Contents

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

A Multi-Band Approach to Automatic Speech Recognition*

Naghmeh Nikki Mirghafori

TR-99-04

January 1999

Abstract

Multi-band approaches have recently generated a great deal of interest in the automatic speech recognition (ASR) community. In this paradigm, each sub-frequency region of the speech signal is treated as a distinct source of information and the streams are combined after each is processed independently. Motivations for the multi-band paradigm include results from psycho-acoustic studies, robustness to noise, and potential for parallel processing.

The main contribution of this dissertation is the systematic exploration of an area of great interest to many in the research community, showing that multi-band ASR is a viable option, not just for improving recognition accuracy in the presence of noise, but also for clean speech. The work focused on the design and implementation of a multi-band system, analysis of some of its characteristics, and development of extensions to the paradigm.

An analysis in terms of phonetic feature transmission showed multi-band processing to be better than a comparable traditional full-band design in many cases. It was observed that some bands were more accurate in discriminating between some phonetic categories. It was hypothesized that combining the confused sub-band classes

*This report is a revised version of the author's thesis, which was submitted to the Department of Electrical Engineering and Computer Science on November 24, 1998 in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of California, Berkeley. This work was supervised by Professor Nelson Morgan. The thesis committee also included Professors Steven Greenberg, Jitendra Malik, and John Ohala.

would reduce the number of input classes and improve generalization. The size of the input space was reduced by almost 30%, and yet the global frame-level phonetic discrimination improved and the word recognition error did not change (the observed improvement was not statistically significant). The results were consistent with the original hypothesis.

The analysis also showed that the phonetic transitions in the sub-bands do not necessarily occur synchronously and are affected by conditions such as speaking rate and room reverberation. Relaxing the synchrony constraints in the sub-bands during word recognition was investigated. The experimental results suggested that removing the synchrony constraints for all phone to phone transitions is unlikely to be advantageous while significantly increasing computational cost.

The combination of the multi-band and the full-band system was studied. This combination reduced the word recognition error rate for the experimental clean speech task by about 23-29% compared to the baseline system. The results obtained are the best that we know of on the Numbers95 experimental database.

List of Figures

List of Tables

Chapter 1

Introduction

Speech is the most natural form of human communication. It is also the most suitable data entry medium for many applications, such as word processing, telephone, hands-free, and handicapped applications, to name just a few. The aim of automatic speech recognition has been to design and implement systems capable of transcribing speech. There have been great advances in automatic speech recognition (ASR) technology since its inception in the 1950s, especially in the last thirty years. In a recent DARPA evaluation, for example, a word error rate of 6% was achieved on a speaker-independent unlimited vocabulary read-speech task [130]. Although impressive, the state of the art in ASR is nowhere close to human speech recognition capabilities [82, 112, 23]. The error rates of automatic speech recognizers are one to two orders of magnitude higher than those of humans for many speech recognition tasks, ranging from a 10-word digit recognition task to a 65,000-word spontaneous continuous speech recognition task. ASR systems are particularly poor in recognizing spontaneous continuous speech, as the error rate on the best system for such a task is roughly 37% and around 41% if the speech samples are from conversations in which the parties are familiar with one another [85]. Variables such as speaking style, speaking rate, accent, variable vocal effort, background noise, room reverberation, and channel effects degrade ASR accuracy dramatically, whereas these factors affect human speech recognition much less [93, 79, 69]. Clearly, ASR is not a solved problem, and with the recent debates on the incremental nature of advancement in the field [9], new and exploratory paradigms are most welcome.

1.1 What is Multi-Band Processing?

Recently there has been much interest generated in the ASR community on the topic of the multi-band paradigm, mainly by Allen's [1] cogent retelling of Fletcher's [37] psycho-acoustic studies, conducted in the 1920s. The multi-band ASR method is a special case of the multi-stream paradigm. The goal of the multi-stream model is the incorporation of different information streams [14], for example, streams of audio and visual information, or different sets of features derived from speech data. In the multi-band paradigm each frequency region is considered a separate source.

What distinguishes the multi-band paradigm from the full-band (or traditional) ASR

paradigm is that in the former, the full frequency band is divided into multiple regions (either overlapping or non-overlapping). Feature extraction and phonetic probability estimation are performed on each sub-band independently, and the streams of information are combined and synchronized at anchor-points (e.g., frame, phone, syllable endings). In contrast, in a traditional full-band approach, features are derived from the full frequency region, and only one speech recognizer is trained on these data.

Figure 1.1 shows an overview of the major components of a four-band multi-band system.

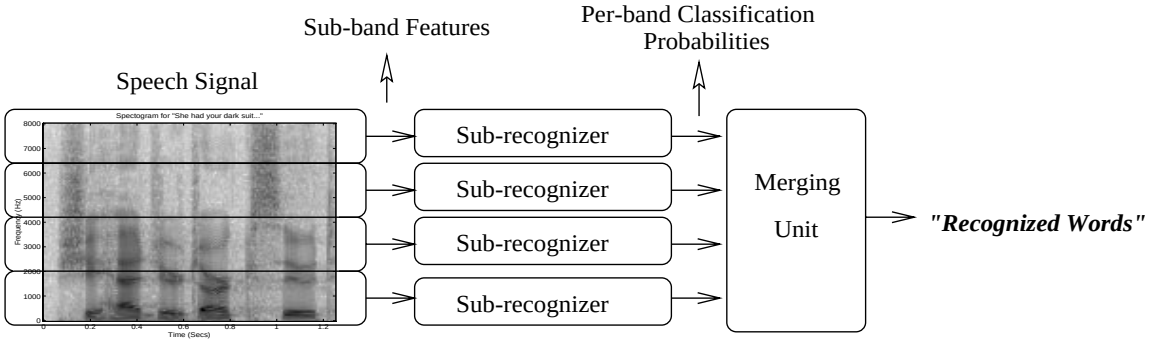

Figure 1.1: A simple overview of the multi-band paradigm.

1.2 Motivation for the Multi-Band Paradigm

The multi-band paradigm could be thought of as simply another approach to ensemble learning [62, 63, 67, 34], where each of the multiple experts are trained on one segment of the data. By extension, the advantages of ensemble learning could be offered as advantages for the multi-band paradigm as well, such as the potential for improved performance and the ability to take advantage of parallel machines. However, there are many speech-specific reasons why multi-band processing is attractive. One category of motivations is based on the signal processing and robustness advantages. The nature of speech and hearing also brings some insight into the utility of this paradigm.

The signal processing and robustness motivations for multi-band processing include:

1. Rao and Pearlman [116] have proven theoretically that the prediction error variance of the full-band always exceeds the total prediction error variance of the combined sub-bands, for a given prediction order p. They proved that the pth-order entropy of the combined sub-bands is closer to the entropy rate of the source than the pth-order entropy of the full-band, for any finite p, where p is the size of the block of source samples. Their proofs provide means of quantifying the sub-band advantages in linear prediction, optimal coding, and DPCM[1] coding in the form of gain formulas. They prove, and shown with simulations, that auto-regressive spectral estimation from sub-bands offers a gain over full-band auto-regressive spectral estimation. Furthermore,

[1] Differential Pulse Code Modulation is a scheme where the difference between a data sample and linear prediction of this sample from past sample values is encoded.

they show that the pth-order linear prediction from sub-bands is superior to pth-order prediction in the full-band, when p is finite.

2. If there are different levels of signal-to-noise ratio (SNR) per band, this technique can increase robustness to noise. For example, imagine a telephone application where the training telephone speech data are noise-free, but the testing data have noise in some selected band(s). If full-band features are extracted from the noisy speech, the full vector would be corrupted, whereas in the multi-band processing case, only the features pertaining to the noisy channel would be corrupted. The multi-band paradigm is more robust to such noise [12, 54], showing a more graceful degradation in recognition performance.

3. Room reverberation causes frequency-dependent smearing of energies in the speech signal, such that typically the high frequencies get smeared less and the low frequencies get smeared more [57]. An intuitive explanation is that the high frequencies get absorbed by the air and most wall materials more readily than low frequency energies, whereas low frequency energy reflects to a greater degree and dissipates more gradually. If each band is processed independently, feature extraction strategies can be tuned to the altered characteristics of speech in each band.

4. The multi-band approach could permit us to have variable-sized bands, the width and frequency placement of which may be adapted to the vocal tract parameters of the speaker. The formants of male speakers, for example, are lower in frequency range than that of female speakers. The multi-band method could be used to perform vocal tract normalization.

The motivations related to the nature of the speech signal and motivated by human hearing are as follows:

1. Some evidence suggests that human speech perception is based on narrow frequency channel analysis and that the recombination of these features is performed at higher processing levels [37, 1, 89]. These psycho-acoustic motivations are discussed in Section 2.2.2.

2. Researchers have hypothesized [13, 135, 138] that phone transitions occur at different times in different bands. Theoretically, the reasons may be[2]:

   - Considering amplitude modulations, the closed vocal tract constitutes a low-pass filter; thus as it opens from a consonant (or mirror image as it closes for a consonant) the frequencies that pass will be progressively higher. For example, in a transition from an unvoiced fricative to a labial stop, /s/ $\rightarrow$ /p/ in *spot*, as the lips close, we see a down-sweep in the middle of /s/. Higher-frequency transitions occur earlier, whereas lower-frequency transitions occur later.
   - Formant transitions come about due to vocal tract constrictions at or near nodes or anti-nodes in the standing waves of the resonant frequencies. The closer

[2]Thanks to Professor John Ohala for these elaborations.

a constriction is to one of these (anti-)nodes the more it influences (lowers or raises) a given formant. Given constrictions, e.g., apical, are right on a node for the second formant but slightly displaced for a node on other formants. Thus the rate of change in the formants will be different as a function of the location of the constriction.

The example in Figure 1.2 from [30] demonstrates that, for example, in the phone transition /g/ to /a/, the transition occurs later in the second formant than in the first formant. Multi-band recognition facilitates asynchronous combination of the sub-band frequency information. This has been highlighted as a reason why the multi-band approach may have the potential to overcome fundamental limitations of current HMM-based systems [121]. This supposition will be examined in Section 4.2.

3. It has been suggested [43] that different frequency regions have different dynamic characteristics. By processing each band separately, one has the option of both developing feature extraction methods, and employing variable-sized temporal windows, tuned to the dynamic characteristics of each frequency region.

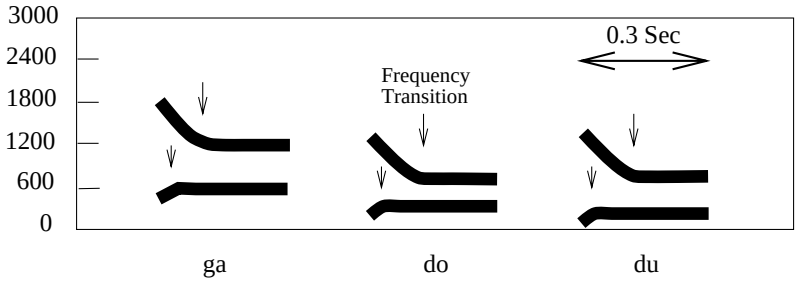

Figure 1.2: Synthetic spectrograms using only F1 and F2 information that produced the voiced stops before various vowels (from [30]).

4. Bilmes' [5] calculation of the time-frequency information density[3] for conversational speech (see Figure 1.3) shows that most of the mutual information is within a sub-band, suggesting that extracting information in a narrow-band region is justified in terms of the concentration of the information content.

5. Redundant information on the identity of the spoken phone is distributed across the frequency band. This redundancy is explicitly represented in the sub-band extracted features, whereas it is only implicitly contained in the full-band features [131]. By keeping a representation of this redundancy on a higher level, we can use the agreement between bands as a confidence measure [126], which may be particularly important for degraded quality speech.

[3]Where information density is defined as the unconditional mutual information, computed using the EM algorithm applied to Gaussian mixture estimates, between pairs of points in the time-frequency spectrum.

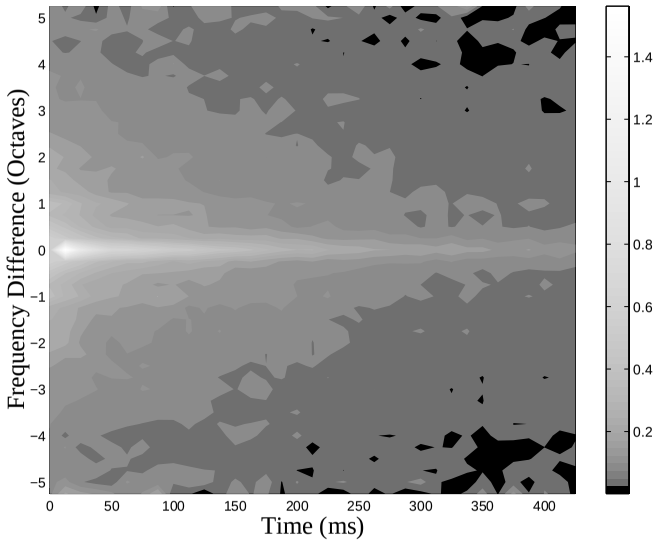

Figure 1.3: The time-frequency information density of a randomly selected 2-hour section of the Switchboard corpus, in bits per unit area (from [5]).

1.3 Thesis Goals and Overview

Multi-band ASR is a relatively unchartered territory. The major thrust of this thesis is to map out some of the terrain. The first goal of this work is to design and implement a multi-band system, exploring various design options in the process. The second goal is to analyze and explore the debated advantages and disadvantages of this paradigm with the aid of confusion matrices and other analysis tools. The third goal is to explore some extensions of the multi-band paradigm, such as deriving reduced sub-band classes and combining the multi-band with a full-band system.

The organization of this thesis is as follows: Chapter 2 contains background information. This includes a discussion of previous work related to multi-band processing and a brief overview of speech recognition technology. Chapter 3 discusses the experimental paradigm and reports the experiments in the design and implementation of a multi-band baseline system. Chapter 4 addresses the following two concerns: (1) whether the multi-band approach is not as good as the full-band approach due to its divide-and-conquer design and (2) whether transitions in different bands occur asynchronously, as has been suggested by many researchers as a potential advantage of multi-band processing. The results of this inquiry motivated the examination of asynchronous merging in the multi-band paradigm, as presented in Chapter 5. Motivated by the observation that some bands perform poorly in discriminating among some phones, Chapter 6 examines merging these oft-confused phone classes in order to derive reduced sub-band super-classes. It has been observed that combining systems with different and complementary characteristics improves word recognition. Combining the full- and multi-band streams is examined in Chapter 7, and an analysis of the two systems is performed. The epilogue (Chapter 8) includes summary, conclusions, and future work. The appendices include supplementary material, as well as a summary

of Allen's article "How Do Humans Process and Recognize Speech," which provided the original motivation for this dissertation.

Chapter 2

Background

This chapter offers a review of the necessary background for this thesis. Section 2.1 is an overview of ASR technology and assumes that the reader is already familiar with the basics of the field. Previous work related to the multi-band approach is discussed in Section 2.2.

2.1 An Overview of Automatic Speech Recognition

Arguably, the first simplified speech recognition system was *Radio Rex*, a celluloid toy dog which would bounce up when its name was called, built in the 1920s [111]. The first "real" speech recognizer came along in 1952 at Bell Labs, built by Davis, Biddulph, and Balashek [29]. This system was a speaker-dependent isolated digit recognizer that achieved a 98% accuracy rate (as long as the speaker did not move his head). ASR research has advanced greatly since then, as can be traced by the difficulty and complexity of the yearly Defense Advanced Research Project Agency (DARPA) benchmarks. In 1971, the evaluation task was a 1000-word vocabulary task, spoken by a handful of speakers [72]. In the 1980s, this benchmark was replaced by Resource Management, another 1000-word vocabulary task which, this time, was collected to cover many different dialect regions and included over a 100 speakers [113]. The Wall Street Journal Corpus (which in 1994 became part of the larger North American Business News (NAB) corpus), a 20,000-word vocabulary task, was used in the 1993 DARPA evaluation [105], at the same time as Switchboard, a 26,000-word vocabulary, spontaneous, telephone dialogue corpus was being collected [25, 44]. The latest DARPA evaluation task, as of 1998, is Broadcast News [45], which is a corpus of speech derived from radio and television news programs, with many challenging properties such as background noise, degraded acoustics, and speaker accents. As may be clear from the development of the test corpora over time, recent ASR systems are capable of recognizing larger vocabularies and have increased robustness to environmental and speaker variation. For tasks such as NAB, word error rates between 7% and 8% are not uncommon using context-dependent phone-level hidden Markov models (HMMs), N-gram language models, and large quantities of training data for vocabularies of up to 65,000 words [149]. Although the algorithms have improved over the years, it is important to note that some commentators attribute the recent advances more to improvements in hardware price-to-performance ratio rather than to breakthroughs in algorithms [24].

SD Baseline	1.5%
SI Baseline	3.0%
Channel	12.0%
Transducer	10.0%
Speaking Rate	15.0%
Noise	30.0%
Dialect	20.0%
Non-Native Speaker	45.0%
Noise + Non-Nativeness	85.0%
Combining All Effects	98.0%

Table 2.1: The word error rates of the system increase dramatically as various field conditions are added to the test data. SD means speaker-dependent and SI means speaker-independent recognition condition (from [41]).

Speech recognition systems have also found their way into the marketplace and into the hands of the average user. Systems with limited capabilities and vocabularies both generate revenue and save costs for companies [115]. Among such limited vocabulary applications are the AT&T Universal Card customer service system and the AT&T operator assistance. The former recognizes spoken continuous digits (credit card numbers) and the latter prompts the user for a choice of keywords ("collect," "operator-assistance," "person-to-person," etc.). There are also command and control systems, such as Wildfire, a continuous-speech engine that works over phone lines and performs as an automated assistant, answering the phone, transferring calls, and providing customized information for callers. Another class of systems aim at recognizing the underlying message rather than identifying the words. AT&T's "How May I Help You" and the AT&T Maxwell System are among such systems. Last, the systems receiving the most attention in the media are the dictation applications, such as Dragon's Naturally Speaking and IBM's ViaVoice. These recent products are continuous-word recognition systems, a substantial improvement over their predecessors which required pauses between every word (discrete-word recognition). These systems have a large vocabulary and the user can also add new words. The users need to wear close-range microphones and train the system individually.

Even though the ASR systems of today are vastly superior to those of even a decade ago, they do not compare favorably with human speech recognition capabilities[1] [82, 112, 23]. As mentioned earlier, the error rates of automatic speech recognizers are one to two orders of magnitude higher than those of humans for many speech recognition conditions, ranging from a 10-word digit recognition task to a 65,000-word spontaneous speech recognition corpus. ASR systems are also much more susceptible to variables such as speaking style, speaking rate, accent, variable vocal effort, background noise, room reverberation, and channel effects [93, 79, 69]. For example, the 3% word error rate on the Resource

[1]Over the past 50 years, various researchers have predicted that the solution to the speech recognition problem is only 5-10 years away.

Management test set increased to 98% when the test data were altered to include some common field conditions (such as channel differences, fast speaking rate, different dialects, and added noise) [41] (See Table 2.1).

As previously mentioned, some have argued that the improvement in ASR technology has become mainly incremental in nature and that new and exploratory paradigms are needed to catapult the state of ASR research out of its current local minimum [9]. It is in such a context that multi-band processing appears as a particularly attractive alternative.

2.1.1 A Brief Review of the ASR Problem

Excellent overviews of the formal ASR problem may be found in any of the following tutorial articles or books [97, 98, 15, 31, 66, 114, 77]. The main problem in automatic speech *recognition* is to find the most likely string of words w^* in a language $\mathcal{L}$ given some acoustic observations X:

$$w^* = \underset{w \in \mathcal{L}}{argmax}\ P(w|X) \tag{2.1}$$

Using Bayes' theorem, we can write:

$$P(w \mid X) = \frac{P(X \mid w)P(w)}{P(X)} \tag{2.2}$$

Since $P(X)$, the *a priori* probability of the acoustics over all word strings, is constant, we need only solve the maximization:

$$w^* = \underset{w \in \mathcal{L}}{argmax}\ P(w)P(X|w) \tag{2.3}$$

$P(w)$ is the prior probability of the word string w, which can be calculated using the language model. The language model parameters are usually estimated from large text corpora or from a finite-state automaton from which N-grams are extracted. An N-gram is the probability of a word given N-1 preceding words. Often, only bigram or trigram models are used. In the equation above, $P(X|w)$ is the likelihood of the word string being pronounced in a particular way. Therefore, in order to calculate $P(X|w)$, we need to know about the pronunciation of the words. $P(X|w)$ is often estimated using the popular forward algorithm [3, 4], or approximated using a Viterbi calculation (which leads to the most likely state sequence).

