurements from the qth band.

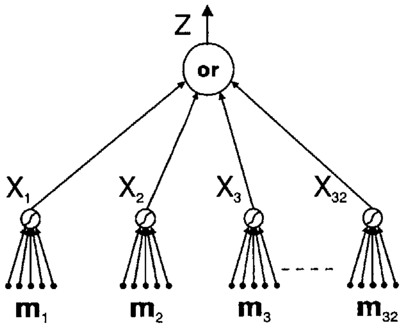

Figure 2-9: Union network.

This measurement is processed by a linear discriminant whose output is the binary random variable X_q which indicates whether the subband has high SNR. A value of unity means the linear discriminant believes the subband has high SNR whereas a value of zero implies that only noise is present. Binary random variable Z represents the overall decision of the network. A value of unity means that speech is present in the input signal while a value of zero indicates otherwise. The probabilities of X_q and Z are defined below where Q represents the number of subbands in the network.

$$\Pr[\mathrm{X}_q = \alpha | \mathbf{m}_q] = \begin{cases} \sigma(\mathbf{w}_q \cdot \mathbf{m}_q) & \text{if} \quad \alpha = 1 \\ 1 - \sigma(\mathbf{w}_q \cdot \mathbf{m}_q) & \text{if} \quad \alpha = 0 \end{cases} \tag{2.6}$$

$$\Pr[\mathrm{Z} = 1 | \mathrm{X}_1, \ldots, \mathrm{X}_Q] = \begin{cases} 1 & \text{if some } \mathrm{X}_q = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

The final decision of the network is Equation 2.7 which is an estimation of the probability of whether speech is present in the input signal given the evidence from the SNR measurements. Both these inferences involve propagating information from the bottom of the network to the top. Other types of inferences can be calculated as well. For example, we can quickly deduce that $\Pr[\mathrm{X}_q = 0 | \mathrm{Z} = 0, \mathbf{m}_1, \ldots, \mathbf{m}_Q] = 1$ from the OR structure of the network. Another posterior probability that is useful is the following quantity which we use in the next section to estimate the parameters of the network.

$$\begin{aligned} &\Pr[\mathrm{X}_q = 1 | \mathrm{Z} = 1, \mathbf{m}_1, \ldots, \mathbf{m}_Q] \\ &\quad = \frac{\Pr[\mathrm{Z} = 1 | \mathrm{X}_i = 1, \mathbf{m}_1, \ldots, \mathbf{m}_Q] \Pr[\mathrm{X}_i = 1 | \mathbf{m}_1, \ldots, \mathbf{m}_Q]}{\Pr[\mathrm{Z} = 1 | \mathbf{m}_1, \ldots, \mathbf{m}_Q]} \\ &\quad = \frac{\sigma(\mathbf{w}_q \cdot \mathbf{m}_q)}{1 - \prod_{q=1}^{Q} (1 - \sigma(\mathbf{w}_q \cdot \mathbf{m}_q))} \end{aligned} \tag{2.8}$$

The fact that the union network is a probabilistic graphical network allows us to make such inferences in a principled way using Bayes' rule and other laws of probability.

2.3.1 Learning Algorithm

The parameters in the union network that require training are the weights of the linear discriminants. We can use maximum-likelihood estimation to determine the weights that are most likely to produce the observed data. Assume our model has Q subbands with weights $\mathcal{W} = \{\mathbf{w}_q\}$ where $q = 1 \ldots Q$. Furthermore, suppose we observe N examples $\mathcal{Z} = \{\mathrm{Z}_j\}$ with measurements $\mathcal{M} = \{\mathbf{m}_{q,j}\}$ for the qth band and jth observation where $q = 1 \ldots Q$ and $j = 1 \ldots N$. Then the likelihood function is as defined in below.

$$\mathcal{L}(\mathcal{Z}|\mathcal{W}, \mathcal{M}) = \prod_{j=1}^{N} \Pr[\mathrm{Z}_j|\mathcal{W}, \mathcal{M}] \tag{2.9}$$

Each term in this product can be derived from Equations 2.6 and 2.7. The weights $\mathcal{W}$ that maximize this equation are the weights that optimize the integrated decision of the network. Maximizing Equation 2.9 is difficult because we need to choose weights for our narrowband detectors that depend in a highly nonlinear way on wideband observations Z_j that specify whether speech is present in the input.

To make maximizing the likelihood function more tractable, we exploit the structure of the network to derive an Expectation-Maximization (EM) algorithm [6]. This iterative algorithm decomposes the problem of maximizing the likelihood function into two main steps. During the first phase, we calculate the posterior probabilities $\Pr[\mathrm{X}_{q,j}|\mathrm{Z}_j, \mathcal{W}, \mathcal{M}]$ for the jth example. In the second phase, we use these posterior probabilities as targets to train the linear discriminant weights in the qth band. We then update the probability densities in the network based on these new parameters in a bottom-up fashion using Equations 2.6 and 2.7. These two steps are repeated for each iteration of the algorithm. The EM algorithm insures monotonic convergence to a local maximum of the likelihood function [6]. Empirical evidence for this is shown in Figure 2-10 which depicts the log-likelihood function for the first one hundred iterations of the EM algorithm. Details of the algorithm are presented in Appendix A.

It is important to note that the EM algorithm is not guaranteed to produce the op-

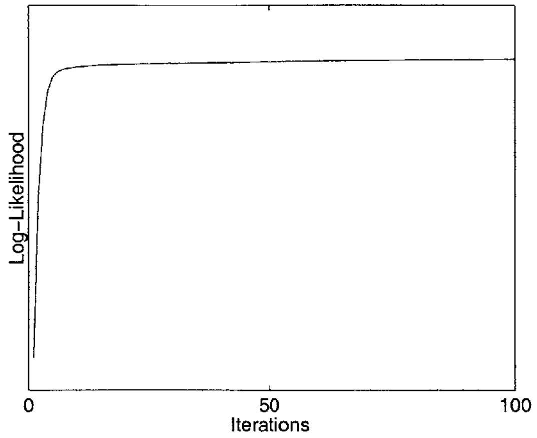

Figure 2-10: Monotonic convergence of the log-likelihood function for the union network EM algorithm.

timal weights $\mathcal{W}$ that maximize the likelihood for the union network. It only promises to find weights $\mathcal{W}$ corresponding to a local maximum of the likelihood function. Since the union network's likelihood function is highly nonlinear with respect to $\mathcal{W}$, the EM algorithm can easily get stuck at a local maximum. An example of a possible local maximum is a set of weights that detect speech using only n subbands where $n < Q$. We might obtain this result if we initialize the linear discriminant weights in certain bands to be very large negative values so that the output from these subbands is always close to zero. To help the EM algorithm find the global maximum of the likelihood function, we need to carefully select a starting point for it to begin its work. We choose to initialize the weights of the linear discriminants in the union model to the weight vector used in the narrowband model. This is a reasonable choice since these weights work very well for detecting speech in narrowband signals.

The EM algorithm for this union network solves a version of the multiple-instance learning problem [14]. In multiple-instance learning, a collection of examples is given a positive label if at least one of the instances is positive. Otherwise, the collection is labeled negatively. This ambiguous labeling is inherent when we train the union network since labels only indicate whether speech is present but fail to specify which subbands are responsible for detecting speech cues. Our EM algorithm solves this

	Training	Testing
Words	tin, berg, pill, knock, bill jog, tote, kame, pan, perch bah, lace, tonne, do	din, jock, dope, piss, half, mood, dad
Examples	5600	14000
(CF, BW) in Hz	(1000, 40), (1000, 100), (1000, 200), (1000, 340), (1000, 400), (1000, 600), (650, 60), (650, 120), (650, 220), (650, 390)	(1000, 40), (1000, 100), (1000, 200), (1000, 340), (1000, 400), (1000, 600), (650, 60), (650, 120), (650, 220), (650, 390)

Table 2.2: Training and testing data for speech detection task.

multiple-instance learning problem by inferring labels for the subbands.

2.3.2 Evaluation

We evaluated the union network by training and testing it against the data from the AT&T speech detection experiments. Our training data consisted of 14 different words. We trained the network to detect speech at two center frequencies and a variety of bandwidths as shown in Table 2.3.2. As before, when we trained the network to match a particular detection threshold t, we presented a signal with 30 dB SNR and a noise signal to serve as prototypical clean and noisy examples. We also produced a signal with SNR $(t-1)$ and $(t+1)$ with labels zero and unity respectively to motivate the union network to match the detection threshold. A total of 5600 training examples were used.

For testing, we used seven different words. These words were presented at SNR levels of 14, 10, 6, 4, 2, 0, -2, -4, -8, and -12 dB under two center frequencies and several different bandwidths (Table 2.3.2). We compare the union model detection thresholds to human performance in Figure 2-11. By using information from multiple critical bands, we see that the union network is able to lower its detection threshold as the speech bandwidth increases. On average, the detection thresholds are all about 1 to 2 dB higher for the union network compared to human performance.

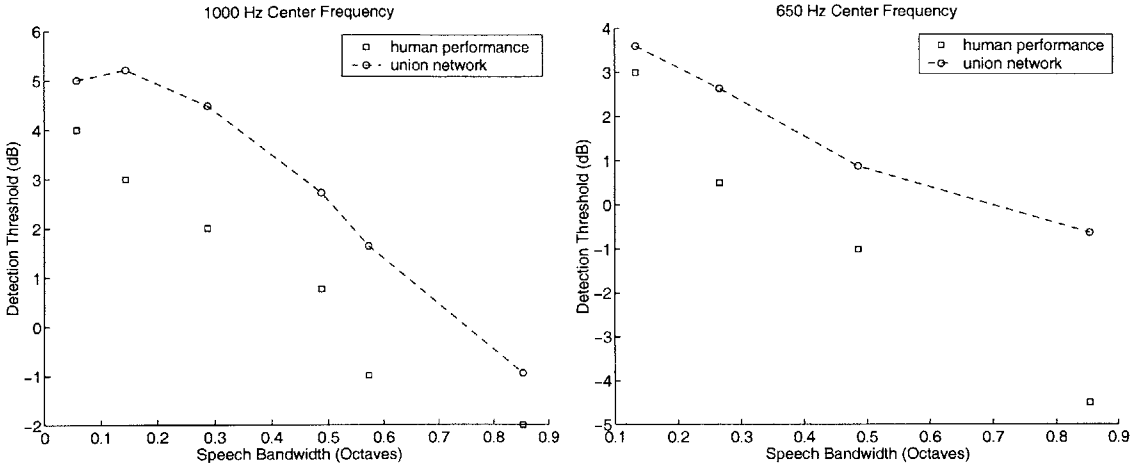

Figure 2-11: Union model and human speech detection comparison.

2.4 Weighted Union Model

In the union network, speech is detected if any of the subbands contain speech evidence. This processing explicitly gives equal weight to all the local decisions from the 32 linear discriminants. However, some bands may be more informative than others for detecting speech and should be weighted more heavily. For example, the lower frequencies of the spectrum often contain more speech information than the higher frequencies. It may also be undesirable to declare speech to be present when just one of the subbands has speech evidence. In some cases, it might be better to detect speech if two or more subbands contain speech cues. This strategy may be appropriate to lower the false positive rate of declaring the presence of speech when speech is actually not present in the input signal.

We construct a weighted union model to address these issues and improve the speech detection performance of the union model. The weighted union model is a generalization of the union model. It is shown in Figure 2-12. Instead of using a union network to process SNR measurements from the 32 subbands, this model uses a weighted union network. All other processing in the weighted union model is identical to the union model.

The weighted union network resembles the union network except for the weights it

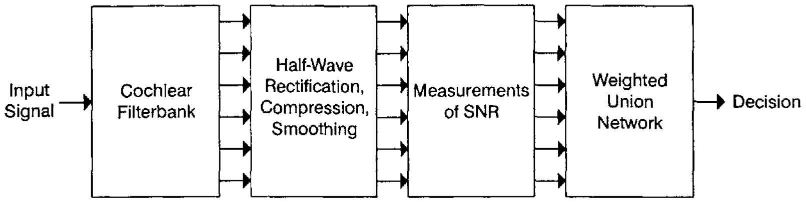

Figure 2-12: Schematic of weighted union model.

gives to the local decisions of the 32 linear discriminants. This weighting enables the network to favor the decisions of some bands more heavily than others. It can also restrain the network from declaring speech to be present if just one band contains speech evidence. As an example, suppose we weight the decision of each band by a value slightly less than 0.5. Then, if only one of the detectors in the subbands detects speech and outputs unity while the other bands output zero, the overall likelihood of detecting speech will be below 0.5 and thus the network declares speech to be absent in the input. However, if two subbands detect speech, then the overall output will be greater than 0.5 and the network will declare the input to contain speech. The weighted union network degenerates to the union network when we weight the decision of each band by unity.

The statistical model for the weighted union network is shown in Figure 2-13. As before, we use a probabilistic graphical network whose nodes are binary random variables and whose edges specify dependencies among the nodes. Again, vector $\mathbf{m}_q$ represents SNR measurements from the qth band, binary random variable X_q indicates whether subband q has high SNR, and binary random variable Z specifies the overall decision of the network. The network weights each of the X_q using binary random variable B_q which specifies whether the information from X_q should be considered. Random variable Y_q gives the weighted result of whether subband q contains speech cues. This weighting is performed using an AND operator on X_q together with B_q. We use an AND operator since Y_q should only be true if X_q is true and B_q indicates that the information in X_q should be used. We define these relationships below where

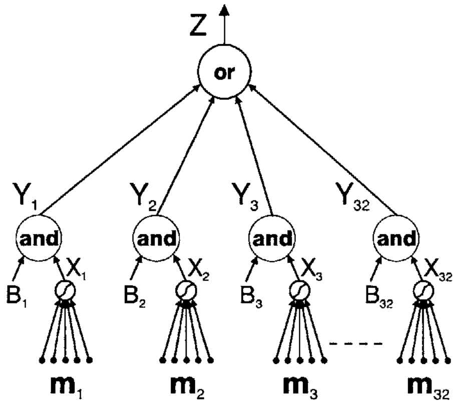

Figure 2-13: Weighted union network.