2.1.2 ICSI's HMM/ANN Hybrid System

This work was performed within the established framework of ICSI's hybrid hidden Markov model/artificial neural network (HMM/ANN) speech recognition system [15]. The main components of this speech recognizer are highlighted in Figure 2.1. The first component is the signal-processing element, where each frame of speech (e.g., each 20 ms segment, overlapped every 10 ms) is processed and relevant speech features (e.g., spectral formants, energy) are derived and non-relevant features (e.g., voice-quality parameters) are de-emphasized. RASTA-PLP processing [52] is the input representation typically used and it is based on an earlier signal processing technique known as PLP [51]. PLP uses three concepts from the psychophysics of hearing to derive an estimate of the auditory

spectrum: (1) the critical-band spectral resolution, (2) the equal-loudness curve and (3) the intensity-loudness power law. The auditory spectrum is then approximated by an autoregressive all-pole model. By the use of simple transformations, RASTA-PLP processing confers robustness to channel variations and acoustic environment changes by suppressing the components of the input spectral trajectories that change either more slowly or quickly than speech. RASTA-PLP processing tends to produce roughly similar word accuracies in a clean environment and significantly improves the recognition accuracy in speech with changed channel characteristics.

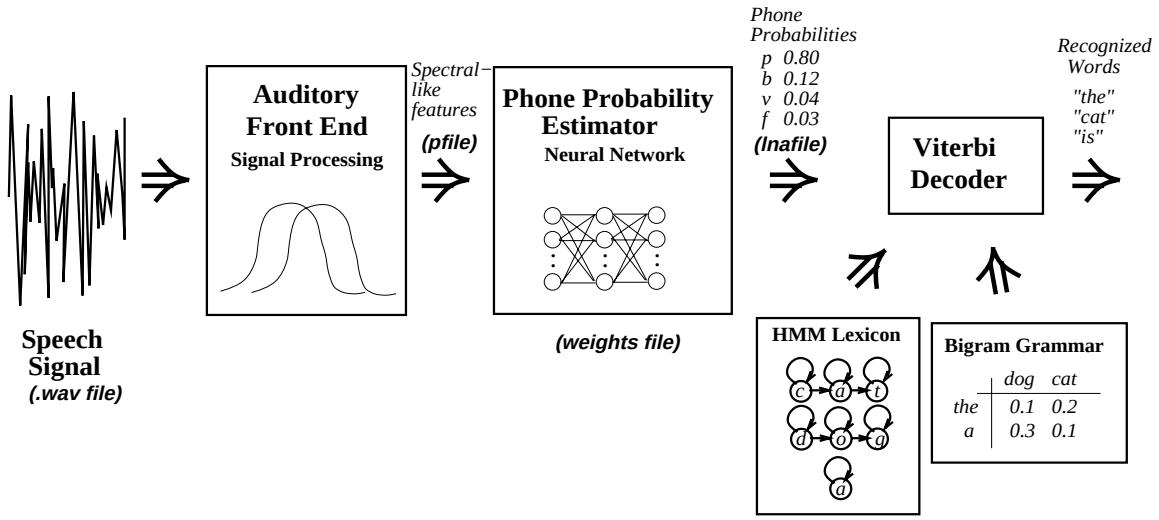

Figure 2.1: The ICSI hybrid HMM/ANN speech recognition architecture.

The next element in this system is the phonetic probability estimator. This is a fully connected multi-layer perceptron trained using the back-propagation algorithm [119] with softmax normalization [19] on the output layer and a relative entropy error criterion [129]. It is used to estimate the probability of each phoneme corresponding to (multiple) frames of speech (Figure 2.2), or $P(q|X)$, where q is a phone in the phone class, and X is an acoustic observation sequence. Next, the phonetic probabilities, along with a grammar[2] and a lexicon[3], are used in a dynamic-programming-based Viterbi search [141], a simplified version of the forward algorithm [3, 4], to find the best strings of words corresponding to the acoustic data.

An HMM/MLP hybrid system is not necessary for building a multi-band-based recognition system, as the systems reported in [35, 108, 138, 87, 22] have all been traditional HMM-based systems. An HMM/MLP hybrid framework, however, simplifies the setup of a multi-band system. Merging the sub-band probabilities on a frame-by-frame basis, for example, is simply performed by multiplying the probability streams from the sub-band MLPs and later feeding the resulting stream into the decoder. By comparison, in a traditional HMM system, multiple speech classifiers are applied to each sub-band, each providing a set

[2]A bigram grammar is often used for small tasks. Such a grammar is specified by a list of words that can follow a specific word, along with associated probabilities.

[3]A pronunciation lexicon is made up of the phonetic pronunciation of each word in an HMM format.

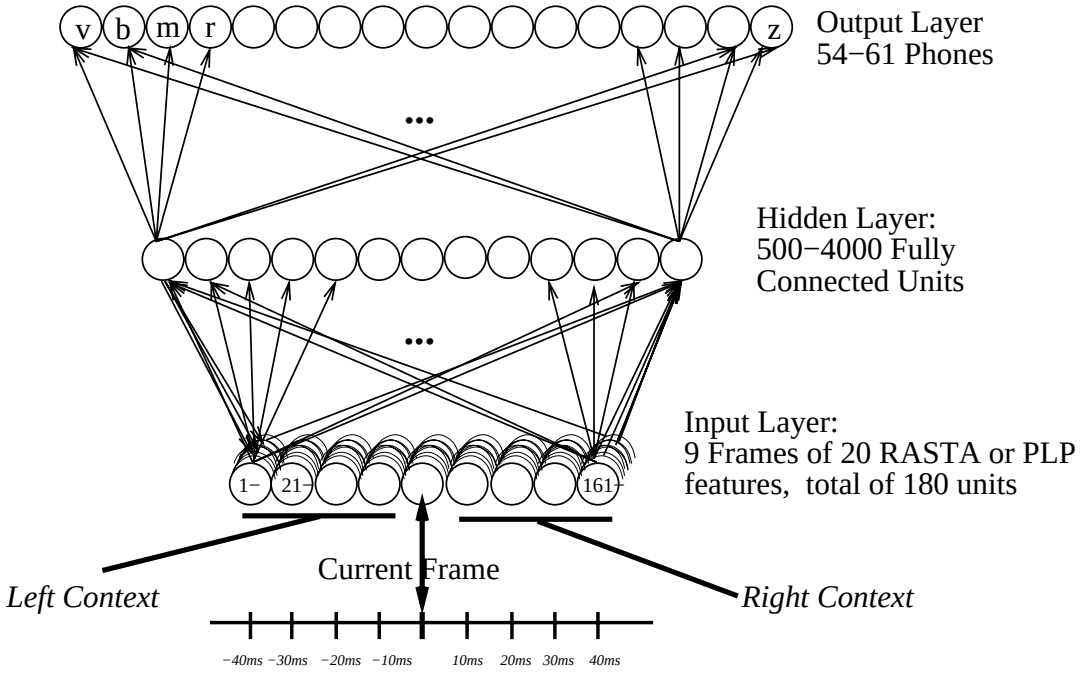

Figure 2.2: An example of a multi-layer perceptron (MLP) phonetic likelihood estimator used in ICSI's speech recognition system.

of recognition hypotheses and scores. The recognition scores are then combined to obtain a global score and recognition decision.

2.2 Related Work

This section is devoted to a discussion of the previous work related to the multi-band approach. Section 2.2.1 summarizes the previous work performed on multi-band by various laboratories, including my earlier work and those of my collaborators. Section 2.2.2 discusses psycho-acoustic work that motivated this thesis. In Section 2.2.3 relevant work on the articulatory features is summarized. In Section 2.2.4 work on missing feature theory, which further motivates work on multi-band, is presented. Finally, Section 2.2.5 summarizes work in other areas that process the data in multiple frequency bands.

2.2.1 Multi-Band ASR

The first work published on multi-band ASR is likely to have been Paul Duchnowski's Ph.D. thesis [35]. His goal was to apply multi-band processing to the task of speaker-independent phone recognition as a cueing aid for the deaf. The focus of his experiments was on comparing different acoustic feature sets (LPC, cepstrum, autocorrelation, and their associated delta features), parameter grouping, quantization methods, and frame-by-frame merging techniques.

Using TIMIT phone recognition as a test-bed, he divided the frequency spectrum into

four non-overlapping bands [100-700 Hz], [700-1500 Hz], [1500-3000 Hz], [3000-4500 Hz] loosely corresponding to the first four formants ([F1], [F2], [F2,F3], [F4]). Four sub-band systems, each based on context-independent three-state phoneme models, were trained. During recognition, the winning phone in each sub-band was chosen. The phone decisions of the sub-bands were then integrated on a frame-by-frame basis according to:

$$\hat{q} = \underset{q \in \mathcal{Q}}{argmax} \; p(q|c_1 c_2 c_3 c_4)$$

where $\hat{q}$ is the collective decision of all sub-bands, and c_i is the frame classification according to the ith channel. Using Bayes' rule, the above may be written as:

$$\hat{q} = \underset{q \in \mathcal{Q}}{argmax} \; p(c_1 c_2 c_3 c_4|q)p(q)$$

Making a simplifying assumption about the independence of the four sub-bands, the first term on the right-hand side may be approximated by multiplying the sub-band probabilities:

$$p(c_1 c_2 c_3 c_4|q) \approx \prod_{i=1}^{4} p(c_i|q)$$

and in turn, using the training set data, $p(c_i|q)$ may be approximated by:

$$p(c_i|q) = \frac{x_{c_i}^q}{N_q}$$

where $x_{c_i}^q$ is the number of frames, the sub-recognizer output label is c_i when phone q is spoken, and N_q is the total number of frames of phone q present in the training data.

Note that "sub-band probabilities" are differently defined in Duchnowski's work than that presented here. To reiterate, sub-band probabilities in Duchnowski's work refer to probability distributions generated from the training set for the co-occurrence of every c_i and q pair. In the present work, the sub-band probabilities, as discussed in Section 2.1.2, are $P(q|X)$, the probability that phone q was uttered given the acoustic observation sequence X.

The frame-by-frame merging of the probabilities was followed by a final "clean-up" stage, consisting of a separate HMM engine to enforce a phone-constancy requirement and avoid frequent label switches. Duchnowski observed that cepstral parameters provided the best performance, followed closely by LPC and autocorrelation. The simple multiplication of probabilities worked as well as more complicated merging schemes (e.g., occurrence weighted interpolation). A phoneme recognition accuracy of 58.5% using an alphabet size of 39 for TIMIT was achieved. This performance was within the range obtained by established phonetic recognizers [78, 59, 150].

More recently, work at ICSI [91] and by our collaborators Bourlard and Dupont [12, 13] and Hermansky and Tibrewala [54, 135, 134] has focused on multi-band for continuous speech recognition. Comparable (or, in a few cases, better) performance for normal speech and superior performance for band-limited noisy speech were demonstrated. The results are briefly summarized below.

The goal of my first set of experiments (in 1995) [91] was to establish a proof of concept for word (as opposed to phone) recognition by developing a two-band system. The features

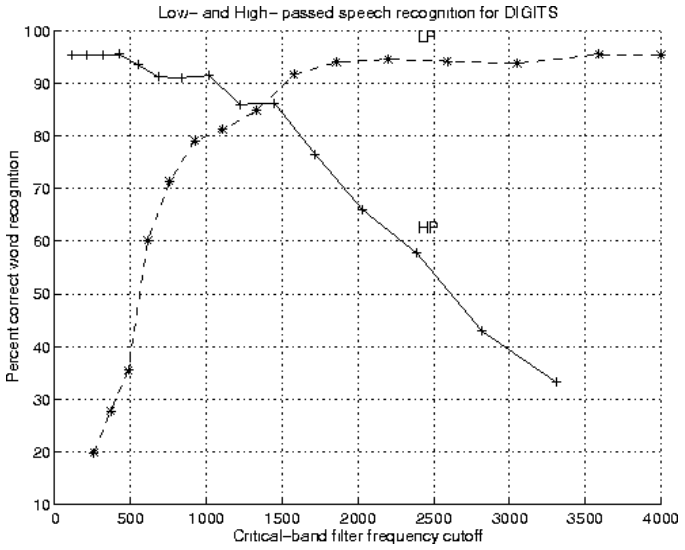

Figure 2.3: Word recognition results for 30 low- and high-pass speech recognizers on the DIGITS corpus. Each system (represented by a point in the graph) has been trained and tested on low- or high-pass speech.

chosen were a vector of 15 power-spectral values, derived from a cascade of PLP critical-band filter analysis, cube-root compression, and equal loudness equalization [51]. To keep the turnaround time of the experiments short, the Bellcore Digits corpus was chosen for testing. This corpus comprises 25 minutes of speech and contains 13 different words: *zero, one, two, ..., nine, oh, yes,* and *no*, spoken by 209 speakers recorded over the telephone (sampled at 8 kHz). To choose the optimal cutoff frequency for the multi-band system[4], 30 separate systems were trained. Each was trained on either a low- or a high-pass filtered set of material. 1400 Hz was the intersection of the low- and high-pass word error curves (Figure 2.3). The intersection point may be thought of as the frequency limit above and below which the same amount of information for speech recognition is available. These curves are similar in shape and point of intersection to the ones reported by Miller and Nicely [89] in their psycho-acoustic experiments.

I trained a two-band and a full-band system using the spectral features described above. The training set consisted of 1720 utterances, cross-validation of 230 utterances, and the final test set had 650 utterances. The full-band MLP (see Figure 2.2) had 135 input, 200 hidden, and 61 output units (40,000 parameters). The low-pass MLP had 63 and the high-pass MLP had 72 input, 200 hidden, and 61 output units (keeping the total number of parameters at 40,000, tantamount to the baseline full-band MLP). The merging was simply performed on the word level by training an MLP on the normalized log likelihoods obtained from the Viterbi decoding distances. More specifically (Figure 2.4), the low-pass and the high-pass probability streams were fed into a Viterbi decoder to produce distances for all competing words. A "merger" MLP was then trained on these likelihood distances to map to the correct word. Note that the merger net was trained on the same data as

[4]This experiment was suggested by Jont Allen.

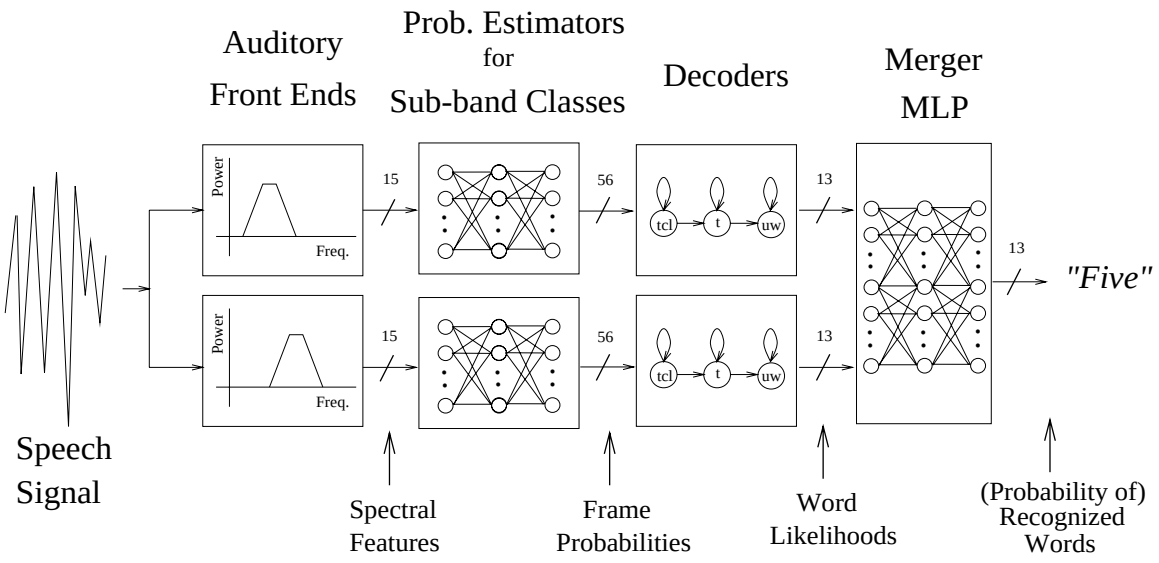

Figure 2.4: The architecture of the initial ICSI multi-band system for experiment with Bellcore Digits.

the classifier. The word error rate of the full-band and the two-band systems were 4.6% and 4.3%, respectively. The improvement was not statistically significant. Clearly, as the number of isolated words in the task increase, this approach to sub-band merging would be difficult to scale up, as the size of the merger MLP input layer would have to increase accordingly. For a similar reason, extension to a continuous word recognition tasks would not be practical.

Hermansky and Tibrewala [54, 133, 135, 134] tested two- and seven-band systems on clean and noisy speech using the Bellcore Digits corpus. They experimented with various forms of corrupted speech: additive 900-Hz sinusoidal noise with SNRs ranging from 30 dB to 0 dB, as well as noise samples from the NOISEX-92 database (factory, destroyer-engine, pink, white, Volvo, babble, and high frequency radio channel). A full-band HMM/MLP recognition system was trained using PLP features extracted from each sub-band and merging was performed on the word level using an MLP trained on the log likelihood distances. They trained 127 MLPs[5] for every possible configuration of n out of seven bands present. To combine these 127 MLPs, they experimented with SNR thresholding (i.e., leaving out the sub-bands which yielded SNR estimates [55] below a certain threshold), majority voting, and adaptation. For noise conditions with local frequency degradations (e.g., additive sinusoidal background, pink noise, and speech babble) the performance of the multi-band system was better than the full-band system. In addition, for low-noise cases, choosing a merging configuration was not necessary, as the seven-band system's performance was as good as the one chosen by the majority voting, SNR-thresholding, etc. This is of note, as it shows the inherent noise robustness of the multi-band approach (as also observed by

[5]The 127 possible configuration of n out of seven bands is calculated by $\binom{7}{1} + \binom{7}{2} + \cdots + \binom{7}{7}$

.

[12]). One does not need to run 127 (or a similar number of MLP merging units) to reap the benefits of noise robustness. This is fortunate because for a large task the required computational power would make this approach infeasible. Another notable result of [135] is that merging on the frame level is as good as merging on the word level, again making their approach scalable to tasks with larger vocabulary sizes.

In experiments associated with the 1996 Johns Hopkins Switchboard Workshop, Tibrewala and Hermansky demonstrated the applicability of the multi-band paradigm to a large vocabulary task. The training set was four hours of male speech (9019 utterances) from the Switchboard corpus and test sentences were 240 male utterances from the development set. The merging network was trained on an independent set of two hours of male speech. Each of the seven frequency channels covered roughly two critical-bands. The MLP probability estimator for each sub-band had 153-207 input, 500 hidden and 56 output units (a total of roughly 760,000 parameters), whereas the baseline full-band system had 234 input, 500 hidden, and 56 output units (145,000 parameters). The per-band likelihoods were merged on the HMM-state level using a non-linear MLP classifier. The word error rate for the full-band baseline system was 60.9%[6], and was 59.0% for the seven-band multi-band system. Tibrewala and Hermansky noted that the 2% improvement on error rate could be attributed to the additional two hours of training data used to train the MLP merger. Nevertheless, it is encouraging that the multi-band system was as good as the full-band system for a large vocabulary continuous speech task, though in fairness, it should be noted that this multi-band system has about five times as many parameters as the baseline full-band system.

The most recent work of Hermansky and Tibrewala [53] uses a one-second temporal vector of critical-band logarithmic spectral energies from each of the 15 critical-band regions. These features are an extreme extension of multi-band features, since they span a very narrow frequency region and a very long time span. Because of the lack of wider frequency context and correlation information, traditionally such extremely narrow-band features have been thought to contain insufficient information about the phone identity. Hermansky and Tibrewala showed that even though independently these features are not as effective for recognition as a more wide-band feature set (full-band PLP [51], in their case), as a supplement to the wide-band features, they improved the performance of the system significantly.

In their experiments Bourlard and Dupont [13, 12] tested critical-band energy features, LPC-cepstra computed on band-limited critical-band values for clean and narrow-band noise, and J-RASTA-PLP features [52] for wide-band noise. They also experimented with the choice of three, four, or six sub-bands. The recombination of the sub-band log likelihoods was performed on either the HMM-state, phone, or syllable level. A multi-layer perceptron (MLP) was used as a merging unit in the frame-level combination experiments, and in syllable- and word-level combination experiments, HMM decomposition [139] was applied in order to force synchronization of the sub-band streams at particular points. They used one of three databases (i.e., a database of German command words, OGI Numbers '93, and Switchboard) for each experiment. Their results showed that for clean speech, the

[6]The error rate for these HMM/ANN systems is particularly high because only four hours of training data are used to shorten the experimental turnaround time. The error rate of the 1996 Johns Hopkins Workshop's HTK system with the same amount of training data was also around 60%.

multi-band paradigm was as good as (in a few instances even better than) the full-band system, and for noisy speech the performance of the multi-band system was superior. Choosing LPC-cepstral features was superior to critical-band-energies (also observed by [135]). Dividing the streams into four narrow bands appeared better than either two or six bands. The experiment on merging level (HMM-state vs. phone, vs. syllable) was inconclusive.

Bourlard et al.'s results on the Switchboard continuous speech corpus [117] are also noteworthy, as they showed applicability of the multi-band paradigm to a large vocabulary task (similar to the work of Tibrewala and Hermansky discussed above). The training and testing sets used in their experiment were similar to the ones used in [133] (as reported above). Each of the four sub-bands had 162-234 input units, 500 hidden units, and 56 output units (a total of roughly 510,000 parameters). The full-band system, to which they compared their results, was made up of four MLPs, each with 342 inputs, 1000 hidden units, and 56 outputs (1,600,000 parameters). They merged the four sub-band log likelihoods linearly using an MLP without a hidden layer. The word error rate on the 240 sentences of the male development set was 61.4% for the four-band system and 63.6% for the full-band system. Note that the number of parameters in the four-band system is one-third of the full-band system's. The best results were obtained, however, by combining the four-band probabilities with the full-band probabilities. The word error rate decreased to 59.7%. Note that in this case, no additional data were used for training the merger MLP.

Tomlinson et al. [138] devised a two-band system with asynchronous-time combination between a high- and a low-pass component of the speech spectrum through a variant of HMM decomposition. Their experiments were performed on a speaker-dependent 500-word ARM (Airborne Reconnaissance Mission) task [120]. They divided the speech into two bands ([<4 kHz][7] and [4 kHz - 8 kHz]) using 25 low-pass and 4 high-pass cosine and one energy term each to describe the lower and upper bands. They allowed asynchronous merging of the streams across the two bands for three-state context-independent HMM phone models and reported a decrease in error rate as compared to the full-band system. However, extending their approach to three bands, as well as relaxing phone-boundary synchronization (by extending the HMM model to allow more model entry and exit states) did not work as well as the more basic system described above. Although it is encouraging that multi-band has been shown to perform as well or better than a full-band system, as mentioned above, the generalizability of the results [138] is uncertain. Furthermore, it is unclear whether the performance gain is purely due to asynchronous merging, since the results for two- and three-band systems are not consistent.

Okawa et al. [108] suggest a feature combination (FC) as opposed to a likelihood combination (LC) approach to multi-band ASR by training only one single system on a vector composed of the features extracted from the sub-bands. They observed better performance for FC compared to LC for the noise conditions they tested. For several noise types such as "babble," "destroyer-engine," and "machine gun" (from the NOISEX-92 database), a multi-band system resulted in higher accuracy. However, the multi-band approach seemed less accurate than the conventional ASR under white- and pink-noise conditions. They also observed modest improvements from weighting the sub-bands likelihoods according to SNR and entropy criteria.

[7]Note that their lower band region alone [<4 kHz] is equal to the full frequency band in this work, since band-passed telephone speech is used.

In [87], McCourt et al. combined various multi-resolution, sub-band feature systems by simply summing the log likelihoods of the systems on a frame-by-frame basis. Their best results were obtained by combining the full-band system with either a two-band or a four-band system. They observed a 2% relative improvement in clean speech and an almost 8% relative improvement in the 15-dB white-noise condition for the TIMIT phone-recognition task.

Cerisara and Haton [22] have developed a multi-band system. Similar to Duchnowski, they tried to merge the *label* of each sub-band, but not the probabilities. In order not to impose synchrony constraints, they grouped the phones and extracted a single phone out of each group. The details of their algorithm are not clear, and they report a "theoretical" final accuracy of 62.5% on a multi-speaker French database known as Aupelf-Uref.

2.2.2 Psycho-Acoustic Studies

The research into multi-band approaches has been first and foremost motivated[8] by the work of Fletcher, as summarized by Allen [37, 1]. The underlying hypothesis of their work is that human speech perception is based on narrow frequency channels. Fletcher conducted human listening experiments using nonsense CVC (consonant-vowel-consonant) sets. Based on the results of the experiments, he proposed a five-layer model for human speech processing. First, the signal enters the cochlea and is broken into critical-bands. Signal-to-noise ratios (SNR_k) are defined for each of the K sub-bands. Next, in each band, features are extracted based on the SNR_k and partial articulation errors e_k are calculated. In the next step, the independent band phone articulation errors are determined as $s = e_1 e_2 \cdots e_k \cdots e_K$. Recognized phones are grouped into CVC syllable units with syllable articulation $S = s^3$. Finally, words are determined with intelligibility $W(A) = 1 - (1 - S(A))^j$, where $W(A)$ is the word articulation, $S(A)$ is the syllable articulation, and the constant $j > 1$ depends on the entropy of the word corpus and may be empirically determined.

Note that in Fletcher's theoretical framework, the merging of multiple-band information is based on the *Articulation Index (AI)*, where the error rate associated with the full band is equal to the product of the band-limited errors. There are two problems with using AI theory for this sort of merging:

1. In order for $s = e_1 e_2 \cdots e_k \cdots e_K$, the system needs to know the relative reliability of the channels. There is no statistical model for the error rate in the full band being equal to the product of the errors in the individual bands, as such a model requires knowledge of the reliability of a band *relative to others*.

2. The work of Kryter [74], Warren et al. [142], Lippmann [81], and Greenberg et al. [46], summarized below, point out the shortcomings of AI theory with respect to human speech perception.

Fletcher's multiple-band model is interesting and warrants simulation and study. It is not clear, however, that using AI theory for the combination of information from the narrow-band channels is correct.

[8] Since [1] was the original motivation for this dissertation, I have included an extended summary of this paper in Appendix A.

Richard Lippmann [81] reports on human perception experiments using low and high frequencies. The common belief has been that high-frequency speech energy above 4 kHz contains relatively little information germane to speech recognition. Lippmann shows that the intelligibility of consonants remains high (roughly 90% correct) when speech energy in the mid frequencies (800 to 4000 Hz) is filtered out of random CVC syllables using sharp high-pass and low-pass filters. He reports 44.3% consonant recognition accuracy when listeners hear only speech low-passed at 800 Hz. Adding a high-frequency pass band above 8 kHz to the low-frequency band increases consonant recognition accuracy by almost 30 percentage points from 44.3% to 73.9%. These results are particularly interesting in that they bring some aspects of the AI theory under question, since one of the most popular methods for calculating AI does not even take speech energy above 6.4 kHz into account [75]. Lippmann further argues that humans use a process for speech recognition that is fundamentally different from the template matching methods most common in HMM ASR systems, pointing out that most recognizers are extremely sensitive to channel variability, filtering, and noise. In contrast, the untrained human subjects achieve high recognition accuracy under highly unnatural conditions.

Warren et al. [142] and Greenberg et al. [46]have both conducted human listening experiments using spectral "slits" (narrow frequency channels) and have observed that the human intelligibility of sentential material far exceeds that predicted by the Articulation Index [60]. For example, in [46] the human recognition accuracy of TIMIT sentences using only two narrow slits in the frequency ranges [750-945 Hz] and [1890-2381 Hz] was as high as 60%.

Another interesting psycho-acoustic study that involves the division of the frequency band is Oded Ghitza's tiling experiments [43] with the Diagnostic Rhyme Test (DRT). He tested the discrimination between various phonetic qualities, such as sibilation (*chair* vs. *care*), voicing (*veal* vs. *feel*), and nasality (*meat* vs. *beat*). He divided up the diphones of the CVCs into 12 tiles: four subsections along the time axis (division on the C$\rightarrow$V, middle of V, and V$\rightarrow$C transitions), and three subsections along the frequency axis (division at 1000 Hz and 2500 Hz). He then interchanged a particular tile from one word with the same tile from another word in the pair. He claims that particular time frequency tiles are responsible for different phone-feature discrimination tasks, for example, voicing and nasality are sensitive to an interchange of the first frequency band, and sibilation to the interchange of the third frequency band of the diphone. Ghitza further argues that there is a direct mapping between phonemic/articulatory features and time-frequency tiles for human perception, and furthermore, that diphone tiles appear to be more important than vowel or consonant tiles. Ghitza's findings serve as motivations for developing specialized phone-like classes for each sub-band, as developed in Chapter 6.

The seminal work of Miller and Nicely [89] on low-frequency and high-frequency masking is also relevant to this thesis. They compared the intelligibility of sixteen consonants in a C-/a/ context in various conditions of low-pass and high-pass filtering and with random masking noise as presented to five listeners (800 syllables in each condition). Their results demonstrated that human speech recognition with limited narrow-band information is not only possible, but surprisingly good (see Table 2.2 for excerpted examples).

They also observed that in the low-pass filtering condition, phonemes were left audible and errors were predictable and similar, whereas high-pass filtering removed most of the

Frequency Band (Hz)	Percent Consonants Correct
200–600	49.5
200–1200	57.2
200–2500	72.8
1000–5000	73.1
2500–5000	38.1
200–5000	83.3

Table 2.2: Percentage of correct consonant recognition from Miller and Nicely 1955, Table XX.

acoustic power, leaving consonants inaudible and consequently causing random confusions. Linguistic features were transmitted differently in the low-pass (LP) and high-pass (HP) conditions: In the LP filtering condition, *voicing, nasality*, and *affrication* features, in descending order, were preserved most clearly. In the HP filtering condition, *duration* was maintained, and all other features degraded as the HP cutoff was increased. From these results, it appears that certain parts of the spectrum specialize in conveying particular linguistic features, supporting the proposal of designing multi-band recognizers that classify sub-band classes.

Note that the Miller and Nicely results were obtained on a set of 16 consonants, and perception of continuous narrow-band speech may be different. For continuous speech recognition, listening experiments on TIMIT sentences conducted at ICSI [2] suggest that band-limited speech recognition is surprisingly good, perhaps due to the presence of contextual information. Clearly, there must be much redundancy in the information content in speech to make recognition of narrow-band speech possible [131].

Finally, French and Steinberg [38] have also performed human speech recognition experiments with nonsense CVCs with high-pass and low-pass speech. The results, summarized in Figure 2.5, re-confirm the ability of humans for narrow-band speech recognition, and further suggest redundancy of information in the speech spectrum. It is interesting to note the similarity between Figure 2.5 and 2.3, the latter of which was obtained from multiple low-pass and high-pass automatic word recognizers.

2.2.3 Articulatory Features

In Chapter 6 an attempt to generate reduced sub-band phone classes is reported. As background, this section discusses previous work on alternative classes for speech recognition.

Deng and Sun[33] describe an ASR system using articulatory features. First, using knowledge about articulatory features, he defines a set of features corresponding to each phone. Then, using phonetically transcribed words and feature overlap rules, he maps out how the features for a word overlap. For each phone in a given context, he generates a state transition graph, and uses the forward algorithm [3, 4] for recognition. The accuracy of his system is higher than traditional context-independent systems trained on small amounts of

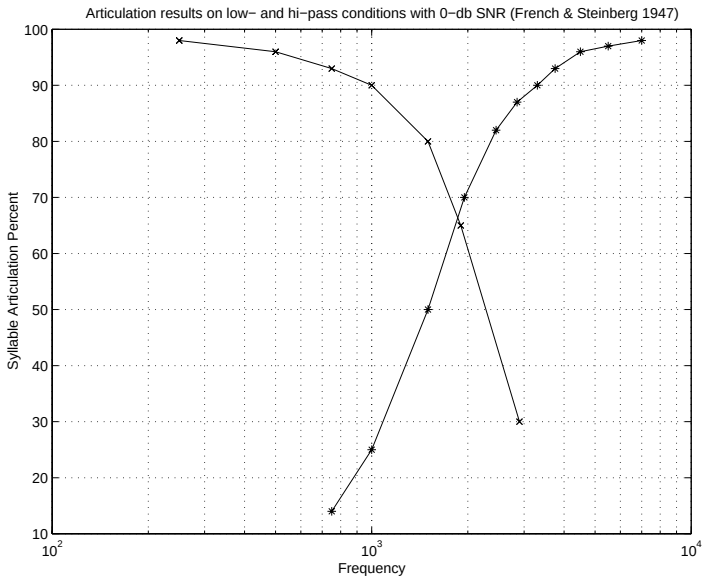

Figure 2.5: Articulation test results for human nonsense CVC recognition on low- and high-pass filters with 0 SNR (from [38]).

data and it performs in the same range as the best systems (as of 1994) for TIMIT[9]. The similarity between their work and this thesis is in the use of an alternative feature set for representing the phones.

Another related study is that of Bitar and Espy-Wilson [6, 7]. They perform broad phonetic category recognition using manner of articulation features (e.g., sonorant, syllabic, frication) derived from acoustic parameters (AP). They process particular frequency regions, looking for specific features (such as the zero-crossing rate, peaks and dips in energy, and auto-correlation coefficients) to determine the acoustic parameters. The accuracy of the AP system is similar to mel-cepstral coefficients and log energy. When training the system on female speech and testing on male speech, the error rates for the AP features do not change significantly (72% vs. 71% percent correct), in contrast to a system trained on mel-frequency cepstral coefficients (72% vs. 67% percent correct).

2.2.4 ASR with Missing Features

Multi-band is another approach for dealing with missing features. The work of Cooke et al. [27] and Lippmann and Carlson [84] have shown that speech recognition with incomplete features may be done using missing-feature theory. The main idea is to ignore the missing pieces of information and compile the best estimate possible for the total likelihood, using the means and variances of the existing data instead. In recognition, the optimal policy is to integrate the required class probabilities over all possible missing values. When all missing values are completely unbounded, this method produces the marginal distribution for data present, which effectively ignores the missing data. Cooke et al. [27] have shown

[9]In 1994, the best HMM systems had 74% accuracy, the best connectionist system had 78% accuracy, and Deng's system was 73% accurate.

that up to 80% of the spectro-temporal regions may be randomly removed without much degradation of the recognition. However, if the neighboring spectral regions are removed, or if the removal is governed by local levels of SNR, the performance deteriorates seriously. Multi-band approaches appear to be more robust in these situations [13, 134].

2.2.5 Other Related Work

A related multi-band work is that of Sankur et al. [124] who applied the approach to respiratory signals for classification of healthy and pathological cases. They decomposed the signal dyadically into M+1 octave bands and extracted a separate cepstral feature vector from each band and formed a time-frequency feature matrix, which they then used in the first stage of a two-stage classifier. The output of the M+1 classifiers were combined using variations on the majority rule. They showed significantly better classification results with a multi-band system.

A somewhat related work to the multi-band approach is that of Sarikaya and Gowdy [125] in which they proposed a new set of feature parameters based on sub-band analysis of the speech signal for classification of speech spoken under stressful conditions. Sub-band-based features are shown to achieve 7.3% and 9.1% increase in the classification rates over the MFCC-based parameters for ungrouped and grouped stress closed vocabulary test scenarios, respectively.

Chapter 3

Designing A Baseline Multi-band System

Building any new recognition system requires a multitude of design decisions. In the design of a multi-band system, some issues are unique to this particular paradigm, whereas others are general to any speech recognition system. Among the issues in the former category are: choosing the sub-band regions, sub-band acoustic features, sub-band probability estimators and, perhaps most importantly, the strategy for combining the multi-band streams. Issues not unique to this paradigm include the choice of the phone set, the recognition units, the lexicon and the grammar. This chapter focuses mainly on the discussion of the design decisions and experiments specific to the multi-band paradigm. The general system choices are the same as those of the ICSI full-band system, as briefly discussed in Section 2.1.2, and will be reiterated and summarized in Section 3.6.

Another point to keep in mind is that it is impossible to make all the design choices in a sequential fashion. An initial set of choices has to be made in order to build a test system. Even though these choices are presented sequentially, the actual process of experimentation was not strictly serial.

3.1 The Experimental Framework

For the experiments reported in this dissertation the OGI Numbers95 corpus [21] was used. The following sections include descriptions of the database and justifications for why it has been chosen.

3.1.1 The Task: Numbers95 Corpus

Numbers95 [21] is a continuous speech corpus recorded over the telephone and sampled at 8 kHz. It was collected and annotated at the Oregon Graduate Institute from census information (phone numbers, birth dates, zip codes, etc.) and includes noise, non-speech sounds, and truncated speech. A portion of the corpus is phonetically hand-transcribed. Table 3.1 summarizes the specifications for this database material. For the experiments in this thesis a 6000-utterance subset (roughly 4.5 hours of speech) of the corpus (the core

	Numbers95 Subset
Number of words	32
Total Utterances	6,000
Duration	4 hrs
Examples	"seventy", "thousand"

Table 3.1: Some features of the Numbers95 subset.

subset[1]) was used. Utterances were chosen to be included in this subset if they fulfilled the following requirements:

1. They are phonetically hand-transcribed.
2. They contain only numbers and digits. Utterances such as "Sears one-day sale" were not included.
3. The entire utterance is intelligible. Since many of the utterances were cut out of a longer context, i.e., callers reciting their addresses, clipped endings were a problem in many utterances.

Table 3.2 lists the vocabulary words in the Numbers95 core subset. A typical sentence from this database is: "six one thirty four." As speech files are fed into the recognition system, the first and last several frames of speech are often ignored due to lack of context. To avoid this problem, 100 ms of artificially created silence was added to the beginning and end of each wavefile.

Two hours of training data (3,500 utterances from 1821 speakers, about 700,000 frames), 40 minutes of development test set (1,206 utterances from 606 speakers, total of 4,673 words, about 230,000 frames), and 40 minutes of evaluation test data (1,227 utterances from 618 speakers, total of 4,757 words, about 230,000 frames) comprise the core subset. The cross-validation data set, which was used as the training stopping criterion, comprised the last 10% of the training set.

3.1.2 Why Numbers95?

Much research in ASR has been done on read speech. Considering the various challenges in ASR, this has been a reasonable choice. However, the structure of read speech is different from fluent and natural speech. Spontaneous speech has posed a major challenge for ASR, particularly in the last few years. Spontaneous speech has been chosen as a test-bed for multi-band ASR for the following reasons:

- Spontaneous speech is quite variable. Since the complexity of the speech material is higher, the pattern recognition task is more difficult. Furthermore, spontaneous speech, as opposed to read speech, is conveyed through a set of efficient minimal cues. It is hoped that by dividing up the space of patterns into smaller units, the ability to

[1]The "core" subset was defined by Mike Shire and Su-Lin Wu at ICSI.

zero	oh	ten
one	eleven	hundred
two	twelve	twenty
three	thirteen	thirty
four	fourteen	forty
five	fifteen	fifty
six	sixteen	sixty
seven	seventeen	seventy
eight	eighteen	eighty
nine	nineteen	ninety
uh	um	

Table 3.2: The vocabulary words in Numbers95 core subset.

find relevant structures in data sub-spaces, and ultimately the generalization power, is increased.

- Since the pattern of minimal cues are manifested in a substantially different way from the template patterns associated with read speech, there is higher potential for spontaneous conversational speech to exhibit a higher degree of temporal asynchrony than read speech. Since one of the potential benefits of the multi-band approach is the asynchronous merging of sub-band data, it may be an appropriate tool for addressing the variabilities in spectro-temporal patterns in spontaneous speech.

Another benefit of the Numbers corpus is its size: it is sufficiently small that the experimental turnaround time is relatively short, and it is large enough to include speech in various contexts and to enable us to see statistically significant differences in various test cases. The task also includes difficult discriminations not present in a digits-recognition task, for example, discrimination between "sixteen" and "sixty" or any two words from the second and third column of rows 4 through 10 in Table 3.2. Additionally, variations such as speaker differences, channel and environmental effects, speaking rate variations, background noise (e.g., crying babies, door slams) are represented in the data. Even though the Numbers task is fairly small, it is non-trivial due to all the variations represented in the samples. One disadvantage of Numbers95 is that it does not offer complete phonetic coverage, i.e., there are no voiced stops (e.g., [b], [d], and [g]) and there is only limited representation of the laterals (e.g., [l] in "eleven").

3.1.3 Reverberation

Some of the experiments reported in this thesis have used reverberant speech. Reverberation is a typical environmental degradation that affects the speech signal. Reverberation is due to sound reflecting off of walls and other solid objects. Its effects can be best heard in large empty rooms. A certain level of reverberation is a desirable property of concert halls and

it enhances the enjoyment of music. Statistically speaking, reverberation is a transient, non-stationary, and relatively slow response of sound in an enclosed environment.[2]

The reflected energy in a reverberant environment masks the energy in the original speech wave and thereby distorts the speech. Recognizing reverberant speech is therefore more difficult for automatic speech recognizers since the acoustic patterns of reverberant speech are different from those of clean speech material used in training the speech recognizer. It is not uncommon to obtain word error rates of one order of magnitude higher for reverberant speech.

For the experiments in this dissertation, mildly reverberant speech produced by Kingsbury and colleagues (at ICSI) [47, 69, 70] have been used. The reverberant data set was generated by convolving the clean set with an impulse response measured in a room having a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 0 dB[3].

It is important to note that, similar to other forms of acoustic degradation, reverberation affects human recognition much less than machine recognition. Kingsbury [68] has performed informal listening experiments with two subjects on 200 sentences of the Numbers development test set. The subjects showed almost perfect recognition of *both* clean and reverberant speech, with an average error rate of 0.3%. Performance of speech recognition systems on such reverberated data is about two orders of magnitude higher.