Q represents the number of subbands in the network.

$$\Pr[\mathrm{X}_q = \alpha | \mathbf{m}_q] = \begin{cases} \sigma(\mathbf{w}_q \cdot \mathbf{m}_q) & \text{if} \quad \alpha = 1 \\ 1 - \sigma(\mathbf{w}_q \cdot \mathbf{m}_q) & \text{if} \quad \alpha = 0 \end{cases} \tag{2.10}$$

$$\Pr[\mathrm{B}_q = \alpha] = \begin{cases} b_q & \text{if} \quad \alpha = 1 \\ 1 - b_q & \text{if} \quad \alpha = 0 \end{cases} \tag{2.11}$$

$$\Pr[\mathrm{Y}_q = 1 | \mathrm{X}_q, \mathrm{B}_q] = \begin{cases} 1 & \text{if} \quad \mathrm{X}_q = 1 \text{ and } \mathrm{B}_q = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

$$\Pr[\mathrm{Z} = 1 | \mathrm{Y}_1, \ldots, \mathrm{Y}_Q] = \begin{cases} 1 & \text{if some } \mathrm{Y}_q = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

As was the case in the union network, a variety of inferences may be made in this probabilistic graphical network. From the AND operation, we see that $\Pr[\mathrm{X}_q =$

$1|\mathrm{Y}_q = 1] = 1$ and $\Pr[\mathrm{B}_q = 1|\mathrm{Y}_q = 1] = 1$. The OR operation also specifies that $\Pr[\mathrm{Y}_q = 0|\mathrm{Z} = 0] = 1$. A host of other inferences may be made as well.

2.4.1 Learning Algorithm

Training the weighted union network is similar to training the union network. We need to determine weight vectors for each of the linear discriminants. In addition, we need to specify the weighting $\Pr[\mathrm{B}_q] = b_q$ as well.

To select the linear discriminant weights and the subbands weights $\Pr[\mathrm{B}_q]$, we maximize the likelihood function. We assume our model has Q subbands with weights $\mathcal{W} = \{\mathbf{w}_q\}$ and $\mathcal{B} = \{b_q\}$ where $q = 1 \ldots Q$. Let us also observe N examples $\mathcal{Z} = \{\mathrm{Z}_j\}$ with measurements $\mathcal{M} = \{\mathbf{m}_{q,j}\}$ for the qth band and jth observation where $q = 1 \ldots Q$ and $j = 1 \ldots N$. Then the likelihood function is as defined in below.

$$\mathcal{L}(\mathcal{Z}|\mathcal{W}, \mathcal{B}, \mathcal{M}) = \prod_{j=1}^{N} \Pr[\mathrm{Z}_j|\mathcal{W}, \mathcal{B}, \mathcal{M}] \tag{2.14}$$

The weights $\mathcal{W}$ and $\mathcal{B}$ that maximize this equation are the weights that optimize the integrated decision of the network. Maximizing this equation is difficult because it is not obvious how to choose weights for our narrowband detectors given the wideband observations Z_j that specify whether speech is present in the input.

We again maximize the likelihood function by exploiting the structure of the network to derive an EM algorithm. The EM algorithm allows us to maximize the likelihood function by specifying the posterior probability $\Pr[\mathrm{X}_{q,j}|\mathrm{Z}_j, \mathcal{W}, \mathcal{B}, \mathcal{M}]$ for training the linear discriminant in the qth band and by specifying the posterior probability $\Pr[\mathrm{B}_q|\mathrm{Z}_j, \mathcal{W}, \mathcal{B}, \mathcal{M}]$ for choosing the weighting in the qth band. The first phase of our EM algorithm is to calculate these posterior probabilities for the jth example. In the second phase of the algorithm, we use these posterior probabilities as targets for training the linear discriminant and subband weights in the qth band. We then update the probabilities in the network based on these new weights according to Equations 2.10, 2.11, 2.12, 2.13 before repeating these steps for another iteration of the algorithm. Like the EM algorithm introduced for the union network, this algorithm guarantees monotonic convergence to a local maximum of the likelihood function as seen in Figure 2-14. Using the EM algorithm to infer training labels for the linear discriminants and the subband weights solves a version of the multiple-instance learning problem. We present further details of this EM algorithm in Appendix B.

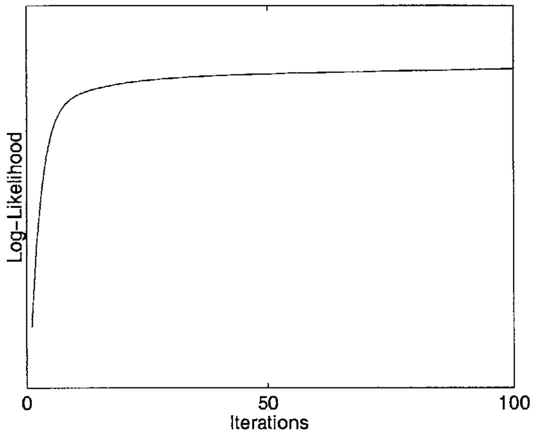

Figure 2-14: Monotonic convergence of the log-likelihood function for the weighted union EM algorithm.

As was the case for the union network, we need to carefully select an initialization of the parameters in the network to help the EM algorithm avoid local maximums in the likelihood function. Again, we decide to use the weight vector from the narrowband model to initialize the weights of the 32 linear discriminants. For $\Pr[\mathrm{B}_q = 1] = b_q$, we choose to set $b_q = 0.5$ for all $q = 1 \ldots 32$. Recall that when these weights are unity, the network is identical to the union network. Initializing $\Pr[\mathrm{B}_i = 1] = 0.5$ favors a strategy where speech is detected if two or more subbands contain speech evidence. Our evaluation will determine whether this behavior is more favorable than the union strategy where speech is detected if just one subband contains speech cues.

2.4.2 Evaluation

We trained and evaluated the weighed union network using the same training and testing procedures used for the union network (Table 2.3.2). The results are shown in Figure 2-15.

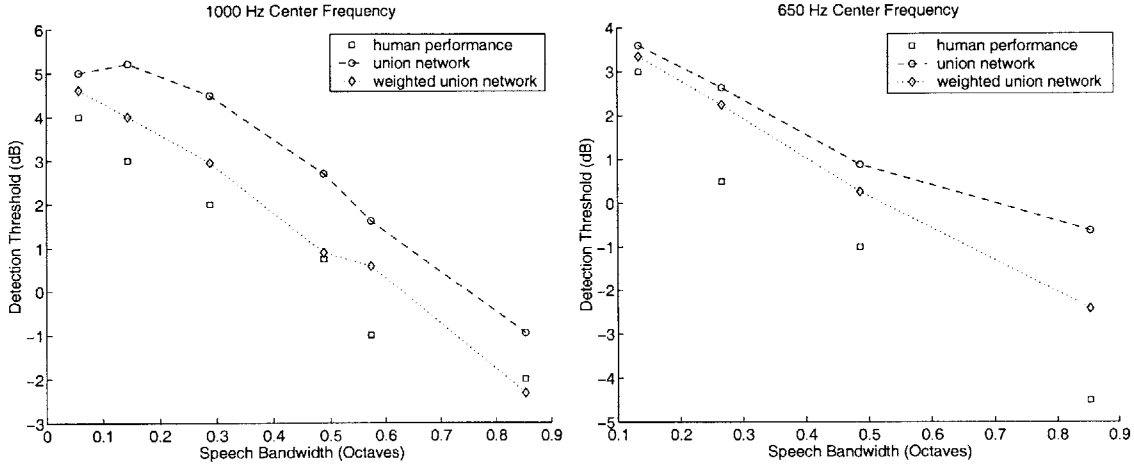

Figure 2-15: Weighted union model, union model, and human speech detection comparison.

By weighting the decisions of each band, the weighted union network is able to match the speech detection thresholds for human listeners much better than the union network. It is interesting to note that the optimal weights $\Pr[\mathrm{B}_i = 1]$ selected by the EM algorithm remain clustered around their initial value of 0.5. The improved results obtained using this weighting suggest that it is better to declare an input to contain speech if two or more subbands contain speech evidence.

2.5 Hierarchical Network

Both the union and weighted union networks detect the presence of speech using independent detectors in different parts of the spectrum. This type of processing is suggested by Fletcher's articulation experiments and product-of-errors rule for the articulation error. Results from the weighted union experiment however hint at an alternative approach to the detection of speech embedded in noise. In the weighted union experiment, we determined that better results are obtained if we declare an input to contain speech only if two or more bands detect speech cues. Since the bandlimited signals we presented to the network were contiguous in frequency, we expect the bands that contain speech evidence to be adjacent to one another. This suggests that informative speech cues often span several neighboring frequency bands. To take advantage of this, we need to process information from combinations of adjacent subbands.

We propose the hierarchical model to improve upon the performance of the union and weighted union models by considering combinations of adjacent critical bands as possible sources of informative speech cues. Subbands in this model are first analyzed in a pairwise manner. The result of this processing is propagated up for subsequent pairwise examinations until a global decision is made. This analysis resembles processing a binary hierarchy from the leaves up to the root node. We use this hierarchical combination strategy because the hierarchical structure of a binary tree is a particularly efficient way of performing pairwise analyses of the leaves. In our case, the leaves are the critical bands that we would like to analyze both individually and in combination for speech cues.

Figure 2-16 shows a block diagram of the hierarchical model. This figure is quite similar to the union model except for the hierarchical combination of the subbands and the use of the hierarchical network. The union model is just a special case of the hierarchical model that ignores all the subband combinations and only considers the individual narrowband measurements themselves.

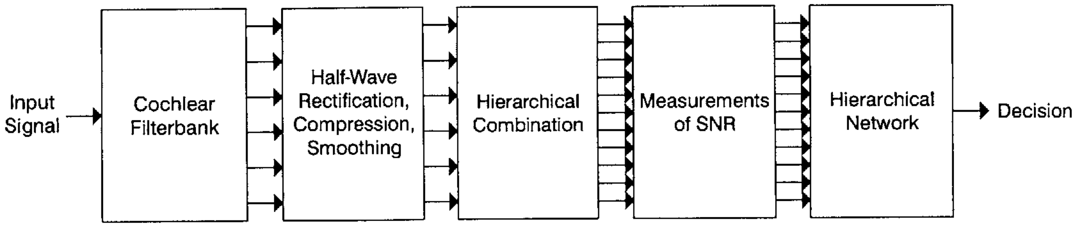

Figure 2-16: Schematic of hierarchical model.

Like the union and weighted union models, the hierarchical model first decomposes the input signal into 32 critical bands with center frequencies evenly spaced on a logarithmic scale from 300 to 1300 Hz with a constant bandwidth of 0.15 octaves. The outputs from these 32 bands are then half-wave rectified, compressed, and smoothed. Next, all 32 bands are hierarchically combined. This combination strategy is illustrated in Figure 2-17. Streams $s_1, \ldots, s_{32}$ correspond to the outputs of the original 32 bands. These streams are combined pairwise to form $s_{1,2}, s_{3,4}, \ldots, s_{31,32}$ which are in turn combined to form $s_{1,4}, \ldots, s_{29,32}$. This procedure continues until $s_{1,32}$ is produced forming a total of 63 streams. Our method for constructing stream $s_{i,j}$ is to average streams $s_i, s_{i+1}, \ldots, s_j$. By averaging narrowband signals, we hope to create new waveforms that may be useful for speech detection. These combined signals have greater bandwidth allowing the network to look for speech cues that span multiple critical bands. Once the 63 streams are formed, SNR measurements are computed for each stream. We compute the same SNR measures used for the narrowband, union, and weighted union models.

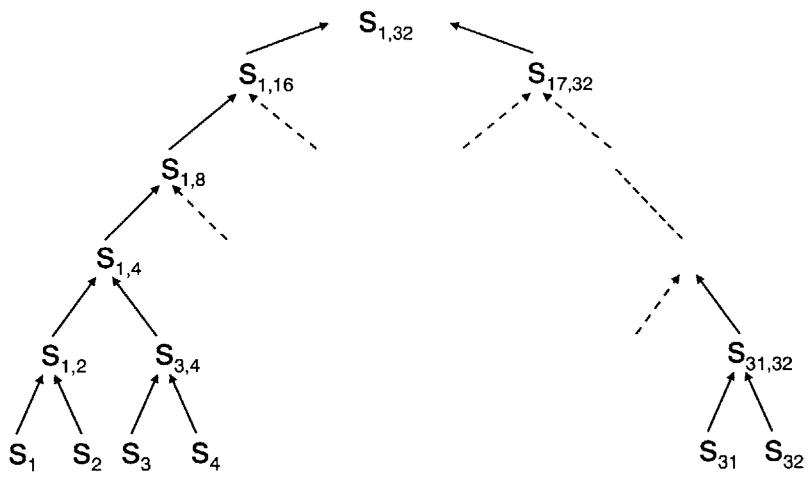

Figure 2-17: Hierarchical combination.

The hierarchical network processes the 63 sets of SNR measurements to determine whether the input signal contains speech. As its name indicates, this network processes these SNR measurements in a hierarchical manner. The network first analyzes measurements from the original 32 streams $s_1, \ldots, s_{32}$ in a pairwise manner. Suppose the network looks at streams s_i and s_{i+1}. Like the union network, if either signal contains speech evidence, the network declares that the input signal contains speech. However, if neither has speech cues, the network examines stream $s_{i,i+1}$ to determine whether speech evidence is present. This strategy of examining pairs of streams, and their combination if necessary, is the basic type of processing the hierarchical network performs.

The statistical model for the hierarchical network is a probabilistic graphical network like the union and weighted union networks. The edges of the network indicate dependencies among binary random variables. A small hierarchical network is shown in Figure 2-18. In this figure, vectors $\mathbf{m}_1, \ldots, \mathbf{m}_4$ represent measurements

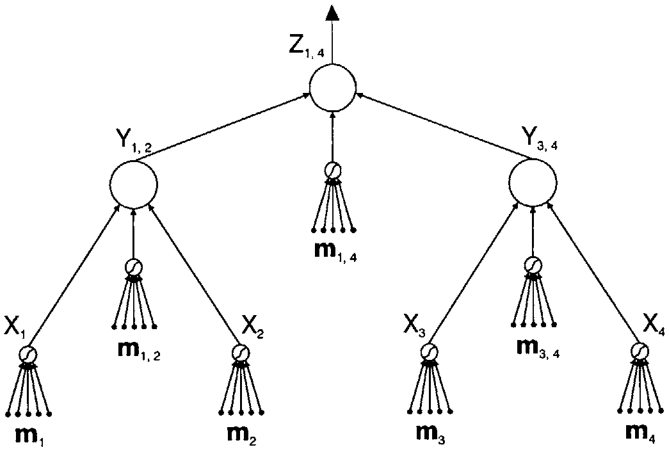

Figure 2-18: Small hierarchical network.

from streams $s_1, \ldots, s_4$ and $\mathbf{m}_{1,2}, \mathbf{m}_{3,4}, \mathbf{m}_{1,4}$ are measurements from the combined streams $s_{1,2}, s_{3,4}, s_{1,4}$ respectively. The actual hierarchical network we use for our experiments is much bigger however. It analyzes measurements $\mathbf{m}_1, \ldots, \mathbf{m}_{32}$ from the original 32 streams $s_1, \ldots, s_{32}$ as well as the measurements from the 31 additional streams formed from hierarchical combinations. Such a large network is unwieldy so we instead illustrate the network using Figure 2-18.