3.1.4 The Evaluation Criteria

Word error rate is the most common metric used in the speech recognition community for the evaluation of speech recognition systems. It has also been used as the evaluation criterion in this dissertation. Word error rate is calculated using a dynamic programming string alignment algorithm which performs a global minimization of a Levenshtein distance function which weights the cost of correct words, insertions, deletions, and substitutions differently (for example, as 0, 3, 3, and 4, respectively). The computational complexity of the dynamic programming algorithm is $O(N^2)$. Further explanation of this algorithm can be found in [123].

3.1.5 A Note on the Reported Results

Almost all of the experimental systems (both full-band and multi-band) reported in this work have been trained using the hand-transcribed phonetic labels, a decision made at the beginning of the work. Later, it was discovered that many of the phonetic-labels were erroneous and that there were mismatches with the data. Much of the work of the other members of the research group at ICSI [147] has been reported with improved phonetic labels by the use of an embedded alignment procedure, along with the retraining of the lexicon, and the baseline word error rates are typically about 1-2% less. The quality of the phonetic labels affect both the baseline and the experimental systems equally in this work. Therefore, all measured word error improvements (or degradations) are valid as long as they are consistent within the experimental framework. However, Chapter 7 describes a

[2]For further reading on the nature of room reverberation see [137], and for a statistical characterization of reverberation in rooms see [96].

[3]Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, now at the University of California, Davis, collected the room impulse responses.

re-implementation of my best experimental system using improved phonetic labels in order to produce word error results that could be compared to the work of others.

3.2 A Formal View of the Multi-Band Paradigm

This section presents a more formal overview of the multi-band paradigm, as developed by Bourlard et al. [14].

First, a series of definitions. We assume that there are K information streams X_k, and that the hypothesized model for an utterance, M, is composed of J sub-unit models M_j ($j = 1, \ldots, J$), where a sub-unit may be an HMM state, phone, syllable, or word which, in turn, is composed of a sequence of HMM states. The goal is to process each stream independently within the chosen sub-unit and impose synchrony on all sub-bands at the end of each sub-unit. In other words, each sub-unit model M_j is composed of parallel models M_j^k that are forced to recombine their respective segmental scores at some common temporal anchor points. These parallel HMMs, associated with each of the input streams, do not necessarily have the same topology. The recombination state (represented as a black circle in Figure 3.1) is not a regular HMM state since it will be responsible for recombining probabilities (or likelihoods) accumulated over a temporal segment for all streams.

Note that forcing asynchrony at every HMM state is equivalent to simply merging on a frame-by-frame basis. In order to merge the streams on any level higher than the state level, some form of a two-pass algorithm must be used. HMM decomposition (or recombination) [139, 42] and two-level dynamic programming are candidates [122]. These algorithms are discussed in more detail in Chapter 5.

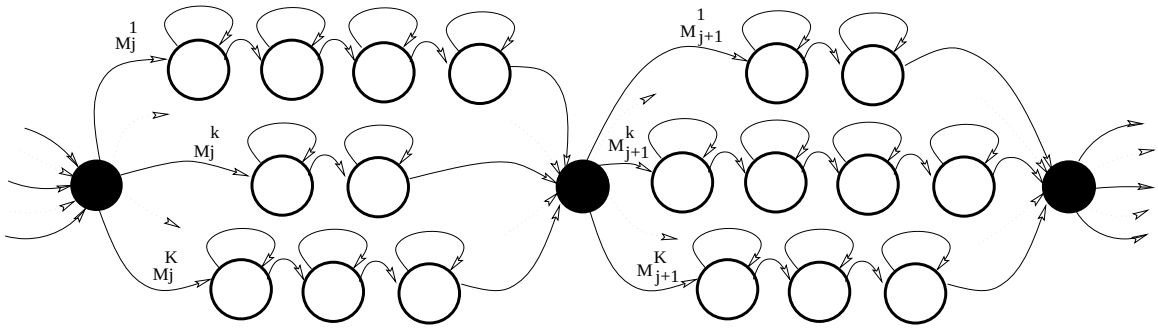

Figure 3.1: The general form of a K-stream recognizer. The black circles between the speech are anchor points with which synchrony is imposed among the streams.

The recognition problem may be formulated either for a posterior-based or a likelihood-based system. The likelihood-based formulation is presented here.

We need to solve the following model M to maximize:

$$M^* = \underset{M \in \mathcal{M}}{argmax} \; P(M|X) = \underset{M \in \mathcal{M}}{argmax} \; P(X|M)P(M)$$

Ignoring $P(M)$, we can focus on acoustic likelihoods $p(X|M)$:

$$p(X|M) = \prod_{j=1}^{J} p(X_j|M_j) \tag{3.1}$$

where $p(X_j|M_j)$ is the probability that sub-unit model M_j produced the observed multiple stream sub-sequence X_j. Assuming that we have different experts E_k for each input stream X^k and that these experts are mutually exclusive and collectively exhaustive, we have:

$$\sum_{k=1}^{K} P(E_k) = 1 \tag{3.2}$$

where $P(E_k)$ is the probability that expert E_k is better than any other expert. We can write Equation 3.1 as:

$$p(X|M) = \prod_{j=1}^{J} \sum_{k=1}^{K} p(X_j^k|M_j^k) P(E_k|M_j) \tag{3.3}$$

where $P(E_k|M_j)$ represents the reliability of expert E_k given the sub-unit M_j.

If we assume that the streams are statistically independent, we can eliminate the estimate of the expert reliability and decompose the likelihood into a product of stream likelihoods for each segment model, simply computing:

$$\log p(X|M) = \sum_{j=1}^{J} \sum_{k=1}^{K} \log p(X_j^k|M_j^k) \tag{3.4}$$

This approach can be generalized to a weighted log likelihood approach and allow bands to be weighted differently, according to their reliability:

$$\log p(X|M) = \sum_{j=1}^{J} \sum_{k=1}^{K} w_j^k \log p(X_j^k|M_j^k) \tag{3.5}$$

And more generally, we may choose any non-linear function (e.g., an MLP) to combine the sub-band probabilities and relax the independence assumption:

$$\log p(X|M) = \sum_{j=1}^{J} f(W, \log p(X_j^k|M_j^k), \forall k) \tag{3.6}$$

where W is a global set of recombination parameters, which may include weighting factors for each sub-band.

3.3 Design Parameters

This section delineates the set of design choices and experimental parameters of concern. In Section 3.4 the details of the experiments are discussed.

3.3.1 Sub-Band Frequency Regions

What is the optimal number of sub-bands and what should the cutoffs be? The answers to these questions are fundamental to the multi-band system's efficacy.

One study relevant to the sub-band boundary decision is that of Houtgast and Verhave [58]. They performed experiments to determine the amount of information overlap in the speech signal and observed that there is more information overlap for higher frequencies (1-4 kHz) on a linear scale. On a logarithmic scale, however, there is less overlap between channels in the higher frequency range, suggesting that we should have more bands in the 1-4 kHz region because of the higher information content. Recall that in this dissertation telephone-quality speech (low-passed below 3.4 kHz) has been used, so dedicating fewer bands above 4 kHz is not an issue.

One concern in choosing the number and width of the sub-band regions is that the smaller the frequency region a sub-band covers, the higher the recognition error rate for that particular sub-band. Although a larger number of narrower sub-bands may better isolate the effects of local narrow-band noise, their lower local accuracy may lead to overall poorer results.

Duchnowski [35] chose four non-overlapping bands [100-700 Hz], [700-1500 Hz], [1500-3000 Hz], [3000-4500 Hz] roughly corresponding to the first four formants ([F1], [F2], [F2,F3], [F4]). Bourlard and Dupont [12] experimented with three, four, and six sub-bands, and observed that four (or perhaps five) was the optimal number of bands for their experiments. They used [17-778 Hz], [707-1631 Hz], [1506-2709 Hz], and [2121-3769 Hz] for their four-band experiment, grouping critical-bands [1-6], [7-10], [11-13], and [13-15]. This division is similar to what Shannon used in his psycho-acoustic experiments on amplitude-modulated noise in four frequency bands [128]. Tibrewala and Hermansky [135] observed that the error rates for a two- and four-band system were lower than that of a seven-band system. They also experimented with fifteen narrow-band long-term features, which proved helpful in combination with full-band spectral features [53].

Based on previous experiments and psycho-acoustic observations, four sub-bands, with cutoffs based on the RASTA-PLP critical-band filters (see Tables 3.3 and 3.4), have been chosen for the work presented here. The first sub-band includes the outputs from four critical-band filters, three to six, inclusive. Note that the first two critical-bands have been excluded. My pilot experiments (not reported here) showed that the first two bands did not help recognition for the task at hand, since telephone speech is typically high-passed above roughly 300 Hz. Further experiments by Tibrewala and Hermansky confirmed this observation [135].

Like the first sub-band, the second sub-band includes four critical-band filters. The last two chosen sub-bands each include three critical-band filters. Although the higher frequency sub-bands include fewer critical-band filters, they span a wider frequency range, because the critical-band filters are logarithmically spread. Each critical-band filter has been included in only one of the four sub-bands, except for filter number 13, which has been included both in the third and fourth sub-bands. Early pilot experiments showed the performance of the last sub-band to be quite poor with only two critical-band filters (14 and 15). Therefore, the last sub-band was expanded to include filter 13. Except for this overlap, the only overlap among the sub-bands is due to the range of the critical-band filters.

RASTA-PLP Critical-band Filter Frequency Cutoffs			
Filter Number	Low Cutoff	Mid Frequency	High Cutoff
1	17.24	97.77	161.27
2	115.28	198.12	264.64
3	216.36	303.70	374.99
4	323.15	417.29	495.23
5	438.47	541.89	628.53
6	565.35	680.78	778.42
7	707.14	837.63	948.84
8	867.59	1016.58	1144.29
9	1050.92	1222.34	1369.93
10	1261.98	1460.35	1631.70
11	1506.32	1736.88	1936.52
12	1790.41	2059.23	2292.43
13	2121.72	2435.90	2708.80
14	2509.01	2876.83	3196.64
15	2962.48	3393.65	3768.80

Table 3.3: The half-power low and high frequency cutoffs for the RASTA-PLP filters when the sampling frequency is 8 kHz.

Chosen Sub-Band Frequency Cutoffs			
Sub-Band Number	Filters Included	Low Cutoff	High Cutoff
1	3 – 6	216.36	778.42
2	7 – 10	707.14	1631.70
3	11 – 13	1506.32	2708.80
4	13 – 15	2121.72	3768.80

Table 3.4: The half-power low and high frequency cutoffs for the chosen four sub-bands. The sampling frequency is 8 kHz.

In terms of spectral information, the first sub-band [216-778 Hz] roughly corresponds to the range of the first formant[4]. The second [707-1631 Hz] and third [1506-2709 Hz] sub-bands together span the range of the second and third formants. The fourth sub-band [2121-3769 Hz], contains information on the higher formants.

In summary, four sub-bands were chosen for the experiments in this thesis, with frequency cutoffs of [216-778 Hz], [707-1631 Hz], [1506-2709 Hz], and [2121-3769 Hz].

3.3.2 Acoustic Features

Various types of acoustic features have been used in multi-band experiments, such as PLP, RASTA-PLP, LPC, MFCC, multi-resolution sub-band cepstral, and critical-band energies, to name but a few [35, 132, 13, 87, 108, 138]. Most recently, TRAPS (TempoRAl Patterns of Spectral energies) [53], as well as spectral sub-band centroids [109], have been proposed.

In pilot experiments motivating this work [91], power spectrum values obtained after PLP critical-band filter analysis, cube-root compression, and equal loudness compensation [51] were chosen. These critical-band features did not perform as well as the LPC-cepstra computed on band-limited critical-band values (also observed by [12, 135]). For a full-band system, RASTA-PLP [52] features perform almost the same as PLP features for clean speech and significantly better than PLP features for reverberant speech on the Numbers95 database. Since the experimental task, Numbers95, has been recorded over the telephone for thousands of speakers, the channel and acoustic environments of speech segments varied. Therefore RASTA-PLP's channel robustness characteristics make it a particularly attractive processing method for this task. Another desirable feature of RASTA-PLP processing is that it emphasizes phone transitions. As mentioned in Section 1.2, it has been posited that phonetic transitions occur asynchronously in sub-bands. For these reasons, most of the experiments in this dissertation relied on RASTA-PLP features. Recently, it was suggested [49] that PLP features may be better suited for multi-band processing. Comparisons between sub-band RASTA-PLP and PLP features were also performed and are reported in Section 7.4.

The division of the frequency range into four sub-bands of [216-778 Hz], [707-1631 Hz], [1506-2709 Hz], and [2121-3769 Hz] resulted in [4, 4, 3, 3] critical-band filter values in the four regions, respectively. It has been shown that often a smoothed representation of the acoustic spectrum improves recognition [51]. Therefore, the number of poles for all-pole modeling in each region was chosen to be one less than the number of critical-band values in that region. In other words, [3rd, 3rd, 2nd, 2nd]-order RASTA-PLP features, respectively, were derived for the four bands. The RASTA-PLP values were supplemented with sub-band energy and delta parameters for all of the above. Delta features describe the evolution of the static parameters over time [40] and have been shown in various experiments to improve the performance of the speech recognizer [77].

[4] A formant is defined as a group of overtones corresponding to a resonating frequency of the air in the vocal tract. Vowels are characterized by three formants [76]. For further reading on formant regions see [76, 80].

3.3.3 Sub-Band Probability Estimators

The overall performance of the multi-band system is dependent on the accuracy of the sub-band systems. Therefore, the sub-band systems must be carefully designed.

Since the work in this dissertation has been performed within the framework of an HMM/MLP system, MLPs were an obvious choice for sub-band probability estimators. One MLP probability estimator was trained on the acoustic features derived from each sub-band. Similar to the full-band MLP, the input layer to each MLP had a context window of nine frames, that is, one current frame, four frames into the past, and four into the future. The total number of inputs was [72, 72, 54, 54], respectively, for each MLP. Hidden layer sizes of [497, 497, 372, 372] were chosen so that the total number of parameters in the four MLPs and a comparison baseline full-band system (153 input units, 1000 hidden units, 56 output units) were roughly equal, since both had around 209,000 parameters. The number of hidden units in each net was chosen to be proportional to the input layer size, so that sub-bands with more input features were allocated more parameters. There were 56 output units, one for every phone, as in the full-band MLP.

The MLPs were fully connected multi-layer perceptrons, trained using a back-propagation algorithm [119]. Softmax normalization [19] and relative entropy error criterion [129] were used on the output layer. The MLPs estimated the probability of each phoneme corresponding to (multiple) frames of speech (Figure 2.2), $P(q_i|X_i)$, where q_i is a phone in the phone class for band i, and X_i is an acoustic observation sequence for sub-band i.

In summary, one MLP probability estimator was trained on the acoustic features derived from each sub-band, and the number of parameters in the four sub-band MLPs was equal to that of the full-band baseline system.

3.3.4 Combining the Streams

An important problem is finding an optimal strategy for combining the sub-bands. If the merging method does not take full advantage of the information contained in each band, the overall performance will suffer.

It is not clear whether the sub-band information streams should be combined on a per-frame level or fluidly over a phone or even a syllable. As summarized in Section 2.2.1, Bourlard and Dupont [12] have run preliminary experiments with state, phone, and syllable combination levels and their results were inconclusive. Tomlinson et al. [138] have claimed that asynchrony was helpful in a two-band system, though their results were not generalizable to a three-band system. Tibrewala and Hermansky's results did not benefit from allowing asynchrony [135].

Merging the sub-band information on the frame level is fairly simple, and it involves estimating the overall phonetic probabilities for each frame given the frame-by-frame probability estimates from the sub-band probability estimators. Merging on the phone, syllable, or word level poses a bigger challenge, since they may be implemented using an algorithm that allows asynchronous merging, such as HMM decomposition [139] or two-level dynamic programming [114]. Increased space and search time requirements can be a problem with these algorithms. Chapter 5 is dedicated to the discussion of asynchronous merging. In the following sections, combining the sub-band systems on the frame level is discussed.

3.3.5 Merging on the Frame Level

Obtaining an overall frame-level probability estimate for the multi-band system entails combining the sub-band probability estimates on a frame-by-frame basis according to a merging function. There has been much work in the statistics community on methods for combining the votes or probability estimates of a group of experts [63, 67, 56, 145, 148, 110, 144, 16, 17, 18, 50, 73]. Combination of multiple classifier algorithms can be divided into two general categories: (1) classifier fusion, which refers to applying all the individual classifiers in parallel, and combining their outputs to reach a group consensus and (2) dynamic classifier selection, which is a method of predicting which single classifier is most likely to be correct for a given sample.

These combination methods attempt to maximize the classification accuracy. In automatic speech recognition, however, improved phonetic classification is not necessarily correlated with improvements in word-recognition accuracy. Hence, choosing a complex classifier combination strategy with respect to only frame-level accuracy can be misleading. To compensate for this, combining the final output of the recognizer based on decoding distance metrics, such as N-best list re-scoring and word-lattice re-scoring, have been performed. Unfortunately, if the accuracy of the sub-systems is poor, which is the case in the multi-band paradigm, such word-level combination methods perform very poorly and are not practical. In this thesis, simple classifier fusion methods have been employed for combining the probability estimates of the sub-bands and building a multi-band system.

Traditionally, non-linear methods have had higher accuracy than linear combination strategies. Note that linear merging in this context refers to applying a linear merging function to the probability estimates of *non-linear* (MLP) functions. In previous multi-band experiments, using an MLP as a merging function has produced lower word error rates than simple merging schemes (e.g., multiplying or averaging the probabilities) [12, 135]. Another issue concerns whether the merging should be done in a posterior or likelihood domain. In all the experiments cited stream were merged in the likelihood domain. In Section 3.4.3 different ways of combining the probabilities (e.g., adding or multiplying the probabilities) and merging in the likelihood and posterior domain are explored.

Simple Merging Functions

A major advantage of simple merging functions is that they are fast and easy to apply. The overall probability estimate is calculated as a weighted combination of the frame-by-frame sub-band probability estimates (Figure 3.2).

In pilot experiments, the product of frame-wise, sub-band-scaled likelihoods was found to be the best simple merging function for word-recognition purposes. This finding has been confirmed recently [71].

The probability stream of each band may be scaled by a weighting factor. Weighting mechanisms such as SNR weighting [108, 135, 12], normalized phone accuracy rates [12], and inverse of conditional entropy [108] can improve the word recognition accuracy in the presence of noise. For clean speech, these weighting mechanisms improve the recognition results marginally [12] and equal weighting appears to be just as good. Since clean speech has been the focus of the experiments in this thesis, equal weights for all sub-bands are applied and averaging versus multiplying the sub-band scaled likelihoods are compared.

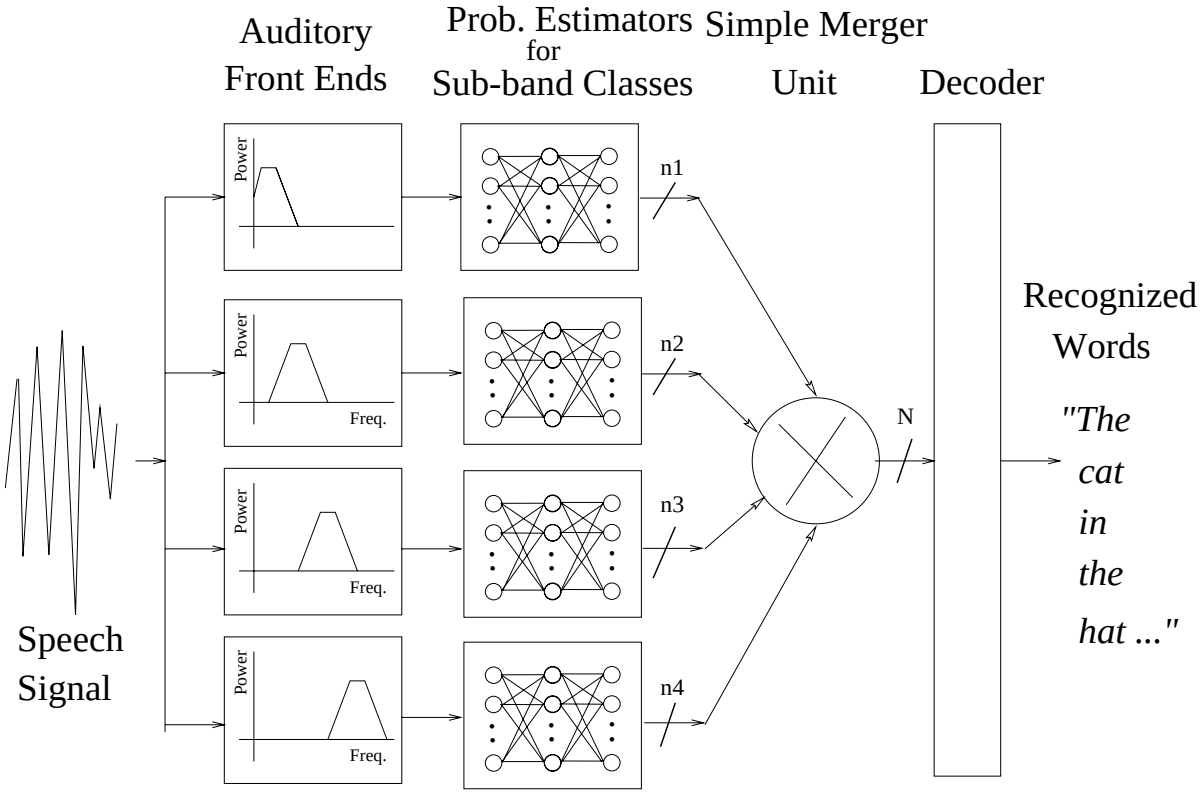

Figure 3.2: Merging on the frame level using a simple merger. The circle with X may be a multiplication, addition, or any other simple function.