To decide whether speech is present in the input signal, the hierarchical network first checks whether measurements from the original streams indicate the presence of speech. Binary random variable X_i specifies whether stream s_i contains speech evidence. If any of the variables X_1, X_2, X_3, X_4 are true, the network will decide that the input contains speech. If not, measurements from the combined streams, $\mathbf{m}_{1,2}$ and $\mathbf{m}_{3,4}$ are examined. If either set of measurements exhibits speech evidence, the network declares the input to contain speech. Else, the network bases the final decision on its analysis of the measurements $\mathbf{m}_{1,4}$ from stream $s_{1,4}$. We characterize these relationships mathematically as follows.

$$\Pr[\mathrm{X}_i = \alpha|\mathbf{m}_i] = \begin{cases} \sigma(\mathbf{w}_i \cdot \mathbf{m}_i) & \text{if} \quad \alpha = 1 \\ 1 - \sigma(\mathbf{w}_i \cdot \mathbf{m}_i) & \text{if} \quad \alpha = 0 \end{cases} \tag{2.15}$$

$$\Pr[\mathrm{Y}_{i,j} = 1|\mathrm{X}_i, \mathrm{X}_j, \mathbf{m}_{i,j}] = \begin{cases} 1 & \text{if } \mathrm{X}_i = 1 \text{ or } \mathrm{X}_j = 1 \\ \sigma(\mathbf{w}_{i,j} \cdot \mathbf{m}_{i,j}) & \text{otherwise} \end{cases} \tag{2.16}$$

$$\Pr[\mathrm{Z}_{1,4} = 1|\mathrm{Y}_{1,2}, \mathrm{Y}_{3,4}, \mathbf{m}_{1,4}] = \begin{cases} 1 & \text{if } \mathrm{Y}_{1,2} = 1 \text{ or } \mathrm{Y}_{3,4} = 1 \\ \sigma(\mathbf{w}_{1,4} \cdot \mathbf{m}_{1,4}) & \text{otherwise} \end{cases} \tag{2.17}$$

These rules allow us to make a variety of inferences in the network. For example, if we know that speech is not present in the signal, then it must be the case that none of the binary random variables detect speech cues. Other useful inferences that we can calculate include posterior probabilities such as $\Pr[\mathrm{Y}_{1,2} = \alpha, \mathrm{X}_1 = 0, \mathrm{X}_2 = 0|\mathrm{Z}_{1,4} = \beta]$ where α and β are either 0 or 1.

2.5.1 Learning Algorithm

We select the weights for the linear discriminants in the hierarchical network by using maximum-likelihood estimation. Assume our model has weights $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_4, \mathbf{w}_{1,2}, \mathbf{w}_{3,4}, \mathbf{w}_{1,4}\}$ and that we observe examples $\mathcal{Z}_{1,4} = \{\mathrm{Z}_{(1,4),l}\}$ with measurements

$\mathcal{M}$. Then the likelihood function is as defined in below.

$$\mathcal{L}(\mathcal{Z}_{1,4}|\mathcal{W},\mathcal{M}) = \prod_{l=1}^{N} \Pr[\mathrm{Z}_{(1,4),l}|\mathcal{W},\mathcal{M}] \tag{2.18}$$

The weights $\mathcal{W}$ that maximize this equation are the weights that optimize the integrated decision of the network. Maximizing this likelihood function with respect to $\mathcal{W}$ is difficult because these weights depend on the observations $\mathrm{Z}_{(1,4),l}$ in a highly nonlinear way.

We maximize the likelihood function by exploiting the structure of the hierarchical network to develop an EM algorithm. Like the EM algorithms developed for the union and weighted union networks, this EM algorithm allows us to maximize the likelihood function by specifying certain posterior probabilities to use as target values for training the linear discriminants. Each iteration of the algorithm proceeds in two phases. In the first phase, we calculate these posterior probabilities. In the second phase, we train the linear discriminants using these posterior probabilities as target values. We then update the probability densities in the network according to Equations 2.15, 2.16, 2.17 before calculating another iteration of the algorithm. The hierarchical structure of this network allows us to develop an efficient recursive procedure that performs these two steps. Monotonic convergence to a local maximum of the likelihood function is shown in Figure 2-19 and is guaranteed by a general EM convergence theorem [6]. This EM algorithm for the hierarchical network solves a version of the multiple-instance learning problem by inferring training labels for the linear discriminants in the network. Appendix C details the development of this algorithm.

To avoid local maximums in the likelihood function, we initialize the weight vectors of the linear discriminants for streams $s_1, \ldots, s_{32}$ to the weight vector from the narrowband model like before. For the other combined streams $s_{i,j}$, we set their weights to the zero vector and their bias terms to a negative value. Setting the weights in this way causes these linear discriminants to initially ignore the speech evidence from the combined streams. This initialization models our strategy to analyze combined

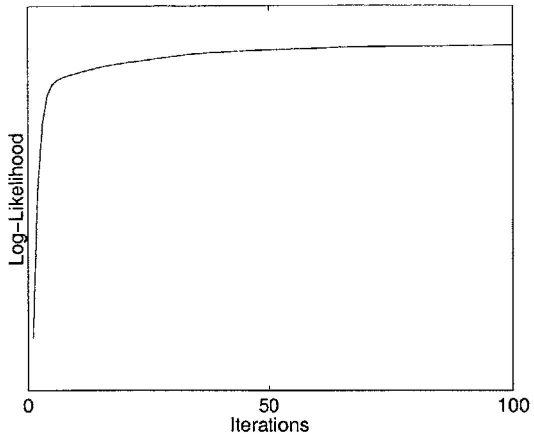

Figure 2-19: Monotonic convergence of the log-likelihood function for the hierarchical network EM algorithm.

streams for speech cues only when the original 32 subbands lack speech evidence. One of the disadvantages of working with the hierarchical network is that it is more prone to getting stuck at local maximums than the union or weighted union networks. This is because the hierarchical network has about twice as many parameters as the other two networks. It has 63 sets of linear discriminant weights whereas the union and weighted union networks only have 32.

2.5.2 Evaluation

We trained and evaluated the hierarchical network in the same way we trained and evaluated the union and weighted union networks. Table 2.3.2 outlines the data used for both training and testing.

The results of the evaluation are shown in Figure 2-20. We see that considering combinations of hierarchically combined subbands as possible sources for speech cues gives a modest improvement over the union model. Though the hierarchical model has lower detection thresholds than the union model, the weighted union model still does slightly better especially for the 1000 Hz center frequency case.

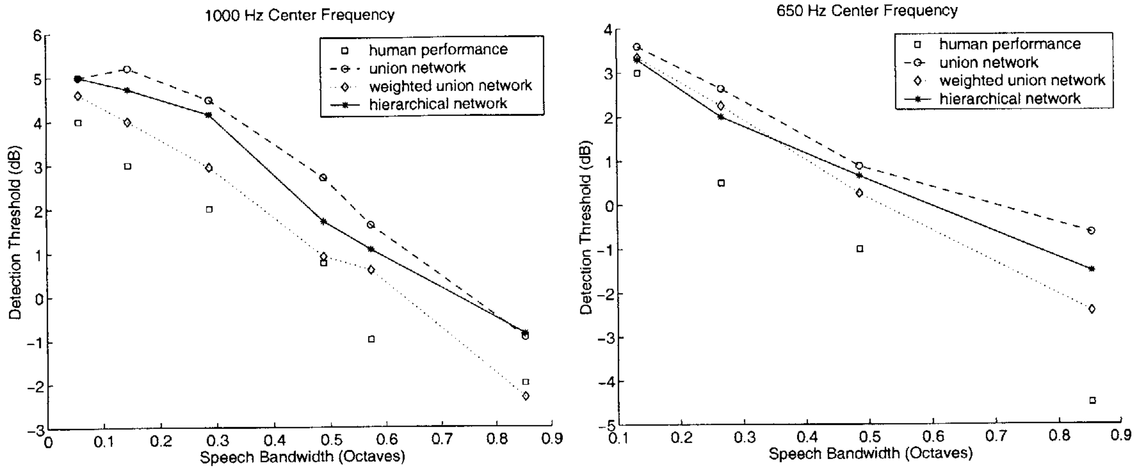

Figure 2-20: Hierarchical model, weighted union model, union model, and human speech detection comparison.

2.6 Discussion

We developed the union, weighted union, and hierarchical models to match human speech detection performance shown in Figure 2-3. Each of these models differs primarily in the probabilistic graphical model used to integrate speech cues from different frequency bands. The network for the union model declares the input to contain speech if it finds narrowband speech evidence in any of the subbands. The network for the weighted union model is more flexible and can be trained to require multiple subbands to contain speech evidence before declaring the presence of speech. Finally, we developed the hierarchical model whose statistical network analyzes combinations of streams for speech cues and integrates this evidence in a hierarchical manner.

Our evaluation of the three models shown in Figure 2-20 would indicate that the weighted union model is best for matching human speech detection. However, this evaluation is only for 650 Hz and 1000 Hz center frequencies so the performance of these models may not completely generalize for speech with different center frequencies and bandwidths. Since we only have measurements of human speech detection performance from the AT&T experiments, we do not have the data to test whether our models match human performance at other center frequencies and bandwidths.

What our results do tell us is that the union, weighted union, and hierarchical networks are able to behave similarly to human listeners. In particular, these models improve their detection thresholds as the speech bandwidth is increased. By emulating the multiband processing of the auditory system, these models are able to come close to matching human speech detection performance for bandlimited speech centered at 650 and 1000 Hz.

Chapter 3

Sonorant Detection

It is important for speech recognition systems to function effectively under poor listening conditions because the acoustic environment is often impossible to control. Although there has been much research in developing robust ASR techniques, modern recognizers are still unable to cope with speech corrupted by moderate amounts of noise. Human listeners, on the other hand, can recognize noisy speech with relative ease even when deprived of linguistic context and other aids [7, 15]. ASR systems stand to gain by emulating the processing that enables human listeners to recognize noisy speech.

Psychoacoustic experiments tell us that the robust detection of basic phonetic features is fundamental to robust human speech recognition. As seen in the work of Miller and Nicely [15], human listeners are able to identify nonsense syllables even when the signal has been filtered and corrupted with noise. More remarkably, the study found that despite poor listening conditions, certain sets of phonemes are difficult to confuse with other groups. These results motivate our study of automatic methods for the robust detection of the phonetic feature [+/−sonorant] which partitions phonemes into two groups: sonorants and obstruents. Like the phonetic groups studied by Miller and Nicely, the sonorant versus obstruent distinction is very resistant to noise and filtering [5]. This chapter details our multiband models for detecting the feature [+/−sonorant] under a variety of acoustic conditions.

3.1 Sonorants and Obstruents

Phonemes are basic phonetic units that correspond to different sounds in language. These building blocks are the smallest elements of language capable of conveying meaning. For example, changing the phoneme /b/ in *bee* to /f/ results in the new word *fee*. There are about four dozen different phonemes used in the English language.

We are concerned with detecting the phonetic feature [+/−sonorant] which distinguishes sonorant phonemes from obstruents. In articulatory terms, sonorants consist of those phonemes that are spoken without obstructing the airflow through the vocal tract. This group includes vowels, nasals, and approximants. Obstruent phonemes on the other hand are produced with at least partial obstruction of the airflow. These include stops, fricatives, and affricates. Sonorants and obstruent are illustrated in Table 3.1. For our purposes, we will treat stops, fricatives, affricates, pauses in speech, and anything that is not a sonorant as [−sonorant]. We will also interchangeably refer to obstruent phonemes as [−sonorant] and vice versa.

[+sonorant]	iy	(beet)	ih	(bit)	eh	(bet)	vowels
	ey	(bait)	ae	(bat)	aa	(bott)	
	aw	(bout)	ay	(bite)	ah	(but)	
	ao	(bought)	oy	(boy)	ow	(boat)	
	uh	(book)	uw	(boot)	ux	(toot)	
	er	(bird)	ax	(about)	ix	(debit)	
	axr	(butter)	ax-h	(suspect)			
	m	(mom)	n	(noon)	ng	(sing)	nasals
	em	(bottom)	en	(button)	eng	(washington)	
	nx	(winner)					
	l	(lay)	r	(ray)	w	(way)	approximants
	y	(yacht)	hh	(hay)	hv	(ahead)	
	el	(bottle)					
[−sonorant]	b	(bee)	d	(day)	g	(gay)	stops
	p	(pea)	t	(tea)	k	(key)	
	dx	(muddy)	q	(bat)			
	jh	(joke)	ch	(choke)			affricates
	s	(sea)	sh	(she)	z	(zone)	fricatives
	zh	(azure)	f	(fin)	th	(thin)	
	v	(van)	dh	(then)			

Table 3.1: Sonorants and obstruents.

Detecting the phonetic feature [+/−sonorant] requires that we capitalize on characteristics that distinguish obstruents from sonorants. Sonorant phonemes are articulated by periodic vibration of the vocal cords. Since the airstream is unobstructed for sonorants, the pitch of the speaker, or the fundamental frequency, should be evident in these signals. Periodicity (if any) in obstruents is generally muddied or destroyed by the partial or total obstruction of the airflow through the vocal tract. We expect only the strong periodicity cues in sonorants to weather corrupting influences in speech. Periodicity in sonorants is difficult to conceal because the energy in these signals is concentrated at the harmonics. It takes more white noise to inundate sonorants than it does for signals with energy spread more evenly across the spectrum. Consequently, portions of speech containing sonorants tend to have higher SNR as well as greater periodicity. As an example, consider the phonetic transcription of the utterance, "Calcium Makes Bones And Teeth Strong" in Figure 3-1. Sonorant regions, marked by a "+" sign, generally have greater energy than obstruent segments. In noise, we can expect sonorants to have higher SNR than obstruents. We use the presence of high SNR along with periodicity cues as heuristics for detecting sonorants in speech.

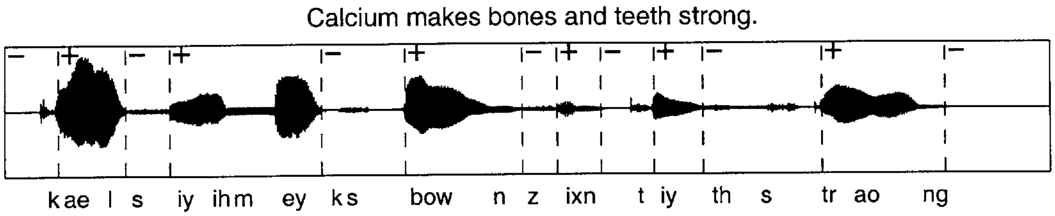

Figure 3-1: Phonetic transcription of "Calcium Makes Bones And Teeth Strong".

3.2 Multiband Models for Sonorant Detection

To explore the robust detection of the phonetic feature [+/−sonorant], we use our multiband models from the speech detection task. We use these models for three reasons. First, as demonstrated in the previous chapter, these multiband models closely match the behavior of human listeners at identifying speech under noisy bandlimited conditions. This is important because the ability to robustly identify sonorants from obstruents requires the capability to distinguish portions of acoustic waveforms that contain speech from those that do not. For example, portions of speech that are silences, pauses, or noise should always be labeled as [−sonorant]. We also apply our multiband speech detection models to sonorant detection because they emulate the processing in the peripheral auditory. It is a good idea to emulate auditory processing when constructing robust automatic methods for detecting sonorants since the recognition capabilities of human listeners is superior to modern ASR systems. The third reason to pursue robust sonorant detection using our multiband models is motivated by the work of Saul et al [21]. This study showed the success of a multiband union model in detecting the phonetic feature [+/−sonorant]. On this task, a union type model outperformed a mel-frequency cepstra Gaussian mixture model under a host of bandlimited noisy test conditions. Since most state-of-the-art recognizers use a mel-frequency cepstra representation for processing speech, this result shows that current recognizers may benefit greatly from a multiband approach to extracting phonetic features for speech processing. We hope to improve on the union model results in [21] by examining the sonorant detection capabilities of the weighted union and hierarchical models.

In this section, we discuss adapting our multiband models from the speech detection task to the task of detecting the phonetic feature [+/−sonorant]. We detail the front-end signal processing for our models, the measurements we extract to detect sonorants, and the statistical networks used.

3.2.1 Front-End Signal Processing

A schematic of our general model for sonorant detection is shown in Figure 3-2. The front-end signal processing for this sonorant detection model is similar to the front-end processing used for speech detection. We pass the input through a cochlear filterbank, apply nonlinearities, and extract measurements from each band. These measurement are then fed to a statistical network back-end which decides whether or not the input

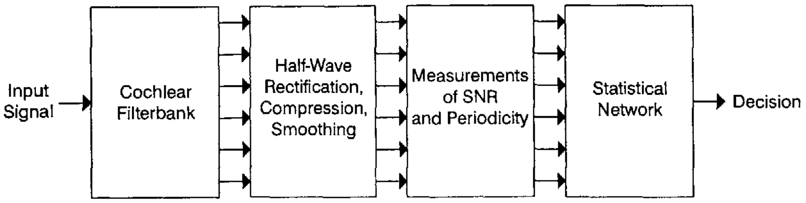
Input Signal
Cochlear Filterbank
Half-Wave Rectification, Compression, Smoothing
Measurements of SNR and Periodicity
Statistical Network
Decision

Figure 3-2: Components of sonorant detection model.

is a sonorant.

The front-end signal processing aids the detection of sonorants by emphasizing periodicity cues and regions of speech with high SNR. The first stage of processing consists of a cochlear filterbank which resolves the input signal into 32 overlapping critical bands. This simulates the multiband decomposition of acoustic signals in the auditory system. We use critical band filters with center frequencies that are evenly spaced on a logarithmic scale from 225 Hz to 3625 Hz with a constant bandwidth of 0.15 octaves. The input waveforms we use for the sonorant detection task are sampled at 8 kHz so the filterbank adequately covers the available spectrum.

Once the input is resolved into 32 critical bands, nonlinearities are applied. Each subband is half-wave rectified to imitate the unidirectional response of hair cells along the basilar membrane and then cube-rooted to simulate the compressive behavior of auditory filters. These two processes also serve to emphasize periodicity cues in the subbands. Should any of the bands have energy at two or more adjacent harmonics, these operations create intermodulation distortions [9] which concentrate energy at frequencies corresponding to the sums and differences of various integer multiples of the speaker's pitch. This processing aids the detection of sonorants since the pitch of the speaker should be more evident in sonorants than obstruents. Our final operation is to smooth the 32 streams by the linear process of downsampling to 1 kHz. Downsampling preserves the signal envelope and greatly reduces the amount of data we need to process. It also retains the energy concentrated near the speaker's pitch which may range anywhere from 50 to 300 Hz for speech.

We illustrate the effects of these nonlinearities with an example. The spectrogram

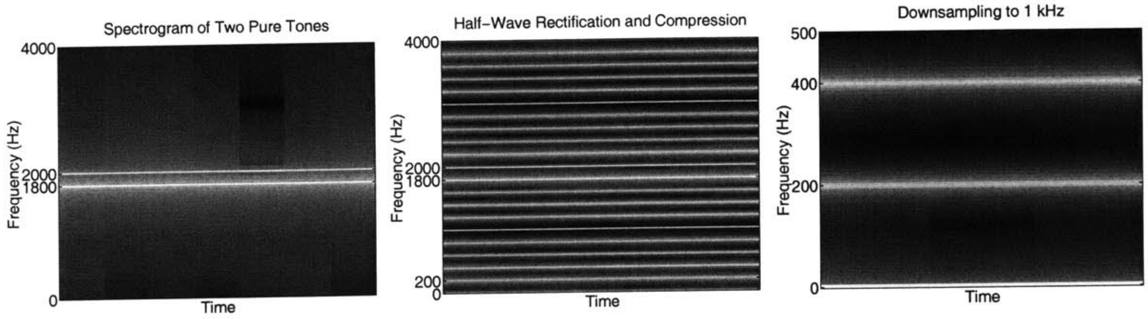

Figure 3-3: Spectrograms illustrating the effects of nonlinear processing.

on the left in Figure 3-3 shows two sinusoids sampled at 8 kHz with frequency 1800 and 2000 Hz. If we imagine these two signals as adjacent harmonics in a particular critical band, then what our nonlinearities do is to extract a 200 Hz signal which corresponds to a multiple of the pitch. We first create energy at 200 Hz through intermodulation distortions produced by half-wave rectification and cube-root compression as seen in the middle spectrogram. Finally, the spectrogram on the right shows the result of downsampling the signal. This final signal captures energy at 200 Hz (and also at 400 Hz) as desired.

For each of the 32 streams of data from the nonlinear processing, we compute 20 measurements for every contiguous, non-overlapping 16ms frame. With a 1 kHz sampling rate, this corresponds to partitioning the streams from each band into frames with 16 discrete samples. The first measure we compute is a local estimate of the SNR. This is calculated by taking the logarithm of the ratio of the current frame energy to the minimum energy of the neighboring twenty frames. We define the energy of a frame as the sum of the squares of the 16 samples. To deal with regions of silence, a small positive offset is added to the denominator of the ratio. The next four measurements are periodicity statistics derived from an autocorrelation sequence. Longer 64 ms frames are used to compute this sequence in order to include multiple pitch periods for deep male voices. We compute the autocorrelation sequence $a[n]$ using the technique advocated by Boersma in [3]. Let $x[n]$ with mean μ be the signal we would like to calculate the autocorrelation of. Assume it is nonzero from

$n = 0 \ldots M$. We can then define the following quantities.

$$w[n] = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi n}{M}\right), \quad \text{for } 0 \leq n \leq M \qquad y[n] = (x[n] - \mu)w[n]$$

$$z[n] = \frac{\sum_{k=0}^{M} y[k]y[k-n]}{\sum_{k=0}^{M} y[k]y[k]} \qquad v[n] = \frac{\sum_{k=0}^{M} w[k]w[k-n]}{\sum_{k=0}^{M} w[k]w[k]}$$

Note that $w[n]$ is just the Hanning window. The signals $z[n]$ and $v[n]$ are the normalized autocorrelations of the zero mean signal and the Hanning window, respectively. These autocorrelations may be efficiently calculated using fast discrete Fourier transform techniques. The autocorrelation sequence $a[n]$ we would like to calculate is just $a[n] = z[n]/v[n]$. As shown in [3], $a[n]$ is much more robust for periodicity detection than the traditional autocorrelation sequence. From the sequence $a[n]$, we extract the maximum peak, the minimum valley, the average of the peaks, and the average of the valleys as our periodicity measures. All four measures give an estimate of the periodicity of the signal. For example, larger peaks indicate greater periodicity as do deep valleys. The remaining 15 measures are computed by including all pairwise multiplications of the five measures described above. This augmentation of the feature space improved the sonorant detection abilities of our models.

3.2.2 Statistical Networks

We use probabilistic graphical networks to integrate the SNR and periodicity measures from the various bands. Recall that measurements are computed for every 16 ms frame of input speech. Our networks must use these multiband measurements to decide which frames are derived from sonorant sounds and which are not. Given a critical band decompositions of the acoustic waveform, our models infer whether the wideband signal is a sonorant by employing hidden variables to represent meaningful cues across the speech spectrum. These hidden variables distributed across frequency are integrated to reach a global decision. This multiband approach to sonorant detection fits the multiple-instance learning paradigm [14] just like the speech detection task in Chapter 2. We examine the performance of four different probabilistic graphical networks for the sonorant detection task.

The first network we consider is the union network described in Chapter 2. For the sonorant detection task, we regard the 32 linear discriminants in the network as sonorant cue detectors. These detectors work in parallel on measurements from individual bands to find sonorant cues. If any of the bands contain sonorant evidence, the network as a whole declares the input frame to be [+sonorant]. This strategy allows the network to ignore bands corrupted by noise and integrate information from the cleaner portions of the spectrum.

We also examine the performance of the weighted union network on the sonorant detection task. This variant of the union network also independently detects sonorant cues from across the spectrum. However, it is able to weight the decisions of the linear discriminants to favor some bands over others. A result of this weighting is that the network no longer is restrained to declare the input to be a sonorant if just a single band contains sonorant evidence. Instead, it can learn a weighting that requires multiple bands to register sonorant cues before asserting the input to be a sonorant. The structure and EM learning algorithm for the weighted union network is described in Chapter 2.

The third network we evaluate is the hierarchical network. It is a probabilistic graphical network that considers hierarchically combined subbands as possible sources for sonorant cues. Specifically, the hierarchical model combines the original 32 bands to form an additional 31 streams of data. To produce intermodulation distortions in these 31 new data streams, we square them. These combined streams have wider bandwidth enabling the network to look for sonorant cues that span multiple critical bands. Often, these wider bands contain sonorant cues that are not evident in individual critical bands. For example, suppose a particular subband captures only one harmonic of speech. With a single harmonic, the nonlinearities in the front-end are unable to produce intermodulation distortions at integer multiples of the fundamental frequency preventing the detection of useful periodicity cues. However, if we average this band with an adjacent band that also contains a single harmonic, we create a two harmonic signal which we square to emphasize frequencies at integer multiples of the speaker's pitch. Examining combined subbands is an improvement over the union and weighted union networks since these networks only analyze the original 32 subbands independently of one another. The hierarchical network gains this advantage at the cost of doubling the computation needed for the union and weighted union networks. A schematic of the hierarchical model is shown in Figure 3-4. This model differs from Figure 3-2 in that signals need to be hierarchically combined after the nonlinearities are applied. The hierarchical combination strategy and other details about the model are detailed in Chapter 2.

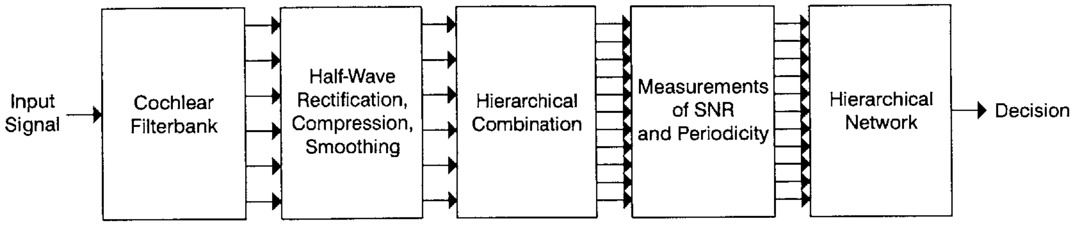

Figure 3-4: Schematic of hierarchical model.

Our final statistical network is a linear discriminant. Given measurements $\mathbf{m}$ and a weight vector $\mathbf{w}$, the output of the network is $\sigma(\mathbf{w}\cdot\mathbf{m})$ where $\sigma(x) = 1/[1+\exp(-x)]$. Like the hierarchical network, the linear discriminant takes measurements from all 63 streams to detect the phonetic feature [+/−sonorant]. The difference between the two models is that the linear discriminant uses no prior information to integrate sonorant cues from across the spectrum. It processes data from all the streams to detect sonorants whereas the hierarchical network registers the presence of a sonorant if just one of its streams has sonorant evidence. Since there are 63 streams each with 20 measurements, the linear discriminant has a total of 1260 parameters. We add one extra parameter to accommodate a bias term. Training the linear discriminant is more computationally difficult than training the hierarchical network. It is hard to train a linear discriminant with over a thousand parameters using fast techniques like Newton's method because of the expensive matrix operations required. Gradient descent procedures are more feasible but they require substantially longer run-times for convergence. The EM algorithm on the other hand is able to separate the training for the hierarchical network into several small logistic regressions each with about 20 weights. This decomposition allowed us to train the hierarchical network significantly faster than the large linear discriminant.

3.3 Experiments

Our task is to evaluate the [+/−sonorant] detection abilities of the union, weighted union, hierarchical, and linear discriminant models. We know that human listeners maintain the ability to identify sonorants even under poor listening conditions. Miller and Nicely's work [15] also illustrates that human listeners can make basic phonetic distinctions when speech is bandlimited and corrupted by noise. Since we are interested in emulating the robust detection abilities of the auditory system, we would like our models to reliably detect sonorants under both quiet and noisy bandlimited conditions.

We evaluated our models using the standard TIMIT speech corpus [8]. This data set contains phonetic transcriptions that are manually aligned with the speech waveforms. We used speech from the first dialect region for training and testing purposes. The training set consisted of 304 sentences with a total of 56077 frames of speech. The testing set contained 88 sentences with 17203 speech frames. Recall that each frame of speech is 16 ms in duration. Frames were labeled as sonorants if the majority of the samples spanned by the frame came from a sonorant phoneme. Otherwise, they were labeled as obstruents. We treat vowels, nasals, and approximants as [+sonorant] and stops, fricatives, affricates, and regions of silence as [−sonorant] with three exceptions. Like in the sonorant detection evaluation in Saul et al, the flapped /d/ ("ladder") was treated as [+sonorant] and the voiceless /h/ ("hay") and devoiced schwa ("suspect") were treated as [−sonorant]. About fifty percent of the frames were sonorants.