A More Complex Merger

MLP probability estimators were used as a non-linear merging function to estimate the overall phonetic probability given the phonetic probabilities of the sub-bands. Recall Equation 3.6:

$$\log p(X|M) = \sum_{j=1}^{J} f(W, \log p(X_j^k|M_j^k), \forall k)$$

where W is a global set of recombination parameters, which may include weighting factors for each sub-band. In a similar approximation, an MLP was trained to estimate the function f and parameters W:

$$p(X|M) = \sum_{j=1}^{J} f(W, p(X_j^k|M_j^k), \forall k) \tag{3.7}$$

As an alternative, an MLP was also trained on the sub-band posterior probabilities, instead of the scaled likelihoods, to estimate overall posterior probabilities:

$$p(M|X) = \sum_{j=1}^{J} f(W, p(M_j^k|X_j^k), \forall k) \tag{3.8}$$

Two parameters in the MLPs were also chosen for experimentation: the size of the hidden layer and the size of the context-input window. Changing the size of the hidden layer changes the number of parameters (or weights) in the MLP. The size of the context input window, in addition to changing the total number of parameters, determines how many frames of the past and the future probability distributions are to be considered for classification of the current frame.

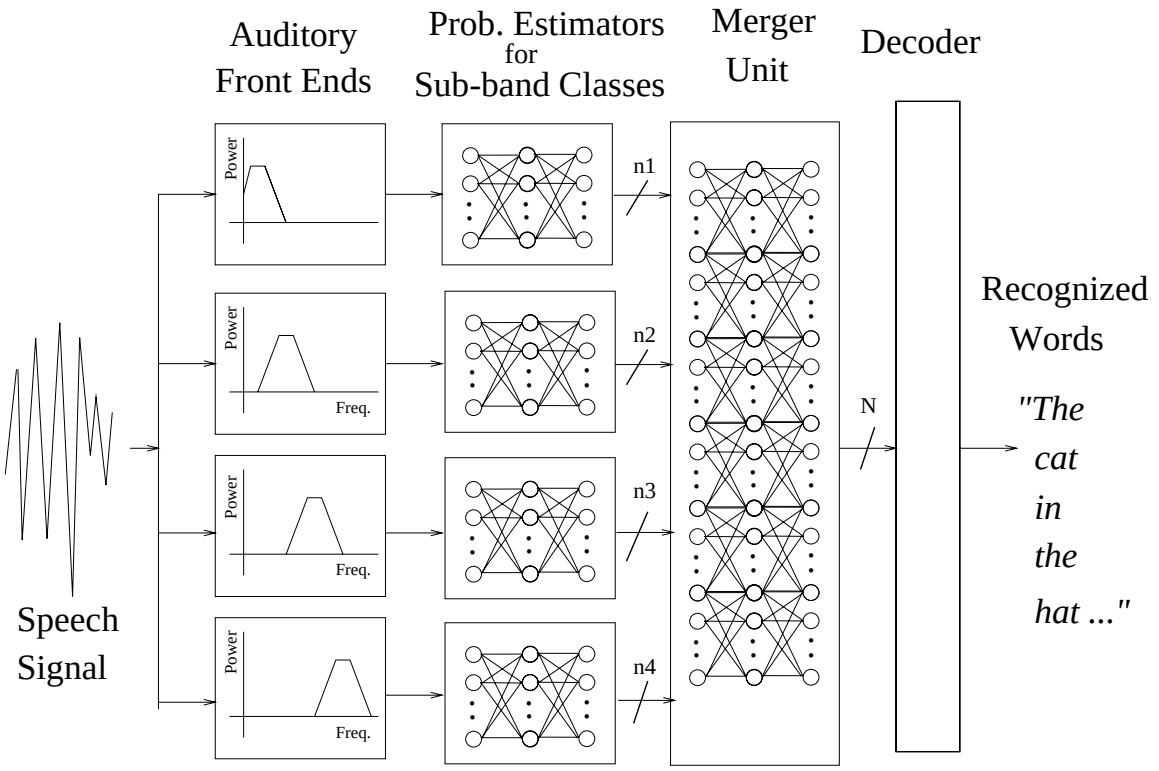

Figure 3.3: Merging on the frame level using an MLP.

3.3.6 Embedded Alignment

Embedded training and alignment [146] is a process for generating phonetic labels from word transcriptions or improving the phonetic labels if the hand-transcribed phonetic labels are unreliable. This procedure, which is a variant of the well-known Generalized Estimation (or Expectation) and Maximization (GEM) [32, 88] algorithm, entails iteratively training the probability estimator and improving the phonetic labels[5], sometimes together with updating the lexicon. This procedure often improves the recognition results. The efficacy of this approach had not been verified in a multi-band paradigm, and is discussed in more detail in Section 3.4.4.

[5] A new set of phonetic labels were generated in each iteration using the forced Viterbi procedure, which attempts to find the best phonetic sequence (or state path) that matches the given word transcription and the generated probability stream.

3.3.7 Other System Issues

The lexicon was created[6] by examining the phonetic hand-transcriptions of the training set and generating multiple pronunciations for each word. These pronunciations provided a 90% coverage of the actual pronunciations in the training set. The pronunciations were then converted to a working lexicon.

In experiments reported in Section 7.4, an improved lexicon was employed. The original lexicon was refined by applying one iteration of forced alignment, which matched the lexicon with the improved phonetic labels by revising the word pronunciations and phone durations. It has been experimentally shown at ICSI that applying more than one iteration of forced alignment to the lexicon results in only marginal improvements.

A bigram language model, with back-off uni-gram probabilities, was generated from the statistics of the training set and used for the experiments in this thesis. For decoding, Y0 [118], a Viterbi decoder with pruning and forced alignment capabilities, was used. The word models were context-independent, multiple pronunciation HMM models. The state-to-state transition probabilities of the word models were uniformly set to $1/T$, where T is the number of transitions out of a particular state. The number of states for each phone in a word model were determined, based on the context-dependent average duration of the phone in the training set.

There were two input window sizes that could be varied in each sub-band system: (1) the signal-processing frame-window size, and (2) the number of acoustic input vectors to the MLP, also known as the MLP input window size. Widening these windows allows more past and future contextual information to be processed in addition to the data segment of interest. Contextual information is particularly important in speech, since coarticulation[7] strongly affects the acoustic representation.

One can imagine using a different window size for each band. For example, for experiments involving room reverberation, the size of the window could be larger for the lower frequencies and smaller for high frequencies, which are typically less smeared. For normal conversational speech, different window sizes may allow capturing different dynamics of speech, such as short-term and long-term variations [132].

Experiments at ICSI have shown that an MLP context window of nine frames (four vectors of context into the future, and four into the past) and an analysis signal processing window of 25 ms to be satisfactory choices, and they have therefore been chosen for the sub-band systems. In some of our experiments with reverberant speech, different window sizes in different sub-bands were used (see Section 7.2 for details).

3.4 Experiments

3.4.1 Acoustic Features Experiments

To show that the narrow-band features contained the information necessary for ASR, all of the sub-band features were concatenated into one long feature vector. *One* MLP was trained

[6] Thanks to Dan Gildea and Eric Fosler-Lussier for creating this lexicon.

[7] Coarticulation is the term used to refer to the change in the articulation and acoustics caused by the influence of neighboring segment in the same utterance [31].

on these features, which also may be viewed as training a *full-band* system on *multi-band* features. The word error rate of this system was 7.9%. The word error rate for the baseline full-band system trained on full-band RASTA-PLP eighth-order features (plus deltas and delta energy) was also 7.9%. It was concluded that for clean speech, RASTA features derived from the sub-band and the full-band contained a similar amount of information relevant to this task.

Four MLPs were trained on sub-band RASTA-PLP acoustic features (as described in Section 3.3.2). The input layer to each MLP had a context window of nine frames – one current frame, four frames into the past, and four frames into the future. The total number of inputs was [72, 72, 54, 54] for each MLP, respectively. Hidden layer sizes of [497, 497, 372, 372] were chosen so that the total number of parameters in the four MLPs and the full-band system were roughly equal (both had roughly 209,000 parameters), and the number of hidden units in each net were proportional to the input layer size. There were 56 output units, one for every phone, as in the full-band MLP. Table 3.5 includes the frame and word errors for each sub-band, full-band, and multi-band systems. The frame error (on the development set) of the four sub-band systems ranged between 37.9% and 50.4%. The word errors followed a similar trend. The word errors of the four sub-band systems ranged between 24.9% and 47.8%. The sub-band streams were merged simply by multiplying the scaled likelihoods on a frame-by-frame basis. The word error rate on the Numbers95 development set was 11.5%.

As a side-note on Table 3.5, notice that when the bands are combined (by multiplying the scaled likelihoods), the word error decreases much more radically than the frame error. As explained in Section 4.1.3 the relationship between the word and the frame error rate is not linear. For a given system, the frame-probability distribution may be relatively accurate overall, but the expected (or correct) phone label may often not have the highest, but the second highest, probability. Since frame error is calculated based on the phone-labels with the highest probability, such frames would be categorized as an error, whereas for word recognition, in the Viterbi search, the relatively high probability of the desired phone contributes to a high score for the accurate word. Another effect may be that the additional contextual information (in this case, the probability of the neighboring frame in the larger utterance context) improves the recognition accuracy. Influences of context are also commonly observed in psycho-acoustic experiments.

For the sake of completeness, 15 critical-band energies and deltas were also used to train a full-band recognition. The word error rate of this system was 7.9% on the Numbers95 development set, which was similar to that of the full-band systems trained on either the full-band or the narrow-band cepstral features. The performance of the multi-band system with these features was worse, however. Four MLPs, with sizes similar to the ones in the RASTA-PLP feature experiments, were trained on critical-band and corresponding delta features. The sub-band streams were merged in a simple manner by multiplying the sub-band scaled likelihoods. The word recognition error rate was 14.1%, which is significantly higher than the 11.5% error observed with RASTA-PLP features. A similar result was observed in my pilot experiments in which spectral features were outperformed by cepstral features.

	Larger MLPs (209,000 params)		Smaller MLPs (95,200 params)	
	Frame Err.	Word Err.	Frame Err.	Word Err.
Band1	40.3	33.7	40.9	34.2
Band2	37.9	24.9	39.6	27.8
Band3	43.3	34.4	44.9	36.6
Band4	50.4	47.8	50.2	48.1
Multi-Band	37.2	11.5	38.1	12.4

Table 3.5: The word and frame error for multi-band systems with different numbers of parameters, measured on the Numbers95 development set. "Larger MLPs" refers to [497, 497, 372, 372] hidden units in each sub-band MLP, respectively. "Smaller MLPs" refers to 200 hidden units in each sub-band MLP. The sub-bands were merged by multiplying the log likelihoods.

3.4.2 Varying the Number of Parameters

To see if the same level of accuracy could be sustained by having fewer parameters, the number of MLP hidden units in the sub-band probability estimators was reduced. Instead of choosing the number of hidden units in the MLP to keep the number of parameters in the multi-band system equal to that of the full-band system, the number of hidden units was reduced to 200. The total number of parameters in the multi-band system was effectively cut in half, reduced from 209,000 to 95,200. As shown in Table 3.5, the frame error, as well as the word recognition error rate, became slightly worse, increasing from 11.5% to 12.4%. The merging of the sub-bands was done by multiplying the scaled likelihoods of the four sub-bands. Interestingly, reducing the number of parameters of the multi-band system by half did not hurt word recognition performance significantly. However, in order to make a fair comparison between the full-band and the multi-band system, it was decided to keep the number of the parameters in the multi-band system equal to that of the full-band system.

In summary, the number of hidden units in the multi-band system was chosen such that the total numbers of parameters in the multi-band and the full-band system were equal. The hidden units for the four bands were [497, 497, 372, 372], respectively. These sizes were chosen in proportion to the size of the input layer, such that an MLP with a larger input space was allocated more parameters.

3.4.3 Merging Experiments

Simple Merging Functions

In pilot experiments in 1995, product of likelihoods proved to be the best simple merging scheme. In this work, the sum of likelihoods and the product of likelihoods for the four sub-bands were compared. The word error rates on the Numbers95 development set were 14.9% and 11.5%, respectively, for sub-band RASTA-PLP features. Product of likelihoods seemed to be the best simple merging scheme. This observation has been confirmed by

[71], who compared product, sum, and normalized maximum and minimum combinations, among other simple functions.

Merging with an MLP

Word Err.	Hidden Units	Input Window Size	Num of Params
9.4%	50	11	126,000
9.7%	50	9	104,000
9.6%	50	7	81,000
10.2%	50	3	36,000
10.0%	50	1	14,000
8.9%	100	9	207,000
9.1%	100	7	162,000
9.2%	100	3	73,000
9.6%	100	1	28,000
9.1%	400	1	112,000
8.3%	300	1	87,000
8.9%	200	1	56,000
9.6%	100	1	28,000

Table 3.6: The word error rate on the Numbers95 development set as the number of hidden units and input window size in the merger MLP were varied. The merger MLPs were trained on the Numbers95 training set.

As mentioned above, an MLP probability estimator was chosen as the non-linear merging function. The size of the input context window and the number of hidden units were varied. As seen in Table 3.6, the number of parameters in the merger MLP increases dramatically as the input window size is increased. It appears that increasing the number of hidden units and keeping the window size small is a more efficient way of allocating extra parameters. This comparison has been demonstrated in Table 3.7, which is a re-arrangements of the rows in Table 3.6. It appears that 50 hidden units are insufficient for capturing the complex function effectively. For example, an MLP merger with 50 hidden units and a window size of seven frames has roughly the same number of parameters as one with 100 hidden units and a window of size three, as well as another with 300 hidden units and a window size of one. However, the word error rate of the first (9.6%) is higher than that of the other two (9.2% and 8.3%, respectively).

Increasing the number of parameters, in a couple of cases, may lead to over-generalization. For example, increasing the size of the hidden units from 300 to 400 for an input window of one (parameters increasing from 87,000 to 112,000), the error rate increases from 8.3% to 9.1%. Similar behavior is observed when comparing the second and third row of Table 3.6, although the error rate continues to decrease further when the number of parameters is increased.

Word Err.	Hidden Units	Input Window Size	Num of Params
9.7%	50	9	104,000
9.1%	100	7	162,000
9.1%	400	1	112,000
9.6%	50	7	81,000
9.2%	100	3	73,000
8.3%	300	1	87,000
10.2%	50	3	36,000
9.6%	100	1	28,000
8.9%	200	1	56,000

Table 3.7: The word error rate on the Numbers95 development set as the number of hidden units and input window size in the merger MLP were varied. This rearrangement of the previous table highlights the difference in performance given roughly the same number of parameters in each system.

Considering the interplay between the number of hidden units and input window size, an MLP with 300 hidden units and a window size of a single frame was chosen.

It was suggested by Fred Jelinek [65] that training the merger MLP on the same training data on which the sub-band MLPs were trained could lead to over-training. As an alternative, a 50-hidden-unit merger MLP, with a window of nine frames, was trained on the cross-validation data. The word error on the development set was 12.6%, which was significantly worse than what was found for the same size MLP trained on the training data (9.7% error rate). The reason may have been that the cross-validation set was too small (12 minutes of speech, about 70,000 frames) and did not provide enough variation of training samples. As Numbers95 is a small corpus and training data are scarce, this hypothesis cannot be adequately tested. Training the merger on an independent data set should be performed on a larger corpus.

Another parameter tested was the method of estimating the overall phonetic posterior probabilities, from either the sub-band scaled likelihoods or posteriors. Again, one merger MLP with 300 hidden units and a window length of one was trained on the scaled likelihoods of the sub-band MLPs, and another was trained on the posterior probabilities. The error rate on the Numbers95 development set was 8.3% for the posteriors and 8.6% for the likelihoods. Although the differences were not significant for this dataset, significant differences may be observed for a larger corpus. Merging posteriors may make more sense, as the implicit prior information may be useful [10].

3.4.4 Performing Embedded Alignment

For multi-band recognition, would the overall accuracy improve if we align the phone labels in each sub-band iteratively and then align the phone labels in the merger MLP iteratively? Embedded alignment on each sub-band independently (without re-estimating the lexicon) was performed. The word error rate on the Numbers95 cross-validation set, averaged over

the four sub-bands, was used as the stopping criterion. This measure reached a minimum after three iterations. Feed-forward probabilities from the separately aligned sub-band systems were generated and embedded alignment was performed on the merger MLP (300 hidden units, one frame window, as described in Section 3.3.5), stopping after two iterations, according to the cross-validation error. The final word error rate for the Numbers95 development set was 8.5%, which was not statistically different from the unaligned system with a word error rate of 8.3%. One reason for the slight degradation may be that the phone labels in high frequency sub-bands (which were less accurate overall) become gravely misaligned and hurt the accuracy of the multi-band system.

Aligning the merger MLP labels alone is an alternative that avoids the above-mentioned problem. Without re-aligning the sub-band systems independently, embedded alignment was performed on the merger MLP alone. The alignment was stopped after the third iteration, after the cross-validation error had reached a minimum. The word error rate of this system was 8.1%, which was slightly better than the unaligned system at 8.3%.

3.5 A Fairness Comparison with a Full-Band System

A valid question is whether the multi-band system has an unfair advantage over the full-band system, since the former benefits from an extra level of training that can further improve the probability estimates. If this is the case, an improvement in the results should be observed if a second MLP is trained to further refine the probability estimates obtained from the first MLP (Figure 3.4). To test this, a second MLP (with 300 hidden units and a window of one, similar to the multi-band merger MLP) was trained on the training data, and training stopped according to the cross-validation frame error. The input to this net was the output of the full-band MLP, and the output was the posterior probabilities for the 56 phone classes. Surprisingly, the word error of this full-band system increased to 9.0% (up from 7.9%). Training on the same data may be causing over-training. Tibrewala [132] observed an improvement when she trained a similar probability estimator on a data set *different* from the training set.

Given our training condition (i.e., training the merger MLP on the same data), the multi-band system does not seem to hold an unfair advantage over the full-band system.

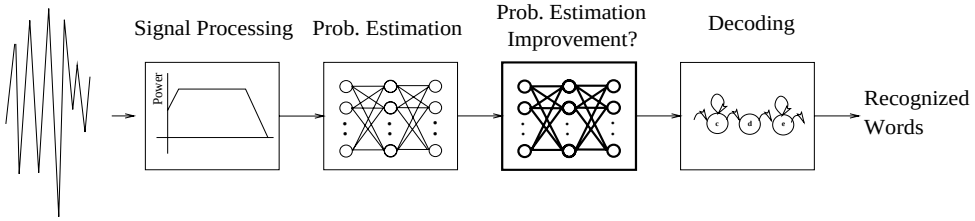

Figure 3.4: The system setup to test if the full-band system would benefit from an extra layer of refinement of the probability estimates. The highlighted box is an MLP probability estimator that is trained on the probability estimates of the first MLP probability estimator.

3.6 Summary of the Multi-Band Baseline System

Most of the chosen systems for the experiments in this dissertation (except as noted) have the following characteristics: The baseline full-band system is an HMM/MLP-based [15] system. The baseline MLP phonetic probability estimator is trained on a nine-frame window of 8th-order RASTA-PLP cepstra [52], energy, and delta-RASTA-PLP cepstral features over a 25-ms window, stepped every 10 ms. The MLP is fully connected and has 153 inputs (nine frames with 17 features per frame), 1000 hidden units, and 56 outputs (one output for each phone[8]), and is trained using back-propagation with softmax normalization at the output layer. The system is trained on hand-transcribed phone labels (without embedded realignment). A multiple pronunciation lexicon (derived from the hand transcriptions), a bigram language model and a synchronous-time decoder called Y0 (described in [118]), which uses a single density per phone with repeated states for a simple durational model are used. The word error rate of this baseline system on the test set is 7.9%.