To have our models detect sonorants under a variety of quiet and noisy conditions, we present training data consisting of both clean and 0 dB SNR wideband speech. The clean speech gives prototypical examples of uncorrupted sonorants and obstruents while the 0 dB SNR signal provides examples of degraded speech. Training using the clean speech examples is straightforward since we have wideband labels from the phonetic transcriptions that specify which frames of speech are [+sonorant] and which are not. For noisy speech, training is not as straightforward. When noise is added to an obstruent frame, the frame remains an obstruent. Unfortunately, when noise is added to a sonorant frame, it may no longer be recognizable even by human listeners. For these noisy sonorant examples, we do not have any labels indicating whether they should be detected as sonorants. To deal with this missing label problem, we recall that human listeners are able to identify basic phonetic distinctions under noisy conditions. It is thus reasonable to assume that at least some sonorant cues from a sonorant phoneme survive when noise is added. We can use this idea to train on noisy speech by requiring at least one frame within a noisy sonorant phoneme to exhibit sonorant cues. Let Φ be a binary random variable indicating whether a noisy phoneme with contiguous frames k through l is a sonorant. Then this idea can be formalized as follows.

$$\Pr[\Phi = 1 \big| \mathrm{M}, \mathrm{Z}_k, \ldots, \mathrm{Z}_l] = 1 - \prod_{i=k}^{l} \big(1 - \Pr[\mathrm{Z}_i = 1 \big| \mathrm{M}_i]\big) \qquad (3.1)$$

Binary random variable Z_i indicates whether frame i contains sonorant evidence. This random variable represents the decision made by a statistical model based on the collection of measurements M_i for the ith frame. Equation 3.1 collects these decisions across the frames of a single phoneme and sets $\Phi = 1$ if at least one of these frames contains sonorant cues. The utility of Equation 3.1 comes from the ease of computing posterior probabilities to derive target labels for noisy frames of speech. The target probability for the ith noisy sonorant frame is just the posterior probability of Z_i given $\Phi = 1$. Similarly, the target probability for the ith noisy obstruent frame is just the posterior probability of Z_i given $\Phi = 0$ which is always identically zero. These posterior probabilities enable us to assign labels to noisy frames of speech that are consistent with our assumption that sonorant evidence survives in at least some frames when noise is added to a sonorant phoneme.

3.3.1 Examples

We illustrate the [+/−sonorant] detection task with a few examples using the utterance "Calcium Makes Bones And Teeth Strong". For each of these examples, human listeners generally have no problem recognizing the corrupted or filtered speech. Figure 3-5 shows the performance of the linear discriminant, union, weighted union, and hierarchical models under quiet conditions. Time is measured across the horizontal axis in 16 ms increments. For each of these 16 ms frames, the models output Pr[+sonorant] as an estimate of how likely the frame is a sonorant. The waveforms in the figure show the time evolution of Pr[+sonorant] for each model processing the input waveform shown in the top figure. Vertical lines indicate the manual [+/− sonorant] segmentation for clean speech. Sections marked with a "+" correspond to frames from sonorant phonemes. Those regions with a "−" indicate non-sonorant sounds. Of the four models, the linear discriminant has the best performance on this clean speech example. In areas marked by "−", its output is consistently near zero.

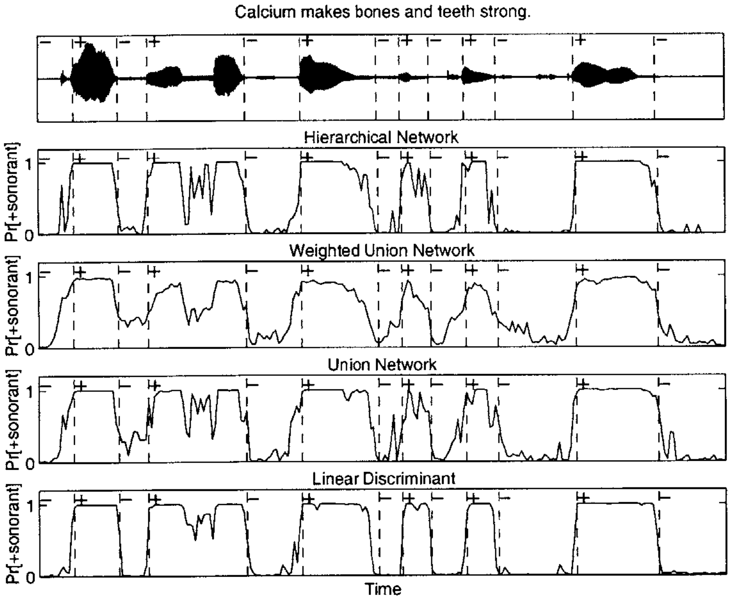

Figure 3-5: Top: Clean speech with TIMIT [+/− sonorant] segmentation for clean speech. Remaining diagrams show Pr[+sonorant] for hierarchical, weighted union, union, and linear discriminant models.

For sections with a "+", it has values close to unity. The other models also behave reasonably well. They generally give values above 0.5 for sonorant frames and output values less than 0.5 for obstruents and regions of silence. The union and weighted union models however seem a bit more prone to making errors in [−sonorant] regions than the hierarchical and linear discriminant models.

The next two examples illustrate how the models perform under noise. Figure 3-6 shows the behavior of the models when 0 dB white noise corrupts the input. Under

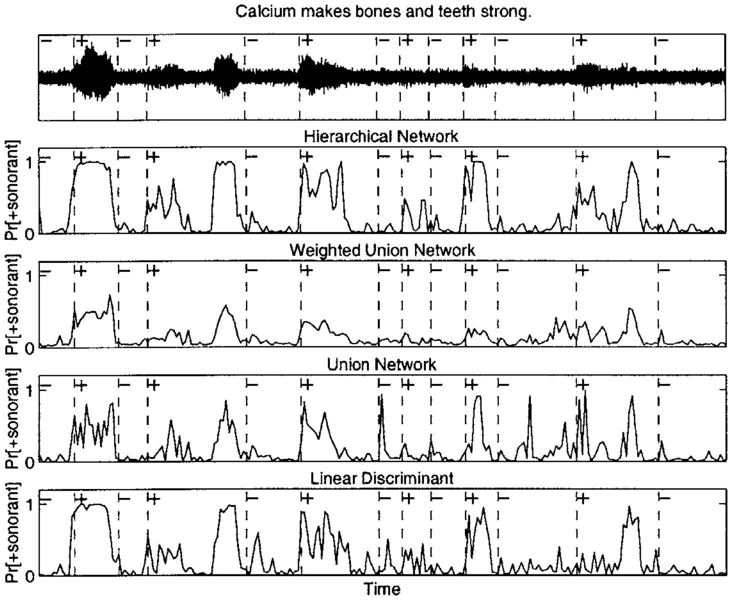

Figure 3-6: Top: Speech with 0 dB white noise with TIMIT [+/− sonorant] segmentation for clean speech. Remaining diagrams show Pr[+sonorant] for hierarchical, weighted union, union, and linear discriminant models.

these noisy conditions, all four models fail to detect a large number of the sonorant frames. This is expected since we trained our models to detect just one or more sonorant frames within a noisy sonorant phoneme to account for the destruction of speech cues by noise. We see that the hierarchical model misses the fewest sonorant frames while the weighted union model misses the most. The linear discriminant again does pretty well while the union model erroneously declares the presence of sonorants in a few [−sonorant] regions. Figure 3-7 shows the outputs of our models on speech corrupted with 1 to 2 kHz noise. Both the hierarchical model and linear

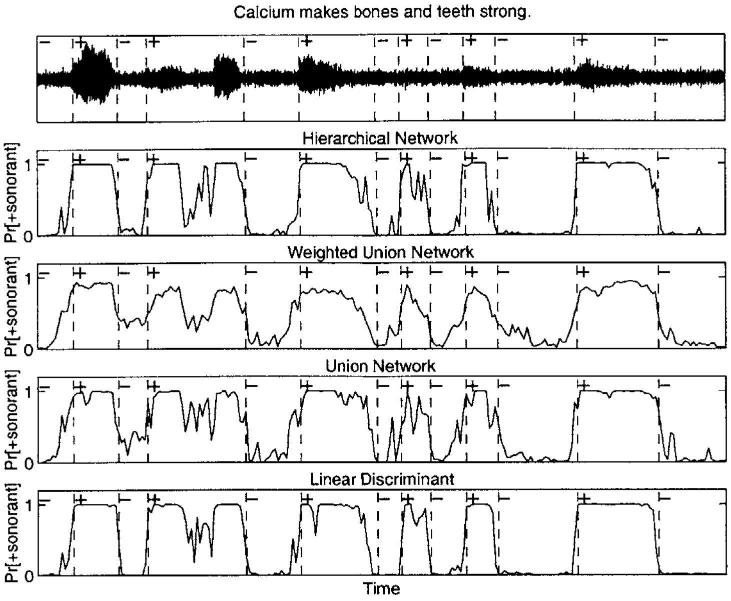

Figure 3-7: Top: Speech corrupted with 1-2 kHz noise with TIMIT [+/− sonorant] segmentation for clean speech. Remaining diagrams show Pr[+sonorant] for hierarchical, weighted union, union, and linear discriminant models.

discriminant are quite resistant to this bandlimited noise. In fact, the responses of all four models are similar to their behavior on clean speech. Apparently the models tolerate this type of bandlimited noise quite well.

In Figure 3-8, we see how the models respond to bandlimited speech from 1 to 4 kHz. The union, weighted union, and hierarchical models are able to integrate cues from across the spectrum and do a good job of detecting sonorant frames. However, the linear discriminant seems to ignore sonorant cues from 1 to 4 kHz since it misses every sonorant frame. We conclude that its good performance on the previous examples is the result of exclusively examining speech evidence below 1 kHz.

To illustrate the multiband processing in our models, we examine sonorant activity from different subbands. Figure 3-9 shows the detection of sonorants in the critical bands of the union model. The numbers on the left of the diagram show the center frequencies of each critical band filter. Each waveform represents $\Pr[X_i = 1|\mathbf{m}_i]$ across time where random variable X_i indicates whether subband i contains sonorant cues as defined in Equation 2.6. On the left, we see the sonorant activity in each

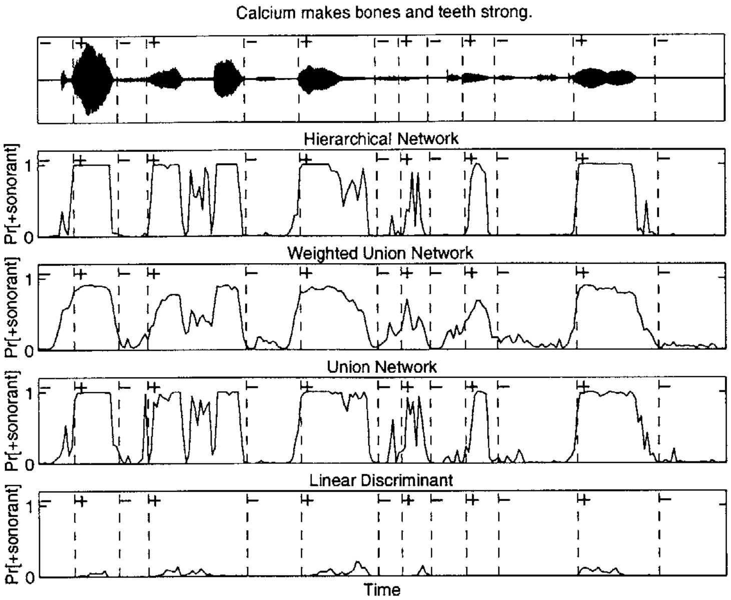

Figure 3-8: Top: Clean speech bandlimited to 1-4 kHz with TIMIT [+/− sonorant] segmentation for clean speech. Remaining diagrams show Pr[+sonorant] for hierarchical, weighted union, union, and linear discriminant models.

band when the input is clean speech. The figure on the right illustrates sonorant activity on speech corrupted with 1 to 2 kHz noise. Notice that bands in the vicinity of 1 to 2 kHz have their sonorant cues destroyed by noise. The union model is only able to decide on the presence of sonorants based on the uncorrupted bands. For the weighted union network, the sonorant detection activity in subbands is similar except the output from band i is multiplied by the scalar $\Pr[\mathrm{B}_i = 1]$.

Figure 3-10 shows the sonorant detection activity for the various streams of the hierarchical model on clean speech. Each plot represents the detection of sonorant cues in particular streams. The original 32 subbands along with their center frequencies are shown in the left column. Columns on the right correspond to sonorant detection in the combined streams $s_{i,j}$. For this clean speech example, the original 32 bands and the combined streams all have sonorant cues that are easy to detect. Figure 3-11 shows the sonorant detection for speech corrupted with 1 to 2 kHz noise. Although noise destroys most of the speech cues from 1 to 2 kHz, the hierarchically combined streams yield a few sonorant cues for these frequencies despite the fact that

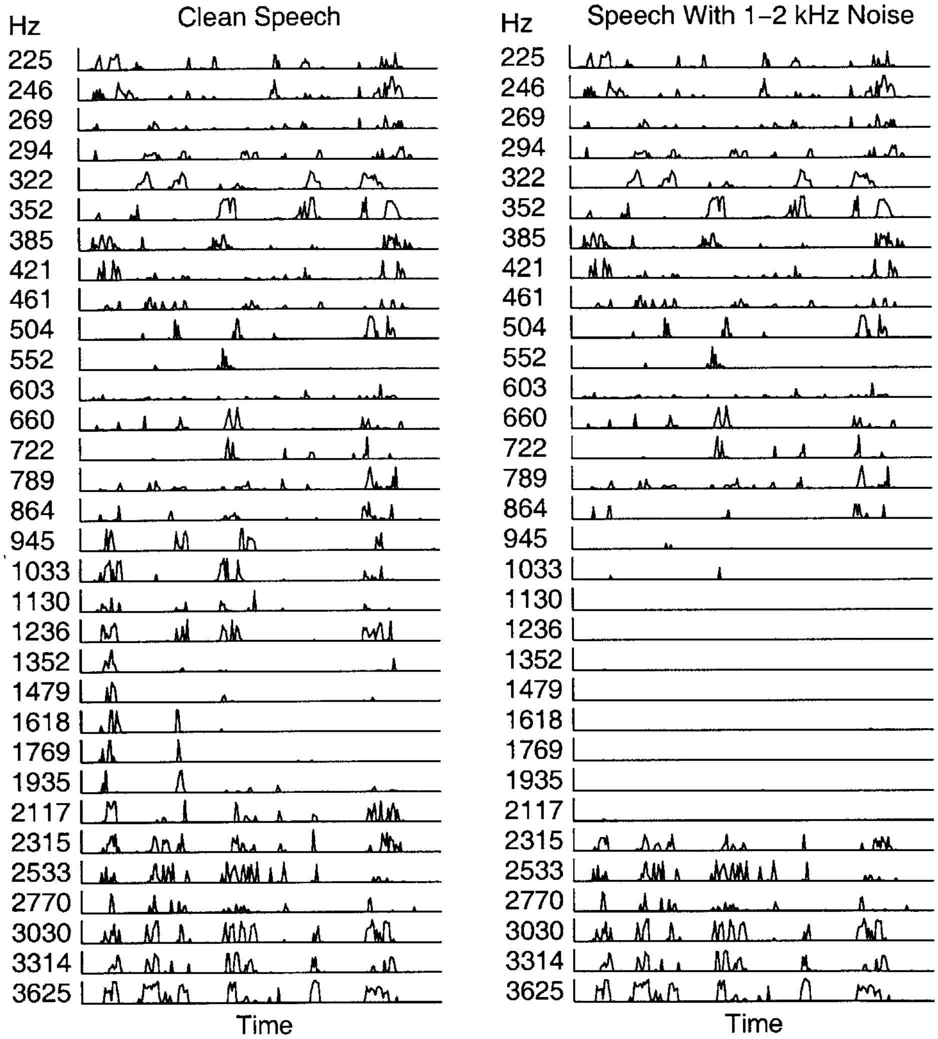

Figure 3-9: Multiband sonorant detection activity in union model for clean speech and speech corrupted with 1-2 kHz noise.

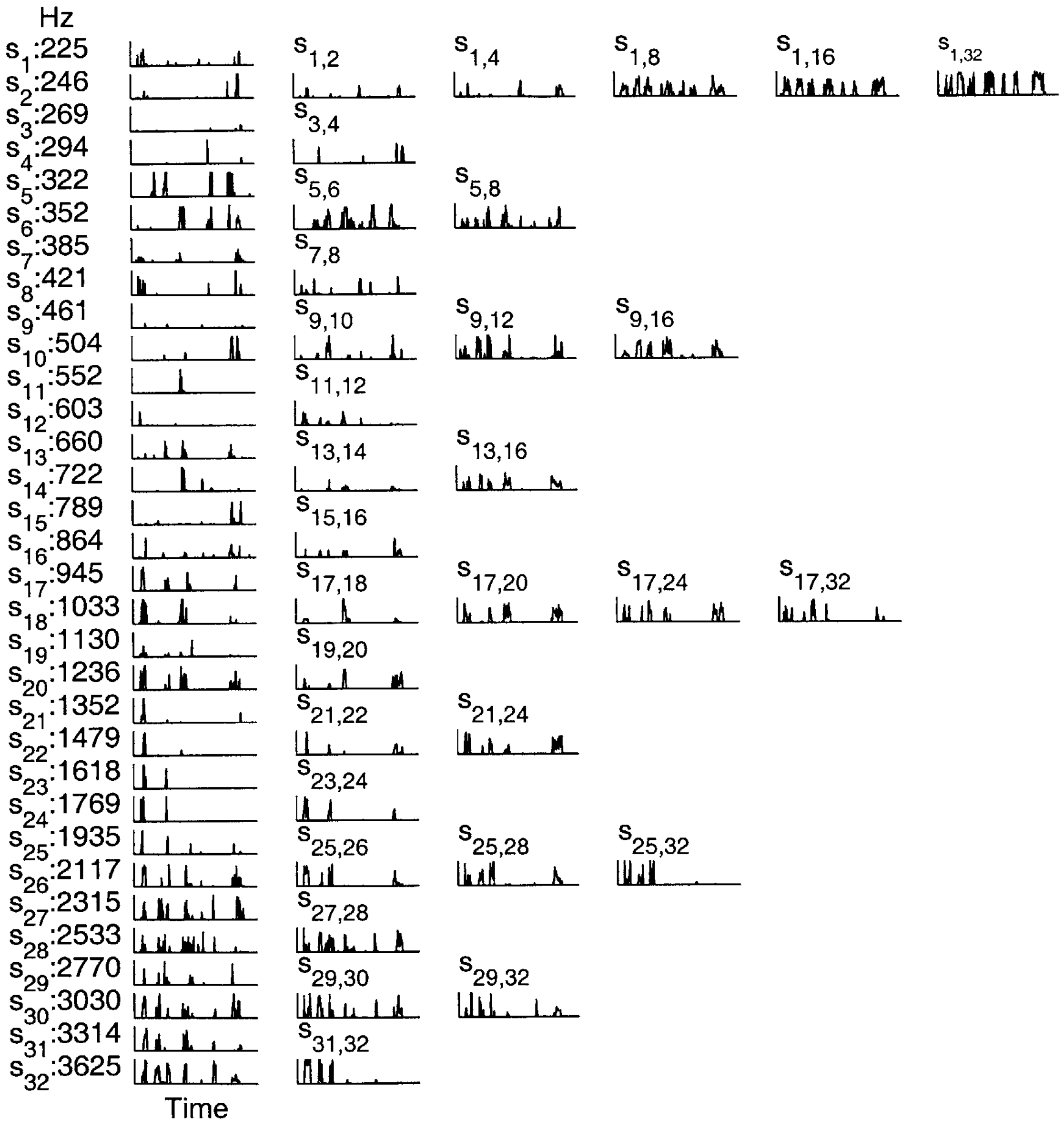

Figure 3-10: Multiband sonorant detection activity in hierarchical model for clean speech.

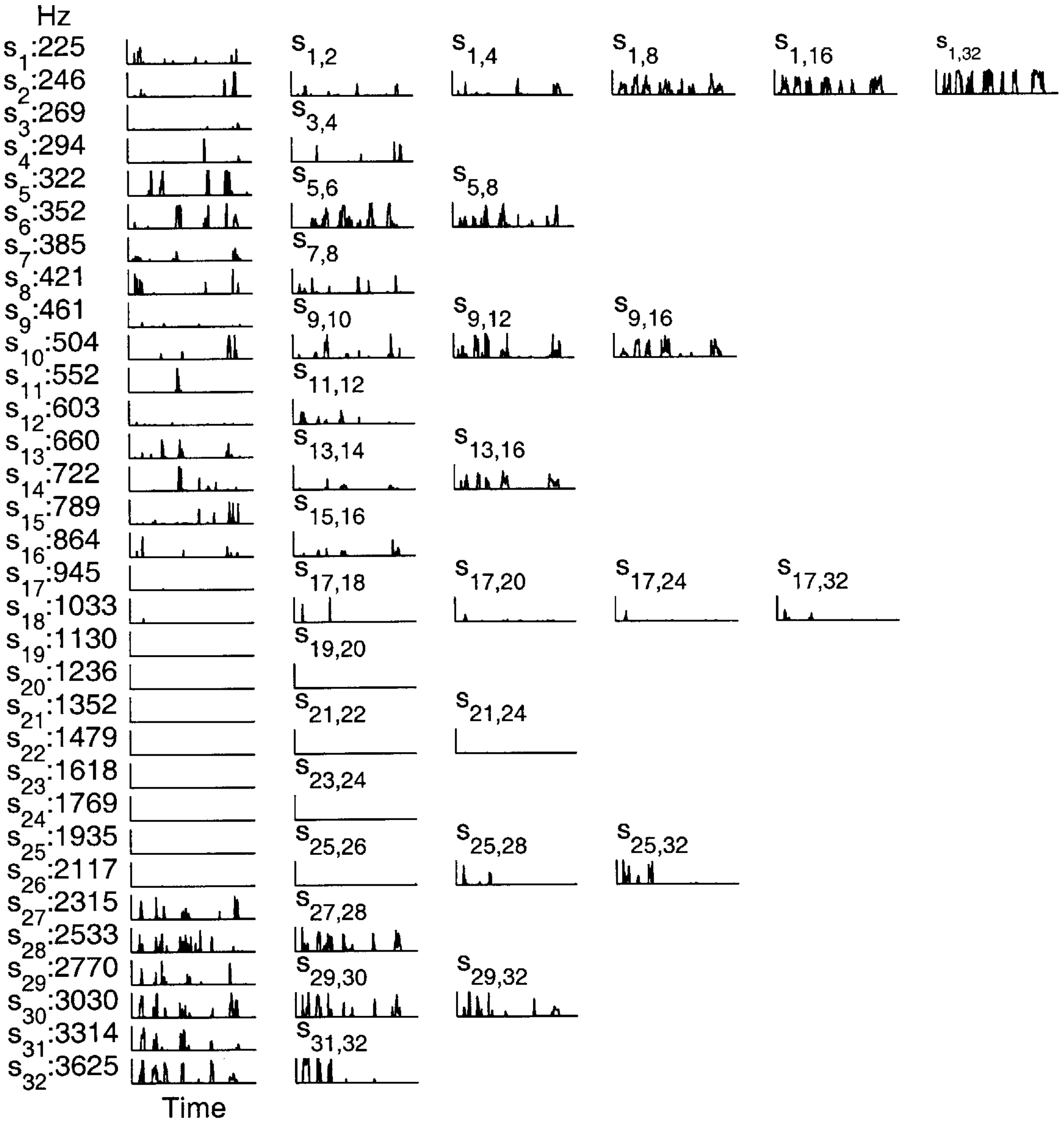

Figure 3-11: Multiband sonorant detection activity in hierarchical model for speech corrupted with 1-2 kHz noise.

sonorant evidence is not detected in streams $s_{17}, \ldots, s_{26}$. The advantage of the hierarchical network is its ability to analyze hierarchically combined streams for sonorant evidence. As evidenced in our examples, this ability allows the hierarchical model to do a good job of detecting sonorants under clean, noisy, and bandlimited conditions.

3.3.2 Results

We systematically evaluated the hierarchical, weighted union, union, and linear discriminant models under several acoustic conditions. These test conditions included clean speech, bandlimited speech, and speech corrupted by 0 dB white noise and 0 dB bandlimited Gaussian noise. We bandlimited signals to frequencies from 0 to 1 kHz, 1 to 2 kHz, 2 to 3 kHz, 3 to 4 kHz, and 1 to 4 kHz. Figure 3-12 shows the error rates for the four models under the test conditions. The frame error rates give the percentage of incorrectly labeled frames by the models. A sonorant frame is incorrectly labeled if Pr[+sonorant] is less than 0.5 and an obstruent or silence frame is wrong

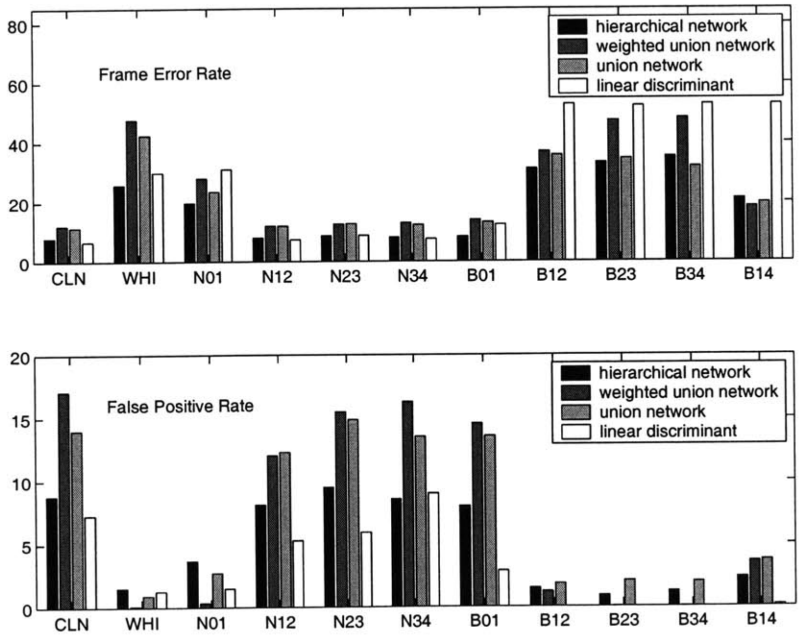

Figure 3-12: Error rates for [+/− sonorant] detection. Test conditions: clean (CLN), white noise (WHI), bandlimited noise (N$f_1 f_2$), and bandlimited speech (B$f_1 f_2$) with passband f_1 to f_2 kHz.

if Pr[+sonorant] is greater than 0.5. Given our sample size of 17203 test examples, it is reasonable to assume that differences in error rates have an approximate normal distribution. Using this assumption, we conclude that differences greater than 1.4% are significant at the 99% level.

The hierarchical model consistently had very low frame error rates in our evaluation. For clean speech, white noise, and bandlimited noise conditions, it had significantly lower error rates than the other multiband models. It also performed better than the linear discriminant except for a few cases where both had similar performance. Under bandlimited speech conditions, the hierarchical model also fared well. Its performance was comparable to the union network but generally much better than the weighted union and linear discriminant models. Note that the linear discriminant performs very poorly on bandlimited speech. Apparently it only analyzes the lower frequencies for sonorant cues. Whenever speech below 1 kHz was removed or corrupted with noise, the error rates for the linear discriminant would increase considerably.

We also computed the false positive rates for the various models. The false positive rate is the percentage of incorrect labelings in obstruent frames. This is a useful measure since it estimates the sensitivity of the models to noise and filtering in frames that should be insensitive to these effects. Given that only half the testing examples are obstruents, error differences greater than 2% are significant at the 99% level. The hierarchical model again performs well under this measure. It generally has lower false positive rates than the other multiband models especially for clean speech and bandlimited noise. The lowest false positive rates belong to the linear discriminant. Its reluctance to identify frames as sonorants is especially noticeable in the bandlimited speech cases.

This evaluation shows the advantages of the hierarchical model. By examining combinations of streams, the hierarchical model is better able to detect sonorants than the other multiband models. This is evident since the hierarchical model generally has lower frame error rates and lower false positive rates than the union and weighted union models. The hierarchical model also does well integrating sonorant cues from across the spectrum. Comparing it to the linear discriminant shows that although both examine the same data, the hierarchical model performs much better on bandlimited speech. Apparently, the linear discriminant only concentrates on sonorant cues under 1 kHz while the hierarchical model is able to effectively look for sonorant cues in higher frequencies as well. By examining combinations of streams and integrating the results in a hierarchical manner, the hierarchical model has good overall performance under a variety noise and bandlimited conditions.

Chapter 4

Conclusions

We argued in Chapter 1 that the robust identification of phonetic features is essential to robust human speech recognition. To study this fundamental problem of robust phonetic detection, we first developed multiband models to find speech embedded in noise. We constructed the union, weighted union, and hierarchical models to match human speech detection performance as measured in an experiment at AT&T Shannon Laboratory. Our models use a cochlear filterbank, auditory nonlinearities, and a measurement extractor coupled with a probabilistic graphical network to detect and integrate speech evidence distributed in frequency. EM algorithms were developed to train the networks. Evaluations showed that our multiband models come close to matching human speech detection thresholds.