In the multi-band system, the frequency range is divided into four bands of [216-778 Hz], [707-1631 Hz], [1506-2709 Hz], and [2121-3769 Hz]. From the sub-bands, [3rd, 3rd, 2nd, 2nd]-order sub-band RASTA-PLP cepstral features, respectively, are derived, as well as energy and corresponding deltas. Four MLPs, one on each sub-band, are trained on these acoustic features. The input layer of each MLP has a context window of nine frames, for total input layer sizes of [72, 72, 54, 54], respectively. Hidden layer sizes of [497, 497, 372, 372] are chosen, so that the total number of parameters in the four MLPs and the full-band system are roughly equal. There are 56 output units, one for every phone, as in the full-band MLP. The frame-by-frame information from the four sub-band streams is combined in two ways: (1) by multiplying the sub-band scaled likelihoods and (2) by using a *merger* MLP. In the case of the first merging scheme, the resulting likelihood stream is decoded using the Y0 Viterbi decoder. The word error rate on the Numbers95 development set is 11.5%. In the second merging scheme, the MLP merger accepts the output of the sub-band MLPs as input, has 300 hidden units, and an output of 56 phones. The word error rate of this multi-band system on the Numbers95 development set is 8.3%, which statistically is not different from the word error of the full-band system (7.9%).

[8] Note that some of the 56 phones do not occur in the Numbers database and have a prior probability of zero.

Chapter 4

Analysis of Common Multi-Band ASR Assumptions

Is multi-band ASR inherently inferior to a full-band approach because phonetic information is lost due to the division of the frequency space into sub-bands? Do the phonetic transitions in sub-bands occur at different times? The first statement is a common objection of the critics of multi-band ASR, and the second, a common assumption by multi-band researchers. This chapter is dedicated to finding answers to both of these questions.

The most common objection to the use of separate statistical models for each band has been that important information in the form of correlation among bands may be lost. Our experience and that of our colleagues has been that recognition performance has not been hurt by this approach. Nevertheless, this chapter examines the estimator performance in a more detailed fashion. In particular, the phonetic feature transmission pattern in each sub-band, the merged multi-band, and full-band probability streams is analyzed. As discussed in Section 4.1, methods similar to those of Miller and Nicely [89] are used to calculate confusion matrices for phone and feature classes, and mutual information is chosen as the measure of information transmission in a channel.

Section 4.2 focuses attention on the following: some multi-band researchers [138, 13, 135, 92, 8] have postulated that transitions in sub-bands occur asynchronously, and that a phone- or syllable-level merging of multi-band streams is necessary to permit independent alignment for each band within the merged unit. However, this hypothesis has not been tested. Neither has there been a study of transition boundary shifts in the presence of speech signal variations (such as room reverberation or speaking rate). Without such evidence, consideration of longer-term merging units for multi-band ASR cannot be substantiated. Section 4.2 examines this assumption by analyzing the transition lags in each sub-band to see if sub-band transitions occur asynchronously.

4.1 Is Phonetic Information Lost?

The perennial objection of skeptics of the multi-band paradigm is that because each sub-band system receives information from only a limited frequency range, phonetic information inherent in the correlation among the bands is lost. This section examines this claim.

4.1.1 Experimental Setup

For this analysis phone and broad-feature confusion matrices are used, similar to the seminal studies of Miller and Nicely [89] on human speech recognition.

A confusion matrix (CM) is simply an extended matrix of *hits* and *misses* for all classes, as in Table 4.1. The column headings represent the features that are *transmitted*, and the row headings correspond to the *received* features. In Table 4.1, for example, 93 instances of [s] are perceived as [eh]. Frame-level phonetic classification on the test set is used to generate phone CMs. To better observe the patterns in the data, the phone CMs are merged according to membership in broad-feature classes (as in Table 4.2), and feature confusion matrices are generated (e.g., Table 4.3). Phonetic classes are classified according to six broad-features[1]: *CV* (consonant, vowel, silence), *duration* (short, long, mid), *frontness* (front, back, neither), *manner* (vowel, diphthong, liquid, glide, stop, closure, nasal, fricative, silence), *place* (high, low, mid, labial, dental, coronal, palatal, retroflex, velar, glottal, silence), and *voicing* (voiced, unvoiced).

Received Phones	Transmitted Phones				
	t	s	eh	sil	$\cdots$
t	5722	252	31	316	$\cdots$
s	258	8495	110	1159	$\cdots$
eh	11	93	3118	37	$\cdots$
sil	436	2733	68	40237	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 4.1: An example of a phone-based confusion matrix.

	vowel	consonant	silence
t	-	+	-
s	-	+	-
eh	+	-	-
sil	-	-	+
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 4.2: An example of binary acoustic features for CV classification.

Inspired by the work of Miller and Nicely [89], the confusion matrix is summarized using a measure of covariance between the input and output. For an input value x, which can

[1] This division has been done by Gary Tajchman at ICSI based on [143]

Received Features	Transmitted Features		
	vowel	consonant	silence
vowel	74393	6962	1816
consonant	6738	61030	5055
silence	2321	8922	49281

Table 4.3: An example of a feature-based confusion matrix.

assume the discrete values $i = 1, 2, \ldots, k$ with probability p_i, the *entropy*[2] of the input is:

$$H(x) = E(-\log p_i) = -\sum_i p_i \log p_i$$

If the base of the logarithm is 2, entropy is the number of bits of information per stimulus. The entropy of the output variable y, which can assume values $j = 1, 2, \ldots, m$, can be defined similarly. The number of decisions needed to specify the particular stimulus-response pair is the *joint entropy*, where p_{ij} is the probability of the joint occurrence of input i and output j. *Mutual information*[3], a measure of the covariance of the input with output, is defined as:

$$I(x;y) = H(x) + H(y) - H(x,y) = -\sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}}$$

Mutual information may be thought of as the transmission from x to y in bits per stimulus. The *relative transmission* can be calculated as:

$$I_{rel}(x;y) = I(x;y)/H(x)$$

Since $H(x) \geq I(x;y) \geq 0$, this ratio varies from 0 to 1. If the transmission is poor, this ratio will be close to 0. If the response can be predicted relatively accurately from the stimulus, then this ratio will be close to 1. We can also define *percent bits transmitted* by multiplying the above ratio by 100.

Since the true probabilities are not known, they must be estimated from the data. The probabilities p_{ij}, p_i, and p_j are estimated from n_{ij}/n, n_i/n, and n_j/n, respectively, where n_i is the frequency of stimulus i, n_j is the frequency of response j, and n_{ij} is the frequency of the joint occurrence of stimulus i and response j in a sample of n observations.

The transmission of each phonetic sub-broad-feature (e.g., sub-feature *fricative* $\in$ broad-freature class *manner*) is further calculated by reducing the full CMs to a 2x2 CM for each $sub\_feature$ and $\overline{sub\_feature}$) (the results for the broad-feature class *manner* are illustrated in Figure 4.2).

[2]Extended discussions of information theory can be found in many sources, e.g., [28].

[3]Note that the chosen measure represents the amount of information transfer for a given *a priori* distribution of the input classes, as opposed to *capacity* of the channel, defined as $C = max_{p(x)} I(X;Y)$, which is the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.

4.1.2 Observations on Feature Transmission

Figure 4.1 shows all features, and Figure 4.2 shows sub-features of *manner*, transmitted as a percentage of the maximum, where the maximum is the entropy of the transmitted feature. The following are observed:

1. Multi-band feature transmission is always as good as or better than the comparable full-band system, except for the feature *frontness*. On average, 60.94% of the features are transmitted for the multi-band system compared to 59.06% for the full-band system for 54,000 acoustic frames.

2. The results are consistent with our knowledge of acoustic phonetics. For example, the low frequency band is expected to contain the most information about *voicing*. The observed patterns for *fricatives* and *nasals* were similar to those of Miller and Nicely [89].

3. Low and sometimes mid frequency bands (often band 1 and sometimes band 2) transmit most of the feature information alone. For example, band 2 transmits 87% of the *frontness* features that are transmitted by the full-band system.

4. There is much redundancy in phonetic information content in the sub-bands. The sum of information transmission over all bands far exceeds 100%. Lippmann [83] has highlighted this redundancy as a source of human robustness to speech degradations.

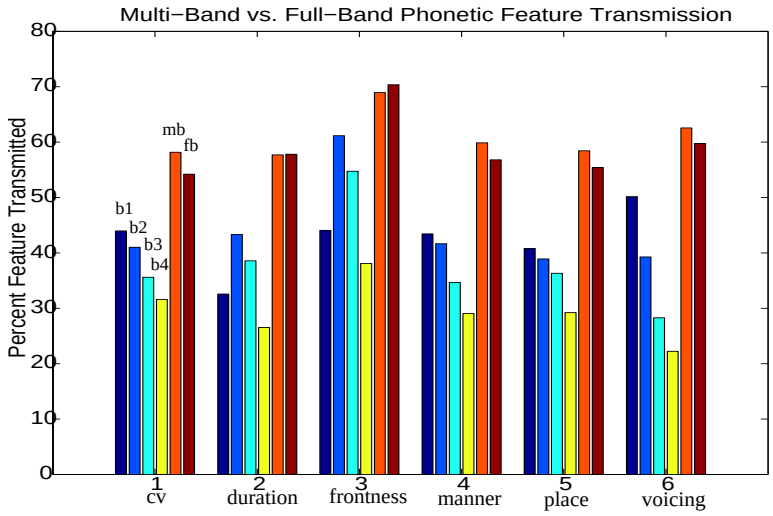

Figure 4.1: Phonetic features transmitted as a percentage of maximum possible, measured by mutual information.

4.1.3 Feature Transmission and Word Error Rate

Since the percent feature transmission has been calculated using frame-level scores, it will be correlated with the frame-level accuracy, but not necessarily with word recognition accuracy. Word- and frame-level accuracy in ASR do *not* have a correlation of 1. In other

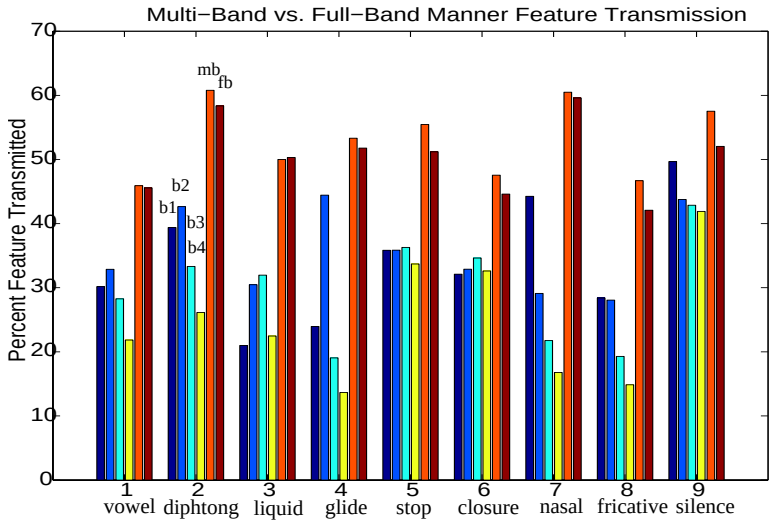

Figure 4.2: Manner of articulation features transmitted as a percentage of maximum possible, measured by mutual information.

words, the phone-level score of system A may be higher than system B, yet the word-level accuracy of system B may be higher than that of system A. This phenomenon occurs because frame-level scores are calculated based only on the phone label with the highest probability (winner-take-all). The entire probability distribution is not considered, whereas the entire probability distribution is utilized in calculating the word-level score. For the example above, system A's *winning* phone label may more often be the *correct* phone label, accounting for a higher frame-level accuracy. However, in the instances when the winning phone label is *incorrect*, the probability assigned to the correct phone label may be very small. In system B, in contrast, the probability of the correct phone label may always be the second-best, yet almost equal to that of the winning phone for all frames. When multiplying the probabilities over all frames to find the best matching word-string, system B may have a more accurate overall likelihood distance for the correct utterance, and hence, a higher word-level accuracy.

4.2 Do Transitions Occur Asynchronously?

Multi-band researchers have posited that transitions occur asynchronously in sub-bands, and a phone- or syllable-level merging of multi-band streams may be necessary. This section is dedicated to the study of this hypothesis.

4.2.1 Experimental Setup

In order to obtain the phone transition boundaries, forced alignment (i.e., Viterbi realignment and retraining the MLP in each iteration) is performed on each sub-band independently. Furthermore, to allow maximum freedom of shifting in transition boundaries, embedded realignment is allowed for six iterations. The word error rate on the Numbers95

cross-validation set is the stopping criterion, and it reaches a minimum value after the second iteration of realignment.

Instead of using the usual multiple-pronunciation word lexicon, whole-sentence models are used in the forced alignment to ensure that identical phone sequences are taken in each sub-band. Whole-sentence models are generated using the phonetic hand-transcriptions and the corresponding average phone durations. An example of the sub-band transition patterns is shown in Figure 4.3.

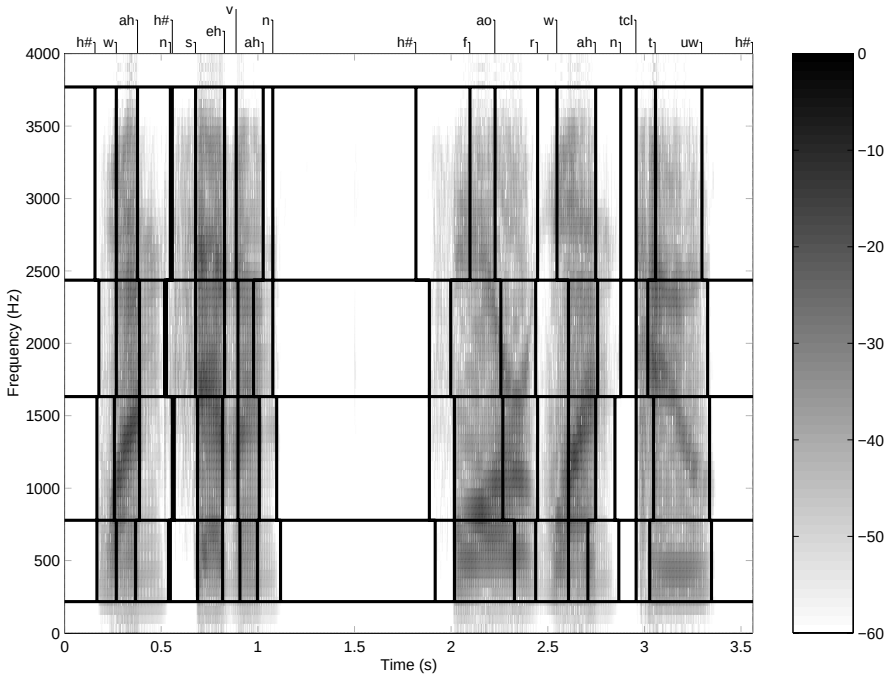

Figure 4.3: The spectrogram for "One seven four one two", showing where the phone label transitions occur in each sub-band, determined by embedded alignment.

Transition statistics are also generated on digitally reverberated versions of the data (as described in Section 3.1.3), as well as on fast and slow speech. The cutoff for fast (slow) speech is set to one standard deviation above (below) the mean rate of the training set. The speaking rates were measured in phones per second and determined from a count of manually transcribed phones over non-silence speech segments (see [95, 94, 93, 90] for a discussion of speaking rate measurement).

For any given phone transition, the transition lags in each sub-band were calculated as compared to (1) the full-band and (2) other sub-bands. Figure 4.9 shows the histograms of average transition lags of the four sub-bands with respect to the full-band for broad phonetic features, where each plot in row *feat1* and column *feat2* corresponds to a *feat1* $\rightarrow$ *feat2* transition. Appendix C includes a list of broad phonetic category membership for the phonetic classes.

4.2.2 Observations on Asynchrony of Transitions

Examining the generated statistics, some evidence for asynchrony is observed. More precisely:

1. Transition lags (with respect to the full-band transition boundaries) appear to have a Gaussian distribution, with a mean close to zero, indicating that, on average, the transition lags happen in both directions, and a standard deviation of [2.8, 3.3, 5.0, 5.6] frames for the sub-bands, respectively. The higher the frequency range, the more shifted the transition boundaries are compared to the full-band.

2. More distant sub-bands have less agreement in transition boundaries. For example, the standard deviation (σ) of transition lags between sub-bands 1 and 4 is 5.9 frames, and between sub-bands 1 and 2 is 3.8 frames. Figure 4.4 shows the transition lags for every sub-band pair. The neighboring bands, represented in the diagonal, have the tightest distributions (tallest zero-difference bar), and in more highly separated bands (for example, sub-band 1 compared to 4) the spread of the transition lags becomes larger. In other words, there are greater lags between bands that are further apart. However, all the transitions in sub-band 4 do not routinely occur before those of sub-band 1.

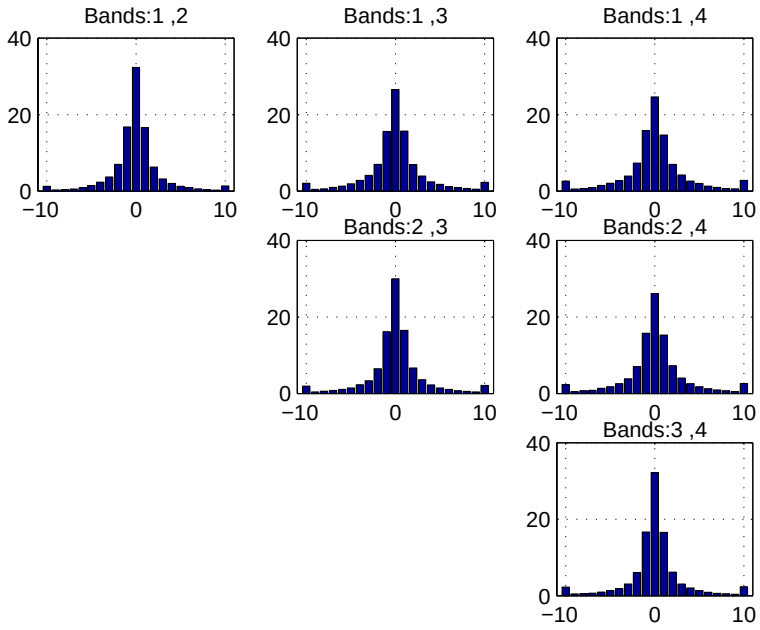

Figure 4.4: Histogram of transition lags for every sub-band pair for all the training data. Each frame corresponds to 10 ms.

3. In some cases, though (e.g., stop→vowel or fric→vowel for sub-band 4 vs. full-band comparison shown in Figure 4.8), there is a skewed distribution, meaning that the

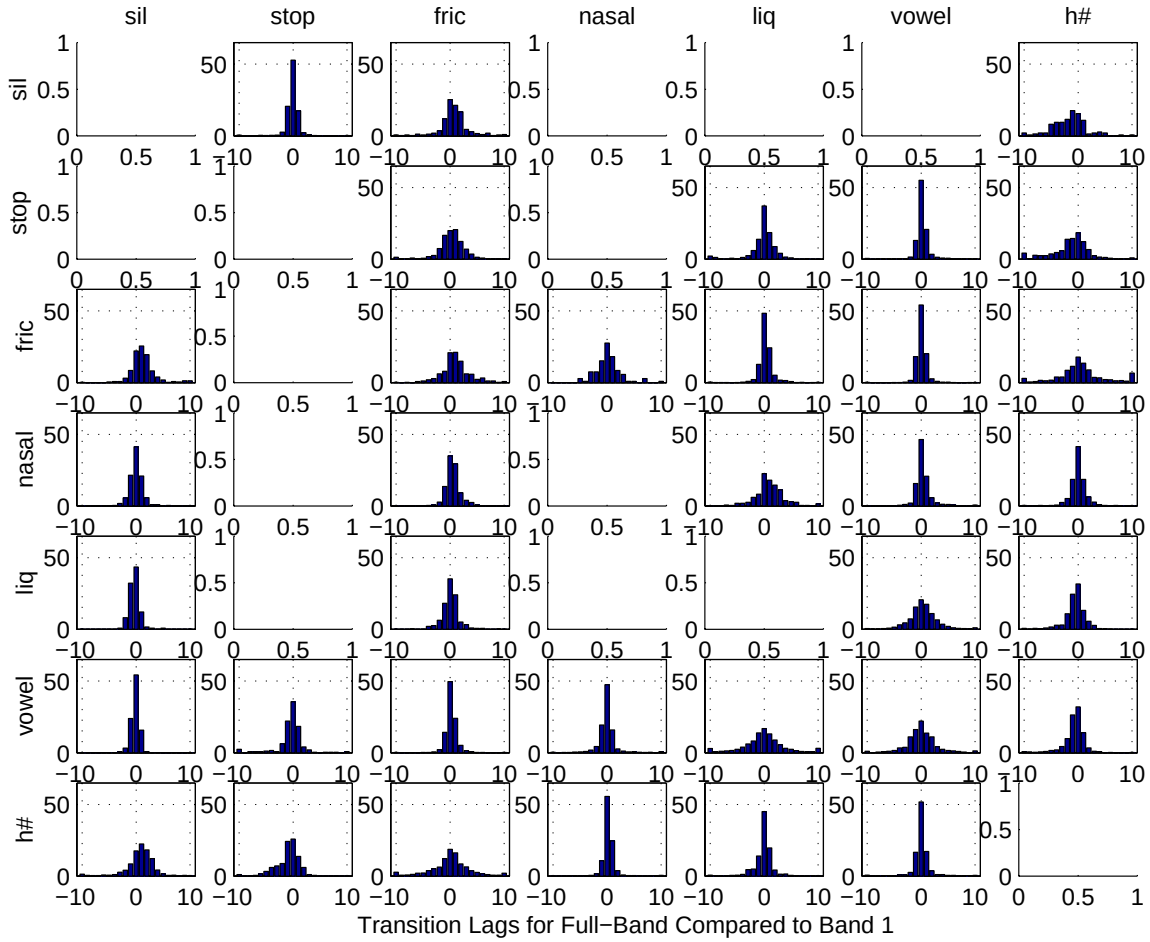

Figure 4.5: Histogram of transition lags for band 1 compared to the full-band, for all the training data. Each frame corresponds to 10 ms. [h#] is silence. "sil" refers to the closure phones. Graphs with fewer than 100 points have not been plotted.