We then applied our multiband models to the task of distinguishing sonorants from obstruents, a phonetic distinction resistant to many corrupting influences in speech. As shown in Saul et al [21], a union type network is good at sonorant detection under a variety of noisy and bandlimited conditions. In our evaluations, we found that the hierarchical model has even better sonorant detection abilities. Unlike the union and weighted union models which independently analyze subbands for speech evidence, the hierarchical model is able to look across frequency bands for speech cues. Its hierarchical structure also enables the use of a recursive EM algorithm that efficiently trains the network to integrate sonorant cues from across the spectrum. This is in contrast to the linear discriminant model which was difficult to train and failed to capitalize on high frequency speech evidence. Our experiments illustrate that the hierarchical model is robust to many types of noise and filtering by virtue of its hierarchical structure and its analysis of combined streams for meaningful speech cues.

We conclude by discussing some possible directions for this work. First, the front-end signal processing for our models requires further study. For example, more research certainly needs to be done to develop principled ways of selecting filterbanks for multiband models that emulate auditory processing. Intuitively, we would like critical band filters that are wide enough to capture phonetic cues yet narrow enough to avoid noise in large parts of the spectrum. In our models, we arbitrarily used filters with a constant bandwidth of 0.15 octaves because this choice yielded good results. A better understanding of human auditory processing should give us greater insight into this problem.

It is also worthwhile to explore how to improve sonorant detection within subbands. We mentioned earlier that augmenting our feature space to include pairwise multiplications of existing measurements improved our sonorant detection results. This suggests that our original measurements from each subband lack important sonorant information. By including more measurements, we may hope to capture sonorant cues missing from our original feature set. Another approach to improving subband sonorant detection is to use detectors that are more sophisticated than linear discriminants. More flexible classifiers may learn decision boundaries that are superior to the hyperplanes used by linear discriminants. Improving sonorant detection in subbands should certainly lead to better overall results.

Another challenge is to incorporate temporal processing into our multiband models. Our current models detect sonorants independently of decisions made in adjacent frames. However, we know a priori that if the previous frame is a sonorant (obstruent), then chances are that the next frame is also a sonorant (obstruent). Modeling this sort of temporal correlation should smooth the outputs of our models shown in Figures 3-5, 3-6, 3-7, 3-8 and reduce the error rates. We might also consider using dynamic programming in conjunction with our models to segment speech into sonorant and obstruent regions. A system capable of this segmentation would be very useful for many speech processing and recognition applications.

More work could also be done developing better ways of combining signals from different frequency bands. One alternative to the hierarchical model's strategy of averaging adjacent bands is to only combine signals with similar periodicity. A network with this ability may be able to process waveforms with different pitches separately and deal with periodic forms of noise. Combination strategies might also be developed to optimize certain measures of SNR or periodicity.

Finally, in addition to multiband speech and sonorant detection, our models are also applicable to other multiple-instance learning problems. For example, consider the problem of detecting pictures containing waterfalls [14]. A modified hierarchical model might attack this problem by decomposing candidate pictures into a collection of small sub-images and analyzing each for evidence suggesting the presence of waterfalls. If necessary, the model may then examine combinations of these sub-images as well. The combination strategy in this case might be to concatenate sub-images to form larger sub-images of the original picture. With this type of processing, a hierarchical network would be able to solve the problem of distinguishing pictures with waterfalls from those that do not. It would be instructive to apply our probabilistic graphical models to such problems to see how they compare against other multiple-instance learning algorithms.

Appendix A

Union EM Algorithm

Equation 2.9 gives the likelihood function we would like to maximize in order to derive the weights $\mathcal{W}$ that optimize the integrated decision of the union network. We can make maximizing this likelihood function more tractable by introducing hidden variables and by using an EM algorithm. The hidden variables we introduce correspond to the random variables in Figure 2-9. We define $\mathcal{X}_1 = \{\mathrm{X}_{1,j}\}$, $\mathcal{X}_2 = \{\mathrm{X}_{2,j}\}$, $\ldots$, $\mathcal{X}_Q = \{\mathrm{X}_{Q,j}\}$ where binary random variable $\mathrm{X}_{q,j}$ indicates whether the jth example in the qth band contains evidence for the phenomenon we are trying to detect. Using this notation, we specify a new likelihood function as shown in Equation A.1.

$$\mathcal{L}(\mathcal{Z}, \mathcal{X}_1, \ldots, \mathcal{X}_Q | \mathcal{W}, \mathcal{M}) = \prod_{j=1}^{N} \Pr[\mathrm{Z}_j | \mathrm{X}_{1,j}, \ldots, \mathrm{X}_{Q,j}] \cdot \Pr[\mathrm{X}_{1,j} | \mathbf{w}_1, \mathbf{m}_{1,j}] \\ \cdot \Pr[\mathrm{X}_{2,j} | \mathbf{w}_2, \mathbf{m}_{2,j}] \ldots \Pr[\mathrm{X}_{Q,j} | \mathbf{w}_Q, \mathbf{m}_{Q,j}] \quad (A.1)$$

This expression is referred to as the complete-data likelihood. The conditional probabilities in this expression are given by Equations 2.6 and 2.7.

To find the weights that maximize the likelihood function, the EM algorithm specifies that we choose $\mathcal{W}$ to maximize the expected value of the logarithm of the complete-data likelihood given the observations $\mathcal{Z}$ and the weights from the previous iteration.

$$
\begin{aligned}
\mathcal{W}^{(k)} &= \operatorname*{argmax}_{\mathcal{W}} \mathrm{E}\big[\ln \mathcal{L}(\mathcal{Z}, \mathcal{X}_1, \ldots, \mathcal{X}_Q | \mathcal{W}, \mathcal{M}) \big| \mathcal{Z}, \mathcal{W}^{(k-1)}, \mathcal{M}\big] \\
&= \operatorname*{argmax}_{\mathcal{W}} \sum_{j=1}^{N} \sum_{q=1}^{Q} \mathrm{E}\big\{\ln \Pr[\mathrm{X}_{q,j} | \mathbf{w}_q, \mathbf{m}_{q,j}] \big| \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{M}\big\} \\
&= \operatorname*{argmax}_{\mathcal{W}} \sum_{j=1}^{N} \sum_{q=1}^{Q} \Pr[\mathrm{X}_{q,j} = 1 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_q \cdot \mathbf{m}_{q,j})) \\
&\qquad\qquad + \Pr[\mathrm{X}_{q,j} = 0 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_q \cdot \mathbf{m}_{q,j}))
\end{aligned}
\tag{A.2}
$$

This algorithm insures monotonic convergence to a local maximum of the likelihood function [6].

Maximizing Equation A.2 can be achieved by maximizing each term individually. To maximize each term, we hope to select a weight vector such that $\sigma(\mathbf{w}_q \cdot \mathbf{m}_{q,j})$, the output of the linear discriminant in the qth band, is equal to the posterior probability $\Pr[\mathrm{X}_{q,j} = 1 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{M}]$. We calculate these posterior probabilities using Bayes' rule.

$$
\Pr[\mathrm{X}_{q,j} = 1 | \mathrm{Z}_j = 1, \mathcal{W}^{(k-1)}, \mathcal{M}] = \frac{\sigma(\mathbf{w}_q^{(k-1)} \cdot \mathbf{m}_{q,j})}{1 - \prod_{q=1}^{32}(1 - \sigma(\mathbf{w}_q^{(k-1)} \cdot \mathbf{m}_{q,j}))} \tag{A.3}
$$

$$
\Pr[\mathrm{X}_{q,j} = 0 | \mathrm{Z}_j = 0, \mathcal{W}^{(k-1)}, \mathcal{M}] = 1 \tag{A.4}
$$

These posterior probabilities in effect are the target labels we use to train the linear discriminant weights in each subband. Thus, our EM algorithm prescribes a means for maximizing the likelihood function with respect to $\mathcal{W}$ by decomposing the problem into several linear discriminant training problems.

Appendix B

Weighted Union EM Algorithm

We would like to maximize the likelihood function (Equation 2.14) to derive weights $\mathcal{W}$ and $\mathcal{B}$ that optimize the integrated decision of the weighted union network. To do this, we introduce hidden variables into our model and use an EM algorithm to make maximizing the likelihood function more tractable. The hidden variables we introduce correspond to the binary random variables shown in Figure 2-13. We define $\mathcal{X}_1 = \{\mathrm{X}_{1,j}\}$, $\mathcal{X}_2 = \{\mathrm{X}_{2,j}\}$, $\ldots$, $\mathcal{X}_Q = \{\mathrm{X}_{Q,j}\}$ where binary random variable $\mathrm{X}_{q,j}$ indicates whether the jth example in the qth band contains evidence for the phenomenon we are trying to detect. We also define $\mathcal{Y}_1 = \{\mathrm{Y}_{1,j}\}$, $\mathcal{Y}_2 = \{\mathrm{Y}_{2,j}\}$, $\ldots$, $\mathcal{Y}_Q = \{\mathrm{Y}_{Q,j}\}$ similarly. Then the complete-data likelihood function is as follows.

$$\begin{aligned} &\mathcal{L}(\mathcal{Z}, \mathcal{Y}_1, \ldots, \mathcal{Y}_Q, \mathcal{X}_1, \ldots, \mathcal{X}_Q, \mathrm{B}_1, \ldots, \mathrm{B}_Q | \mathcal{W}, \mathcal{B}, \mathcal{M}) \\ &= \prod_{j=1}^{N} \Pr[\mathrm{Z}_j | \mathrm{Y}_{1,j}, \ldots, \mathrm{Y}_{Q,j}] \prod_{q=1}^{Q} \Pr[\mathrm{Y}_{q,j} | \mathrm{X}_{q,j}, \mathrm{B}_q] \cdot \Pr[\mathrm{X}_{q,j} | \mathbf{w}_q, \mathbf{m}_{q,j}] \cdot \Pr[\mathrm{B}_q | b_q] \end{aligned} \tag{B.1}$$

The conditional probabilities in this expression are given by Equations 2.10, 2.11, 2.12, 2.13.

To select the weights that maximize the likelihood function, the EM algorithm specifies that we choose $\mathcal{W}$ and $\mathcal{B}$ to maximize the expected value of the logarithm of the complete-data likelihood given the observations $\mathcal{Z}$ and the weights from the previous iteration.

$$\begin{aligned}
&\operatorname*{argmax}_{\mathcal{W},\mathcal{B}} \mathrm{E}\left[\ln \mathcal{L}(\mathcal{Z}, \mathcal{Y}_1, \ldots, \mathcal{Y}_Q, \mathcal{X}_1, \ldots, \mathcal{X}_Q, \mathrm{B}_1, \ldots, \mathrm{B}_Q | \mathcal{W}, \mathcal{B}, \mathcal{M}) \middle| \mathcal{Z}, \mathcal{W}^{(k-1)}, \mathcal{M}\right] \\
&= \operatorname*{argmax}_{\mathcal{W}} \sum_{j=1}^{N} \sum_{q=1}^{Q} \Pr[\mathrm{X}_{q,j} = 1 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_q \cdot \mathbf{m}_{q,j})) \\
&\qquad\qquad + \Pr[\mathrm{X}_{q,j} = 0 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_q \cdot \mathbf{m}_{q,j})) \\
&\qquad\qquad + \Pr[\mathrm{B}_q = 1 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] \ln(b_q) \\
&\qquad\qquad + \Pr[\mathrm{B}_q = 0 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] \ln(1 - b_q)
\end{aligned} \tag{B.2}$$

Like in the case of the union network learning algorithm, this EM procedure guarantees monotonic convergence to a local maximum of the likelihood function [6].

We maximize Equation B.2 by maximizing each individual term. Maximizing each term requires that the outputs of the linear discriminants and subband weights match certain posterior probabilities in Equation B.2. As an example, to maximize the first term we would like $\sigma(\mathbf{w}_q \cdot \mathbf{m}_{q,j})$, the output of the linear discriminant analyzing subband q, to equal the posterior probability $\Pr[\mathrm{X}_{q,j} = 1 | \mathrm{Z}_j, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}]$. The posterior probabilities in Equation B.2 can be calculated using Bayes' rule. First, we define the following.

$$\begin{aligned}
x_{q,j} &= \Pr[\mathrm{X}_{q,j} = 1 | \mathbf{m}_{q,j}] = \sigma(\mathbf{w}_q^{(k-1)} \cdot \mathbf{m}_{q,j}), &\qquad b_q &= \Pr[\mathrm{B}_q = 1] = b_q^{(k-1)} \\
y_{q,j} &= 1 - x_{q,j} b_q, & z_j &= 1 - \prod_{q=1}^{Q} (1 - x_{q,j})
\end{aligned}$$

Then,

$$\Pr[\mathrm{X}_{q,j} = 1|\mathrm{Z}_j = 1, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] = \frac{x_{q,j}(1-b_q)}{y_{q,j}}(1 - \frac{x_{q,j}b_q}{z_j}) + \frac{x_{q,j}b_q}{z_j} \tag{B.3}$$

$$\Pr[\mathrm{X}_{q,j} = 0|\mathrm{Z}_j = 0, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] = \frac{1 - x_{q,j}}{y_{q,j}} \tag{B.4}$$

$$\Pr[\mathrm{B}_q = 1|\mathrm{Z}_j = 1, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] = \frac{b_q(1-x_{q,j})}{y_{q,j}}(1 - \frac{x_{q,j}b_i}{z_j}) + \frac{x_{q,j}b_i}{z_j} \tag{B.5}$$

$$\Pr[\mathrm{B}_q = 0|\mathrm{Z}_j = 0, \mathcal{W}^{(k-1)}, \mathcal{B}^{(k-1)}, \mathcal{M}] = \frac{1 - b_q}{y_{q,j}} \tag{B.6}$$

We use posterior probabilities B.3 and B.4 as targets to train the weight vector for the linear discriminant in subband q. The posterior probabilities B.5 and B.6 are used to determine the weighting $\Pr[\mathrm{B}_q = 1]$ for the qth band. By selecting $\mathcal{W}$ and $\mathcal{B}$ according to these posterior probabilities, we maximize the likelihood function in Equation 2.14.

Appendix C

Hierarchical EM Algorithm

We maximize the likelihood function (Equation 2.18) to derive weights $\mathcal{W}$ that optimize the integrated decision of the hierarchical network. To make this maximization easier, we introduce hidden variables into our network and use the EM algorithm. The hidden variables we introduce correspond to the binary random variables in Figure 2-18. We define $\mathcal{X}_1 = \{\mathrm{X}_{(1),l}\}$, $\mathcal{X}_2 = \{\mathrm{X}_{(2),l}\}$, $\ldots$, $\mathcal{X}_4 = \{\mathrm{X}_{(4),l}\}$ where binary random variable $\mathrm{X}_{(i),l}$ indicates whether stream s_i for the lth example contains evidence for the feature we are trying to identify. We also define $\mathcal{Y}_{1,2} = \{\mathrm{Y}_{(1,2),l}\}$, $\mathcal{Y}_{(3,4)} = \{\mathrm{Y}_{(3,4),l}\}$, $\mathcal{Z}_{(1,4)} = \{\mathrm{Z}_{(1,4),l}\}$ similarly to represent whether random variables $\mathrm{Y}_{(p,q),l}$ and $\mathrm{Z}_{(r,s),l}$ contain the cues we are looking for in lth example. Using this notation, we represent the complete-data likelihood function as follows.

$$\mathcal{L}(\mathcal{Z}_{1,4}, \mathcal{Y}_{1,2}, \mathcal{Y}_{3,4}, \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4 | \mathcal{W}, \mathcal{M}) = \prod_{l=1}^{N} \Pr[\mathrm{Z}_{(1,4),l} | \mathrm{Y}_{(1,2),l}, \mathrm{Y}_{(3,4),l}, \mathbf{w}_{1,4}, \mathbf{m}_{(1,4),l}]$$
$$\cdot \Pr[\mathrm{Y}_{(1,2),l} | \mathrm{X}_{(1),l}, \mathrm{X}_{(2),l}, \mathbf{w}_{1,2}, \mathbf{m}_{(1,2),l}] \cdot \Pr[\mathrm{Y}_{(3,4),l} | \mathrm{X}_{(3),l}, \mathrm{X}_{(4),l}, \mathbf{w}_{3,4}, \mathbf{m}_{(3,4),l}]$$
$$\cdot \Pr[\mathrm{X}_{(1),l} | \mathbf{w}_1, \mathbf{m}_{(1),l}] \cdot \Pr[\mathrm{X}_{(2),l} | \mathbf{w}_2, \mathbf{m}_{(2),l}] \cdot \Pr[\mathrm{X}_{(3),l} | \mathbf{w}_3, \mathbf{m}_{(3),l}] \cdot \Pr[\mathrm{X}_{(4),l} | \mathbf{w}_4, \mathbf{m}_{(4),l}] \tag{C.1}$$

The conditional probabilities in this expression are given by Equations 2.15, 2.16, 2.17.

To select weights that maximize the likelihood function, the EM algorithm specifies that we choose $\mathcal{W}$ to maximize the expected value of the complete-data log-likelihood given the observations $\mathcal{Z}_{1,4}$ and the weights from the previous iteration.

$$\begin{aligned}
\mathcal{W}^{(k)} &= \operatorname*{argmax}_{\mathcal{W}} \mathrm{E}\left[\ln \mathcal{L}(\mathcal{Z}_{1,4}, \mathcal{Y}_{1,2}, \mathcal{Y}_{3,4}, \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4 | \mathcal{W}, \mathcal{M}) \middle| \mathcal{Z}, \mathcal{W}^{(k-1)}, \mathcal{M}\right] \\
&= \operatorname*{argmax}_{\mathcal{W}} \sum_{l=1}^{N} \Pr[\mathrm{X}_{(1),l} = 1 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_1 \cdot \mathbf{m}_{(1),l})) \\
&+ \Pr[\mathrm{X}_{(1),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_1 \cdot \mathbf{m}_{(1),l})) \\
&+ \qquad \qquad \vdots \\
&+ \Pr[\mathrm{X}_{(4),l} = 1 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_4 \cdot \mathbf{m}_{(4),l})) \\
&+ \Pr[\mathrm{X}_{(4),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_4 \cdot \mathbf{m}_{(4),l})) \\
&+ \Pr[\mathrm{Y}_{(1,2),l} = 1, \mathrm{X}_{(1),l} = 0, \mathrm{X}_{(2),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_{1,2} \cdot \mathbf{m}_{(1,2),l})) \\
&+ \Pr[\mathrm{Y}_{(1,2),l} = 0, \mathrm{X}_{(1),l} = 0, \mathrm{X}_{(2),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_{1,2} \cdot \mathbf{m}_{(1,2),l})) \\
&+ \Pr[\mathrm{Y}_{(3,4),l} = 1, \mathrm{X}_{(3),l} = 0, \mathrm{X}_{(4),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_{3,4} \cdot \mathbf{m}_{(3,4),l})) \\
&+ \Pr[\mathrm{Y}_{(3,4),l} = 0, \mathrm{X}_{(3),l} = 0, \mathrm{X}_{(4),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_{3,4} \cdot \mathbf{m}_{(3,4),l})) \\
&+ \Pr[\mathrm{Z}_{(1,4),l} = 1, \mathrm{Y}_{(1,2),l} = 0, \mathrm{Y}_{(3,4),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(\sigma(\mathbf{w}_{1,4} \cdot \mathbf{m}_{(1,4),l})) \\
&+ \Pr[\mathrm{Z}_{(1,4),l} = 0, \mathrm{Y}_{(1,2),l} = 0, \mathrm{Y}_{(3,4),l} = 0 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}] \ln(1 - \sigma(\mathbf{w}_{1,4} \cdot \mathbf{m}_{(1,4),l}))
\end{aligned} \tag{C.2}$$

By a general convergence theorem [6], we know that this algorithm converges monotonically to a local maximum of the likelihood function.

We can maximize Equation C.2 by maximizing each term individually. To maximize each term, we hope to select weight vectors such that the outputs of the linear discriminants match the posterior probabilities in Equation C.2. For example, for the first term in the equation, we would like the output of the linear discriminant $\sigma(\mathbf{w}_1 \cdot \mathbf{m}_{(1),l})$ to match the posterior probability $\Pr[\mathrm{X}_{(1),l} = 1 | \mathrm{Z}_{(1,4),l}, \mathcal{W}^{(k-1)}, \mathcal{M}]$. These posterior probabilities can be computed using Bayes' rule. First, let us define the following.

$$\begin{aligned}
x_{(i),l} &= \Pr[\mathrm{X}_{(i),l} = 1 | \mathbf{m}_{(i),l}] = \sigma(\mathbf{w}_i^{(k-1)} \cdot \mathbf{m}_{(i),l}) \\
y_{(i,j),l} &= \Pr[\mathrm{Y}_{(i,j),l} = 1 | \mathrm{X}_{(i),l}, \mathrm{X}_{(j),l}, \mathbf{m}_{(i,j),l}] = 1 - (1 - x_{(i),l})(1 - x_{(j),l}) \\
&\qquad + \sigma(\mathbf{w}_{i,j}^{(k-1)} \cdot \mathbf{m}_{(i,j),l})(1 - x_{(i),l})(1 - x_{(j),l}) \\
z_{(i,j),l} &= \Pr[\mathrm{Z}_{(i,j),l} = 1 | \mathrm{Y}_{(i,p),l}, \mathrm{Y}_{(p,j),l}, \mathbf{m}_{(i,j),l}] = 1 - (1 - y_{(i,p),l})(1 - y_{(p,j),l}) \\
&\qquad + \sigma(\mathbf{w}_{i,j}^{(k-1)} \cdot \mathbf{m}_{(i,j),l})(1 - y_{(i,p),l})(1 - y_{(p,j),l})
\end{aligned}$$

We can then express the posterior probabilities in Equation C.2 as follows.

$$\Pr[\mathrm{X}_{(i),l} = 1 | \mathrm{Z}_{(1,4),l} = 1, \mathcal{W}^{(k-1)}, \mathcal{M}] = \frac{x_{(i),l}}{z_{(1,4),l}} \tag{C.3}$$

$$\Pr[\mathrm{X}_{(i),l} = 0 | \mathrm{Z}_{(1,4),l} = 0, \mathcal{W}^{(k-1)}, \mathcal{M}] = 1 \tag{C.4}$$

$$\begin{aligned}
&\Pr[\mathrm{Y}_{(i,j),l} = 1, \mathrm{X}_{(i),l} = 0, \mathrm{X}_{(j),l} = 0 | \mathrm{Z}_{(1,4),l} = 1, \mathcal{W}^{(k-1)}, \mathcal{M}] \\
&\quad = \frac{\sigma(\mathbf{w}_{i,j}^{(k-1)} \cdot \mathbf{m}_{(i,j),l})(1 - x_{(i),l})(1 - x_{(j),l})}{z_{(1,4),l}}
\end{aligned} \tag{C.5}$$

$$\Pr[\mathrm{Y}_{(i,j),l} = 0, \mathrm{X}_{(i),l} = 0, \mathrm{X}_{(j),l} = 0 | \mathrm{Z}_{(1,4),l} = 0, \mathcal{W}^{(k-1)}, \mathcal{M}] = 1 \tag{C.6}$$

$$\begin{aligned}
&\Pr[\mathrm{Z}_{(1,4),l} = 1, \mathrm{Y}_{(1,2),l} = 0, \mathrm{Y}_{(3,4),l} = 0 | \mathrm{Z}_{(1,4),l} = 1, \mathcal{W}^{(k-1)}, \mathcal{M}] \\
&\quad = \frac{\sigma(\mathbf{w}_{1,4}^{(k-1)} \cdot \mathbf{m}_{(1,4),l})(1 - y_{(1,2),l})(1 - y_{(3,4),l})}{z_{(1,4),l}}
\end{aligned} \tag{C.7}$$

$$\Pr[\mathrm{Z}_{(1,4),l} = 0, \mathrm{Y}_{(1,2),l} = 0, \mathrm{Y}_{(3,4),l} = 0 | \mathrm{Z}_{(1,4),l} = 0, \mathcal{W}^{(k-1)}, \mathcal{M}] = 1 \tag{C.8}$$

These equations serve as training labels for our linear discriminants. Equations C.3 and C.4 specify the targets for training the linear discriminant analyzing stream s_i. Likewise, C.5 and C.6 give the targets for training the linear discriminant analyzing stream $s_{i,j}$ and C.7 and C.8 are the labels for training the linear discriminant processing $s_{1,4}$. Procedurally, each iteration of this EM algorithm proceeds in two phases. The first phases computes these posterior probabilities. In the second phase, these posterior probabilities are used as target values for training the linear discriminants in the network. The probabilities of the hidden variables in the network are then updated using Equations 2.15, 2.16, 2.17 before we begin another iteration of the algorithm.

Training the hierarchical network in Figure 2-18 is relatively straightforward as outlined in this section. However, dealing with the larger hierarchical model used for processing 32 subbands and 31 hierarchically combined streams requires much more computation. Updating the probabilities of the hidden variables in the larger network is straightforward using equations similar to 2.15, 2.16, 2.17 but calculating the posterior probabilities is more complicated. To make this computation easier, we can exploit the structure of the network. Notice that computing the posterior probabilities in C.3 through C.8 requires only local information. For example, if the network in Figure 2-18 were only a small part of a larger hierarchy with training label Z_l for the lth observation, we can calculate posterior probabilities as follows.

$$\begin{aligned}
&\Pr[\mathrm{Y}_{(1,2),l} = 1, \mathrm{X}_{(1),l} = 0, \mathrm{X}_{(2),l} = 0 | \mathrm{Z}_l = 1, \mathcal{W}^{(k-1)}, \mathcal{M}] \\
&\quad = \Pr[\mathrm{Y}_{(1,2),l} = 1, \mathrm{X}_{(1),l} = 0, \mathrm{X}_{(2),l} = 0 | \mathrm{Z}_{(1,4),l} = 1, \mathcal{W}^{(k-1)}, \mathcal{M}] \qquad \text{(C.9)} \\
&\qquad \cdot \Pr[\mathrm{Z}_{(1,4),l} = 1 | \mathrm{Z}_l = 1, \mathcal{W}^{(k-1)}, \mathcal{M}]
\end{aligned}$$

$$\Pr[\mathrm{Y}_{(1,2),l} = 0, \mathrm{X}_{(1),l} = 0, \mathrm{X}_{(2),l} = 0 | \mathrm{Z}_l = 0, \mathcal{W}^{(k-1)}, \mathcal{M}] = 1 \qquad \text{(C.10)}$$

These equations show that posterior probabilities for the internal hidden variables can be found using posterior probabilities calculated from the hidden variables in the next level up. Thus, we can determine all the posterior probabilities needed by computing posterior probabilities in a recursive manner resembling a pre-order traversal of the network. This recursive EM algorithm allows us to efficiently train the network to select weights $\mathcal{W}$ that maximize the likelihood function.

Bibliography

[1] Jont B. Allen. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, October 1994.

[2] Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, 1995.

[3] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 17:97–100, 1993.

[4] Hervé Bourlard and Stéphane Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. *Proc. ICSLP '96*, 1:426–429, 1996.

[5] John Clark and Colin Yallop. *An Introduction to Phonetics and Phonology.* Blackwell Publishing Ltd, Oxford, U.K., 1995.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[7] Harvey Fletcher. *Speech and Hearing in Communication.* D. Van Nostrand, New York, 1953.

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM (printed documentation), 1992. NTIS order number PB91-100354.

[9] William M. Hartmann. *Signals, Sound, and Sensation.* Springer-Verlag, New York, 1998.

[10] Frederick Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, Cambridge, Massachusetts, 1997.

[11] Michael I. Jordan, editor. *Learning in Graphical Models.* MIT Press, Cambridge, Massachusetts, 1999.

[12] Wayne A. Lea. *Trends in Speech Recognition.* Prentice-Hall, Englewood Cliffs, New Jersey, 1980.

[13] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–16, 1997.

[14] Oded Maron and Tomas Lozano-Perez. A framework for multiple-instance learning. *Neural Information Processing Systems*, 10, 1998.

[15] George A. Miller and Patricia E. Nicely. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352, March 1955.

[16] Ji Ming and F. Jack Smith. Union: a new approach for combining subband observations for noisy speecy recognition. *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, 1999.

[17] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

[18] Joseph W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993.

[19] James W. Pitton, Kuansan Wang, and Biing-Hwang Juang. Time-frequency analysis and auditory modeling for automatic recognition of speech. *Proceedings of the IEEE*, 84(9), 1996.

[20] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[21] Lawrence K. Saul, Mazin G. Rahim, and Jont B. Allen. A statistical model for robust integration of narrowband cues in speech. *Computer Speech and Language*, 2001.

[22] Georg von Békésy. *Experiments in Hearing.* McGraw-Hill, New York, 1960. translated and edited by E. G. Wever.

[23] Victor W. Zue. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, 73(11):1602–1615, 1985.