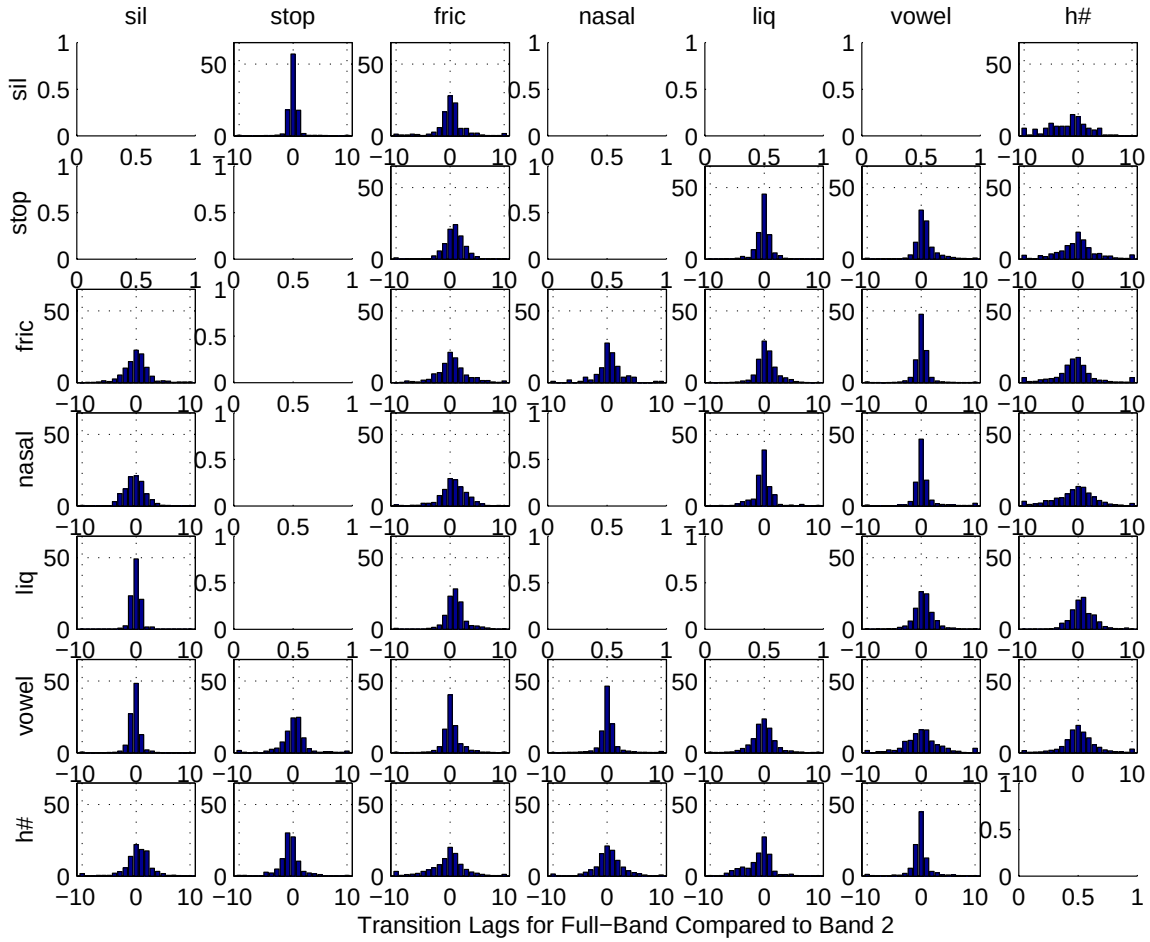

Figure 4.6: Histogram of transition lags for band 2 compared to the full-band, for all the training data. Each frame corresponds to 10 ms. [h#] is silence. "sil" refers to the closure phones. Graphs with fewer than 100 points have not been plotted.

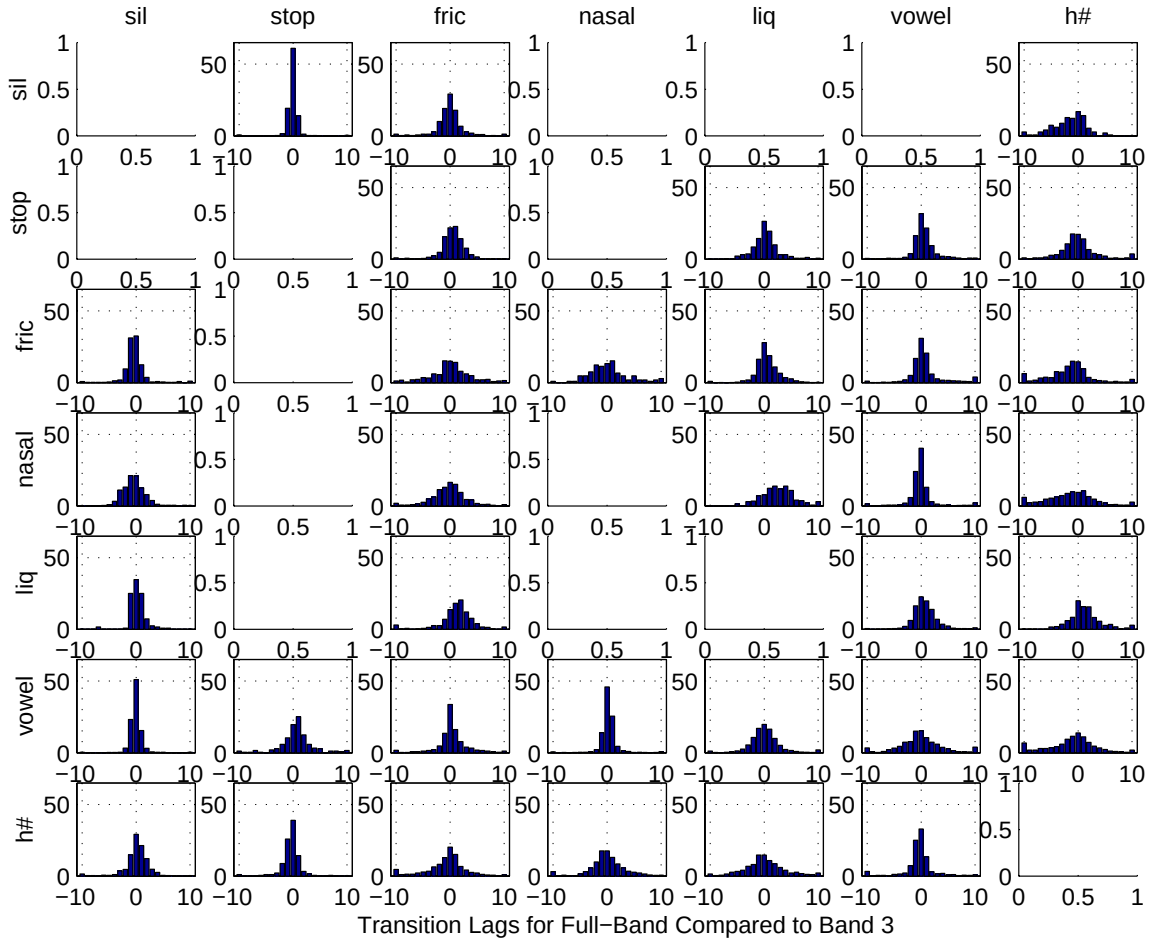

Figure 4.7: Histogram of transition lags for band 3 compared to the full-band, for all the training data. Each frame corresponds to 10 ms. [h#] is silence. "sil" refers to the closure phones. Graphs with fewer than 100 points have not been plotted.

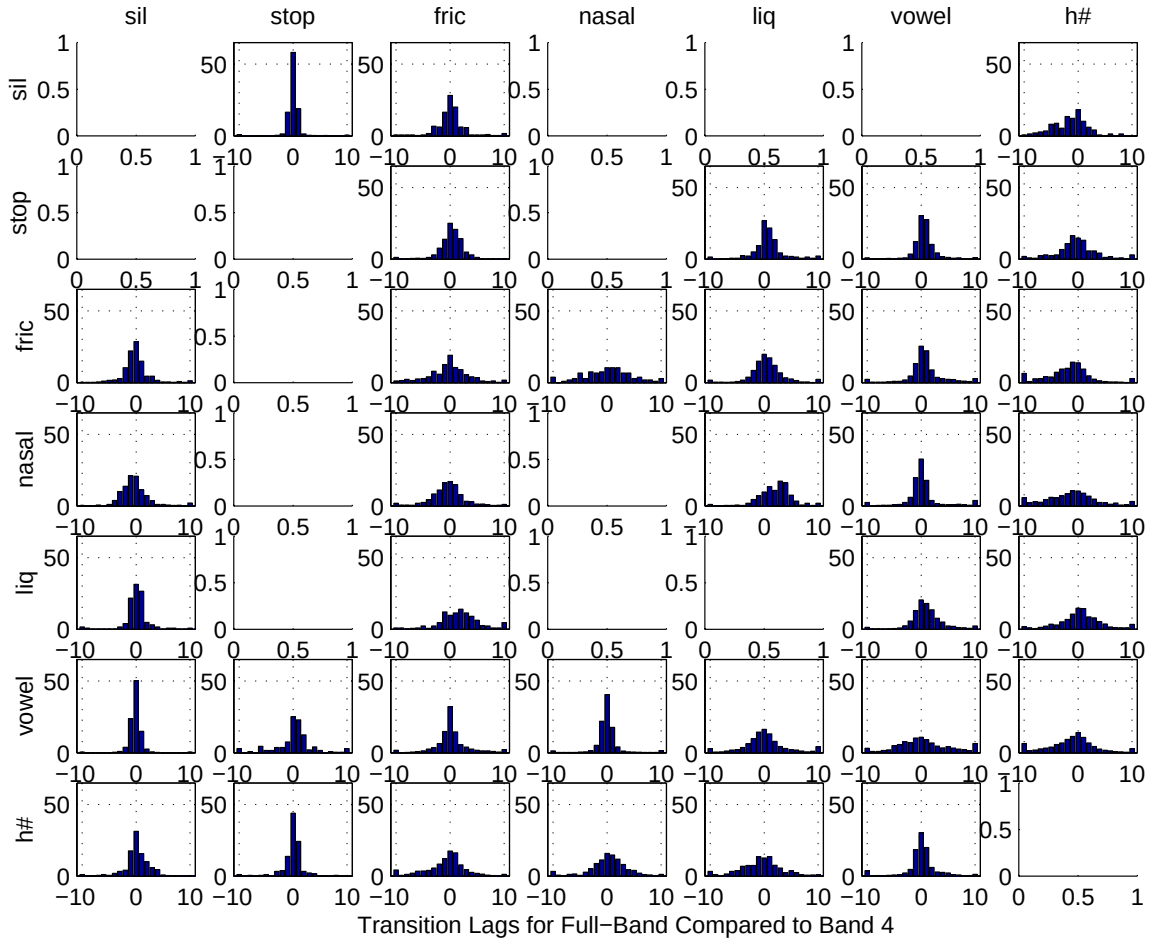

Figure 4.8: Histogram of transition lags for band 4 compared to the full-band, for all the training data. Each frame corresponds to 10 ms. [h#] is silence. "sil" refers to the closure phones. Graphs with fewer than 100 points have not been plotted.

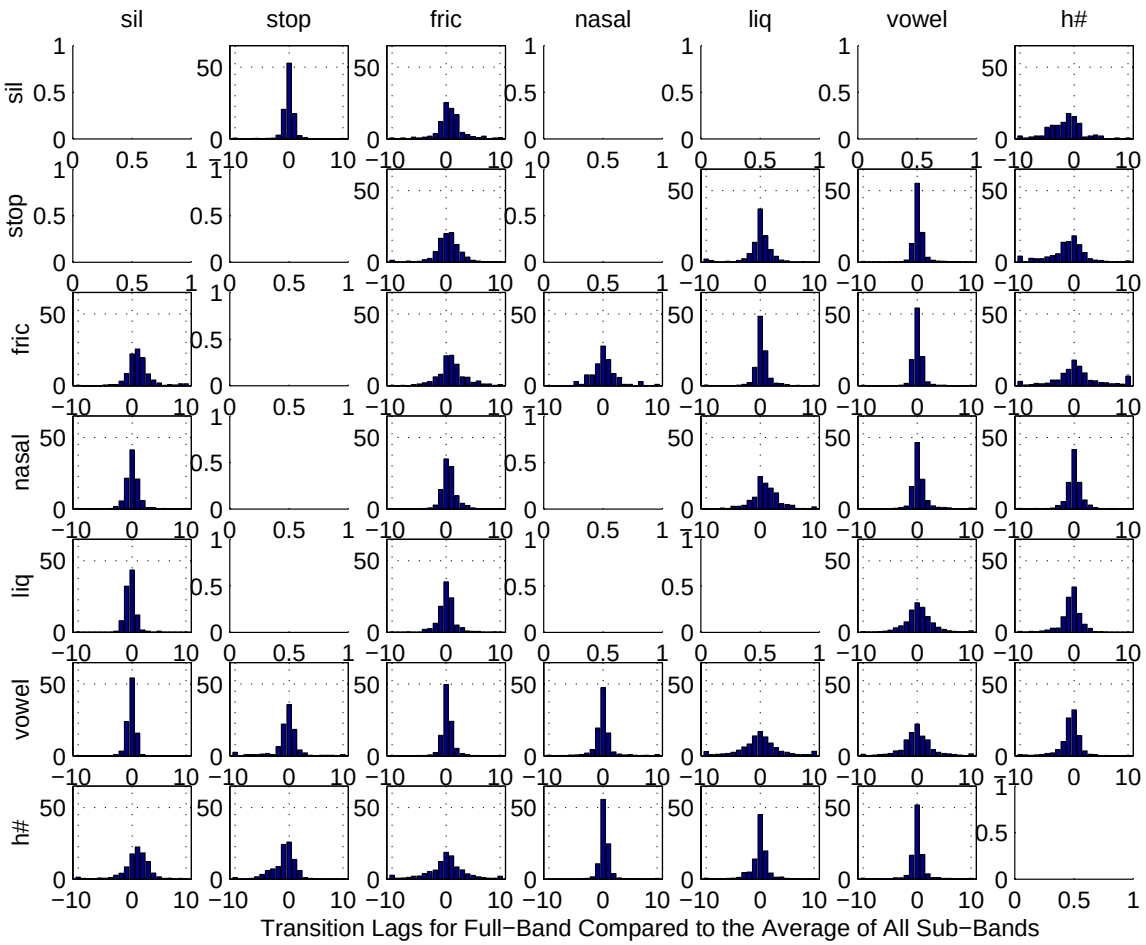

Figure 4.9: Histogram of average transition lags for broad phonetic features for the full-band compared to the average of the four sub-bands. Each frame corresponds to 10 ms. [h#] is silence. "sil" refers to the closure phones. Graphs with fewer than 100 points have not been plotted.

Condition	band 1	band 2	band 3	band 4
Slow	3.7	3.6	9.8	9.2
Medium	2.8	3.1	4.2	5.1
Fast	2.1	4.1	2.8	3.6
Reverb	4.0	4.4	5.5	6.3
Clean	2.8	3.4	5.0	5.6

Table 4.4: Standard deviation for sub-band transition lags as compared to the full-band transition boundaries.

transitions for these broad phonetic features occur later than in the full-band. More such examples can be seen in Figures 4.5 through 4.8.

4. About 30% of the sub-band transitions do not occur within 50 ms of each other. Similarly, 44%, 41%, and 21% of the transitions for reverberated, slow, and fast data, respectively, do not occur within 50 ms of each other.

5. Some broad feature transitions are sharp (e.g., sil $\rightarrow$ stop), and some have a relatively flat distribution (e.g., vowel $\rightarrow$ liquid) (see Figures 4.5 through 4.8 for more examples).

For contrast conditions of speaking rate and room reverberation, strong changes in transition timing are observed, as reflected in a difference in the variance rather than a systematic difference in the means. Table 4.4 shows that for three out of the four bands, the standard deviation of the per-band lag decreases as speaking rate increases, which conforms to the intuition that phone durations decrease with rate. The table also suggests that the higher frequency transitions are the most sensitive to speaking rate variations.

Table 4.4 further confirms the original intuition that reverberation should affect transitions more at low frequencies than at high frequencies, since most common room boundary materials are less absorptive at low frequencies, leading to longer reverberation times at those frequencies.

4.2.3 Real Asynchrony or Alignment Noise?

One valid concern is whether the observed asynchrony is a real effect or simply a product of alignment noise. It is possible (and in fact, likely) that some random noise or "alignment jitter" is being included in the measurements. It would be expected, however, that such randomness would be small, whereas the observed differences are often much larger than a 1- to 2-frame average difference. In addition, if the measurements were only a manifestation of alignment noise, the observations for fast, slow, and reverberant speech would not have been as consistent with the characteristics of such speech signals. Finally, the spread of random noise would be Gaussian, with samples both before and after the true transition point, and would not give rise to skewed distributions. Figures 4.5 through 4.8 show instances of skewed distributions for particular broad class transitions. For example, the liquid $\rightarrow$ fricative transition in bands 3 and 4 (Figures 4.7 and 4.8, respectively) occurs consistently

later in these sub-bands than in the full-band. Similar results have been observed by Morris and Pardo [101] when analyzing the transition patterns in each sub-band according to energy. They observed that the patterns of onsets and offsets of the phone transitions across frequency tended to be quite stable for each transition.

4.3 Conclusions

Two common assumptions about multi-band ASR have been examined: (1) the objection of the critics of multi-band ASR that it is inherently inferior to a full-band approach because phonetic information is lost due to the division of the frequency space into sub-bands; and (2) the assumption by multi-band ASR researchers that transitions in bands often occur asynchronously (i.e., at different times than the full-band transition).

To study the first point, phonetic feature *transmission rates* for each sub-band was calculated. The above objection was not substantiated; in fact the contrary was observed. The second hypothesis was confirmed by analyzing the *transition* lags in each sub-band.

The exploration of the first question further showed that even when using a simple multi-band merging method, phonetic features are transmitted better (60.94% for our database) than the comparable full-band system (59.06%) for roughly 54,000 frames.

For the second question, no single band's phone transitions were found to be consistently delayed or advanced compared to others, as the per-band transition lags had a mean close to zero. In other words, the transitions in one band did not always occur before another band. However, some broad phonetic feature transitions did show a skewed distribution in some bands, meaning that the acoustic cues often appear later (or earlier) than in other bands.

The spread of the transition lags were dependent on both frequency and contrast conditions (speaking rate and reverberation). Roughly one-third of the sub-band transitions in the control condition do not occur within 50 ms of each other. Furthermore, the high frequency band timings have a spread that is strongly dependent on speaking rate.

It appears that sub-band alignments can have significant timing deviations from the full-band alignments; thus, it would be expected that there is a potential for improvements in acoustic modeling if longer time-scale information stream merging (e.g., phone or syllable) is used. This will be examined in the next chapter.

Chapter 5

Asynchronous Merging of the Sub-Band Streams

Section 4.2 discussed the results of the analysis that showed that the phone label transitions occur at different times in different bands. In this chapter, ways to combine the sub-band streams asynchronously are studied to see whether taking advantage of this inherent asynchrony improves the recognition accuracy.

First of all, what does it mean to asynchronously merge the bands? Such an approach permits a phone-state transition to occur later – or earlier – in one band than another when there are sufficient acoustic cues to warrant such a decision. For example, if maximum evidence for phone transition $\alpha \rightarrow \beta$ is available in frame t in sub-band i, and the same phone transition, $\alpha \rightarrow \beta$, is best supported by the acoustic evidence in frame $t + \delta$ in sub-band j, evidence from frame t from band i is combined with evidence from frame $t + \delta$ from band j. As mentioned, it is suspected that by allowing the acoustic cues to be realigned temporally independently according to the inherent asynchrony across the sub-bands, recognition accuracy will improve.

The first algorithm considered was HMM-recombination, which is better known as either Parallel Model Combination or HMM-decomposition [139, 42]. Later, two-level dynamic programming [122] was implemented for this task. The following sections discuss these two algorithms and their experimental results for this task.

5.1 HMM-Recombination

HMM-recombination was the first of the two algorithms I applied for asynchronously merging the sub-bands. Section 5.1.1 describes the algorithm, and Section 5.1.2 discusses the implementation and the experimental results.

5.1.1 Algorithm Description

HMM-decomposition has traditionally been used for speech recognition in the presence of noise [139, 42]. The main idea is to separate noise and speech into two streams, assuming that each is produced by a separate model. Similarly, HMM-recombination can combine several independent streams into a single model. The most intuitive way to think of the

algorithm for multi-band purposes is to consider a two-band system. If each sub-band stream is decoded independently with its own sub-band model, the acoustic data in the sub-bands may best match different words. A reasonable constraint is to assume that the acoustic information in every sub-band must match the same model. Hence, this approach would force both streams to consider the same word model given a start and finish time, yet allow freedom for each band to transition from state to state within a word as the sub-band acoustic information content deems necessary. In the simple example of the two uni-dimensional models in Figure 5.1, two separate streams of data, one for each sub-band, may be decoded independently using each model. Combining the two models and clamping the enter and exit states (represented as black circles) creates a two-dimensional model (as in Figure 5.2). The two-dimensional model could be expanded and more clearly expressed, as shown in Figure 5.3, where each new state is a product of the two old states. In other words, assuming independence between the two bands, we have: $p(X|S_1, S_a) = p(X_1|S_1)p(X_2|S_a)$, where X_1, X_2 and X are acoustic information for band 1, band 2, and the full-band, respectively, and S_i is a state in the HMM. Or more generally, the likelihood may be estimated as:

$$p(X|M) = \prod_{k=1}^{K} p(X_k|M_k). \tag{5.1}$$

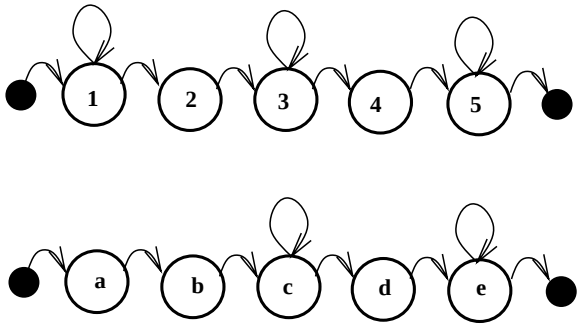

Figure 5.1: Two uni-dimensional HMM models.

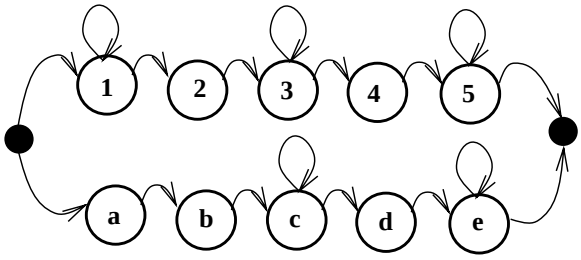

Figure 5.2: An unexpanded multi-dimensional HMM model.

As the reader might suspect, the size of these multi-dimensional models can get prohibitively large for realistic word models. This problem can be alleviated, to some extent, by enforcing a maximum number of states of asynchrony. For example, if a maximum

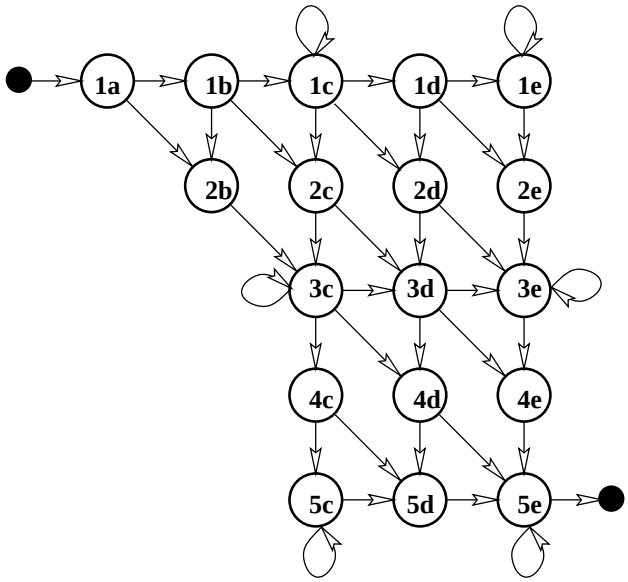

Figure 5.3: An expanded multi-dimensional HMM model.

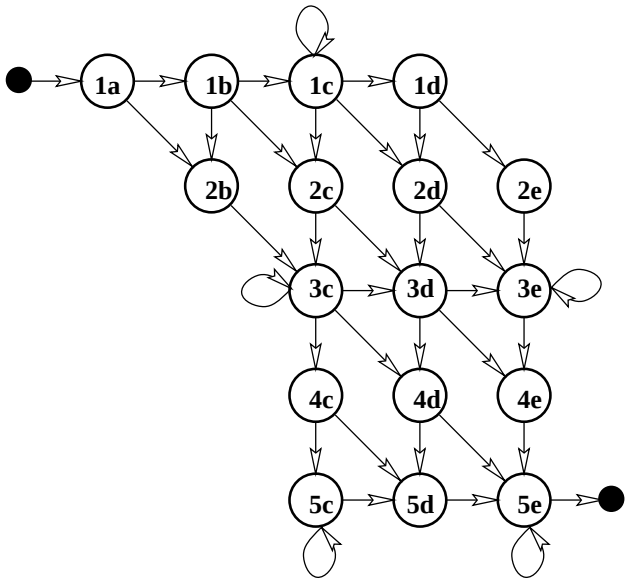

Figure 5.4: An expanded multi-dimensional HMM model, with maximum asynchrony limit of three states.

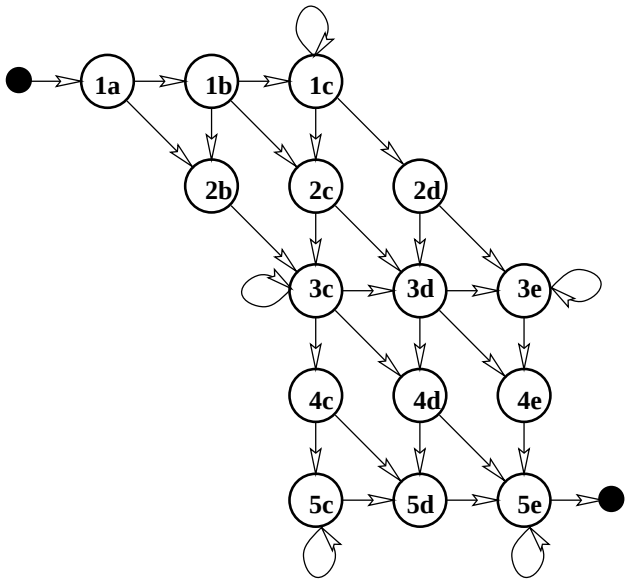

Figure 5.5: An expanded multi-dimensional HMM model, with maximum asynchrony limit of two states.

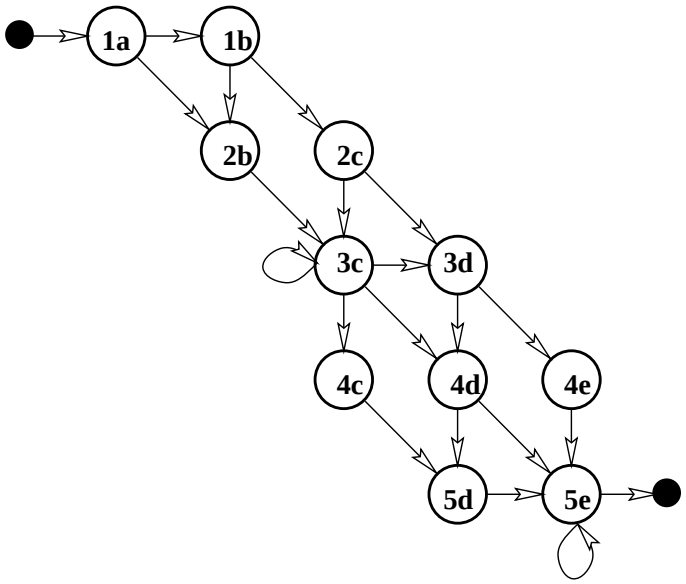

Figure 5.6: An expanded multi-dimensional HMM model, with maximum asynchrony limit of one state.

asynchrony of three states is allowed, state **1e** is pruned away (Figure 5.4). If asynchrony is further limited to a maximum of two states, states **1d** and **2e** are further pruned (Figure 5.5). Finally, if asynchrony is limited to only one state, **1c, 2d, 3e** and **5c** are pruned (Figure 5.6). The practical issue of the explosion of number of states is discussed in Section 5.1.2.

5.1.2 Experimental Results

The implementation[1] of HMM-recombination was performed in the following way:

1. Multi-dimensional word models with a given maximum asynchrony constraint were created.

2. The scaled likelihoods for each new state were calculated by multiplying the likelihoods of the old states.

3. Viterbi decoding was run on the new multi-dimensional model, given the newly generated data likelihoods.

As suggested in Section 5.1.1, an explosion in the size of the models proved to be a problem. As before, a phone set of size 56 elements was used, 32 phones of which had non-zero priors for Numbers95. For a four-band system, allowing just a single (and later two) state of asynchrony, the number of multi-dimensional phone states increased drastically (to 1810 and 3410, respectively) Accordingly, the size of the acoustic likelihood file increased from 48 MB to 1.6 GB and (an estimated) 3.0 GB. The size of the word models, similarly, increased drastically. The multiple-pronunciation word model for the number *seventeen* had 372 states originally, and for one state of asynchrony, the number of states increased to 5549. A comparable explosion in the number of states was observed by Dupont [36].

Because of the prohibitive time and space requirements for the combination of four parallel models, I decided to focus on a two-band system. The size increase of this system proved to be more manageable. Allowing one, two, or three states of asynchrony, the total number of two-dimensional states increased by a factor of 4 to 6, reaching 286, 314, and 336, respectively. The multiple-pronunciation word model for the number *seventeen*, for example, had 1121, 1837, and 2531 states, respectively, instead of 372 states.

The frequency ranges for the two-band system were [216-1631Hz] and [1506-3769Hz]; each band of the two-band system was made up of two contiguous bands of the four-band system. The 6th- and 3rd-order RASTA-PLP features for the lower and higher sub-bands were derived and MLPs with 820 and 470 hidden units were trained for the sub-bands. For a baseline, two-band comparison, a simple merging of the two bands was performed by simply adding the log likelihoods and the resulting stream was decoded using Y0 decoder. The word error rate was 9.0%. The word error rates for HMM-recombination are listed in Table 5.1. Experiments with reverberant speech were also conducted, since evidence for higher levels of asynchrony had been observed. It was surmised that if there were any gains to be made for asynchronous decoding using this method, they would be visible in reverberant speech. The results are reported in Table 5.2.

[1] The HMM-recombination program was implemented together with Su-Lin Wu at ICSI, based on original code from Stéphane Dupont at IDIAP.

HMM-Recombination on Clean Numbers	
Max. Asynchrony	word error rate
None	9.0%
1	9.1%
2	9.5%
3	9.8%

Table 5.1: Word error rates for HMM-recombination asynchronous merging algorithm on clean numbers as the maximum states of asynchrony is increased for a two-band system.

HMM-Recombination on Reverberant Numbers	
Max. Asynchrony	word error rate
None	35.4%
1	37.1%
2	37.0%

Table 5.2: Word error rates for HMM-recombination asynchronous merging algorithm on reverberant numbers as the maximum states of asynchrony is increased for a two-band system.

Increasing the level of asynchrony tolerance did not improve the word recognition performance, and in fact, it hurt it slightly. It may be that the synchrony requirement provides needed constraints and relaxing this requirement increases the confusion because of the explosive increase in the number of alternative states. Also note that as the number of states in the word models grow with the increased asynchrony allowance, the word error also increases.

An alternative algorithm for asynchronous merging is two-level dynamic programming, which as a bonus does not have prohibitive space requirements. That algorithm allows us to experiment with a four-band system as well.

5.2 Two-Level Dynamic Programming

Streams can also be merged asynchronously by incorporating the two-level dynamic programming algorithm [122]. This approach does not have the potential time and space limitation of the HMM-recombination algorithm. In Section 5.2.1, the algorithm is briefly described and in Section 5.2.2 implementation and experimental results are discussed.

5.2.1 Algorithm Description

The main idea of the two-level dynamic programming [122, 114] is to perform the decoding in two stages: the first level matches each individual word (or other unit) model against an

arbitrary portion of the test string. The second level of the computation pieces together the individual reference pattern scores to minimize the overall accumulated distance over the entire test string.

The two-level dynamic programming algorithm is rarely used for connected word recognition nowadays, and has often been replaced with the one-pass algorithm [140, 20, 103]. The one-pass algorithm computes best paths to every reference pattern frame at every test frame and finds the best word sequence by backtracking. The main advantage of the one-pass algorithm is that the computation is frame-synchronous, which makes real-time implementation of the algorithm possible.

5.2.2 Experimental Results

In our implementation[2] of this algorithm, synchrony is enforced on the word level. In the first stage of the algorithm, for every word and every sub-band a distance matrix is calculated. Every entry (i, j) in the distance matrix signifies how likely it is for the word to have been uttered, starting at frame i and ending at frame j. Our chosen structure is different from that of the original algorithm. In the original algorithm, mainly for space considerations, only the calculated distance score for the best reference word was stored in the matrix and therefore only one matrix was sufficient for storing all the information. In our implementation, since the distances for all words in the bands have to be combined, it is necessary to store all the information. Synchrony at the start and end of the word unit is enforced by adding all the distance matrices for the sub-bands for a given word. Next, word lattices are created for each utterance. And finally, a search on the lattices[3] [104, 102, 107] is performed to generate the string with the least distance.

Without any pruning, the produced lattice can be unmanageably large, since there would be an arc for every word for every start and finish time. On-line garbage model pruning [11] was used. This is a simple on-line pruning method that has proved as effective as some of the more sophisticated off-line approaches that require training. Dynamically, the top n scores are averaged; the scores above this threshold are kept, and the rest are pruned. Experimentally, it was determined that an n of 10 produced lattices of reasonable size in an acceptable amount of time. The word error rates are reported in Table 5.3.

Similar to the HMM-recombination algorithm, allowing increased asynchrony using two-level dynamic programming did not improve the recognition results. Perhaps the conclusion to draw is that, for clean speech, asynchronous merging does not have a significant effect on the word error rate. The word error rates are very similar to when the streams are simply log linearly merged. It is not too surprising since, in both methods, the probability streams are multiplied. If alternative asynchronous paths are not exploited, the two methods of stream merging are essentially equivalent, which would explain the results.

For reverberant speech, the phone-transitions in the sub-bands are more temporally spread, as observed in the analysis in Section 4.2, so the results should be more affected if the sub-bands are allowed to merge asynchronously. The results, however, may depend on the size of the sub-band. If the sub-bands are large enough to include sufficient acoustic cues, the

[2] Thanks to Eric Fosler-Lussier for the in-house implementation of this algorithm.

[3] A lattice, for the purposes of this work, is defined as a directed acyclic graph where each edge corresponds to a word with a distance score and each node corresponds to a point in time.

Two-level Dynamic Programming on Numbers95		
Condition	Simple Merge WER	2-level DP WER
Clean, 4 bands	11.5%	11.1 %
Clean, 2 bands	9.0%	9.3%
Reverb, 4 bands	39.1%	45.9 %
Reverb, 2 bands	35.4%	37.7 %

Table 5.3: Word error rates (WER) for two-level dynamic programming for clean and reverberant speech on the Numbers95 development set.

probability stream would be relatively accurate, and the freedom to merge asynchronously hurts only slightly – analogous to the clean speech case. However, if the sub-bands are too narrow, in the presence of reverberation, acoustic cues would be gravely degraded, and in combination with relaxed constraints, the word error rate may degrade significantly.

5.3 Conclusions

I observed that relaxing the synchrony constraints when merging the multi-band streams did not improve word recognition accuracy. The results were consistent both for word-level (in the case of two-level dynamic programming) and multiple state-level (for HMM-recombination) relaxation of synchrony constraints.

As summarized in Section 2.2.1, Tibrewala and Hermansky [135] had also observed differences in optimal sub-band paths and conjectured that relaxing the temporal synchrony requirement among sub-bands would improve the word error rate. The word error rate degraded slightly (though, not statistically significantly) when relaxing the synchrony requirement over a word (using Viterbi decoding on each stream), compared to enforcing synchrony at every state for an isolated digit recognition task both for a four-band and a seven-band system. In a similar experiment for isolated German word recognition, Bourlard and Dupont [12] observed no improvement when the synchrony was relaxed from the frame level to the phone level, and only a small (not statistically significant) improvement when the merging was performed on the syllable level for a three-band system using HMM-recombination. Finally, Tomlinson et al. [138] have reported a slight improvement ($p < 0.1$) when synchronization is performed on a three-state, instead of a per-state, level for a two-band system, and no improvement for a three-band system using HMM-recombination.

It is interesting to note that although the system choices made in each of the asynchronous merging experiments reported above are different in terms of choice of asynchronous merging algorithm, number of sub-bands, level of merging (multiple states, phones, syllables or words), acoustic features, database, etc., the results are similar in that the effects of allowing asynchronous merging in a sub-band system for clean speech have been marginal.

It may be that by disregarding the synchrony information, important information is being lost. As observed in Section 4.2.2 in Figures 4.5 through 4.8, some broad phone

category transitions in some sub-bands occur systematically earlier or later than the full-band average. If at all, synchrony requirement should perhaps be relaxed only for particular phone transitions, and not indiscriminately for all phone transitions, as it has been done in previous work. Furthermore, the amount of the permitted asynchrony may have to depend on the phone-class transitions and the training-data statistics. Along these lines, Morris and Pardo [101] have also observed that the patterns of onsets or offsets of the phone transitions across frequency bands tend to be quite stable for each transition, suggesting that warping the sub-bands to align the transitions might remove the potentially useful information that this characteristic transition pattern could provide.

In summary, various reasons may justify the observed results. As mentioned above, useful transition information may be lost. Or perhaps the transition constraint is aiding the Viterbi search by reducing the number of potential paths and transition options. On the other hand, it may be that there is a "gain" from relaxing the synchrony assumptions, but only for a limited number of phone to phone transitions, and that by allowing synchrony relaxation for all transitions, the above-mentioned "gain" is being lost.

Based on the significantly higher computational costs and the currently available evidence, I am forced to conclude that relaxing the synchrony constraint, in this form, is unlikely to be advantageous in a multi-band ASR system.

Chapter 6

Deriving Reduced Sub-Band Phone Classes

This chapter focuses on deriving reduced phonetic classes for each sub-band. The intuition is that some frequency bands contain more information for distinguishing particular acoustic classes. The corollary of this hypothesis is that some bands possess little or no information for distinguishing among some of the classes. Therefore, using only relevant information to distinguish among *limited* classes may be better than trying to categorize *all* classes using either limited or all the available (including noisy) information.

6.1 Motivation

If phone classes are highly confused with each other in a sub-band, it might be desirable to merge them into a single super-class for that particular sub-band. Note that the merged classes may be different for each sub-band. For example, all fricatives may be merged into one class for the low-frequency band, and not so for the high frequency band. In fact, it is desirable that the sub-band super-classes be different so that the identity of the detailed phone classes can be reconstructed on the full-band level. The main motivation for using sub-band categories is that energies from different acoustic features (e.g., fricatives, vowels) can manifest themselves in different frequency bands. Hence, some bands may be more useful than others for discriminating among particular categories, whereas other bands may be devoid of sufficient information for such discrimination. In other words, different bands may be "experts" for different phonetic features.

In order to determine the relative accuracy of phone discrimination for each sub-band, band-limited phone recognition was performed using the systems described in Sections 3.6 and confusion matrices[1] created, as described in Section 4.1.1. Figure 6.1 shows the diagonals of the confusion matrices for the sub-band, multi-band, and the full-band systems. The figure illustrates the phone discrimination accuracy differences among bands. For example, phone [ow] is recognized most accurately in band 2 (72% accuracy) and least accurately (36% accuracy) in band 4, and is mostly confused with phones [h#] and [ay]. Phone [f], on the other hand, is recognized correctly as much as 53% and as little as 21% of the time

[1] The full confusion matrices are in Appendix E.

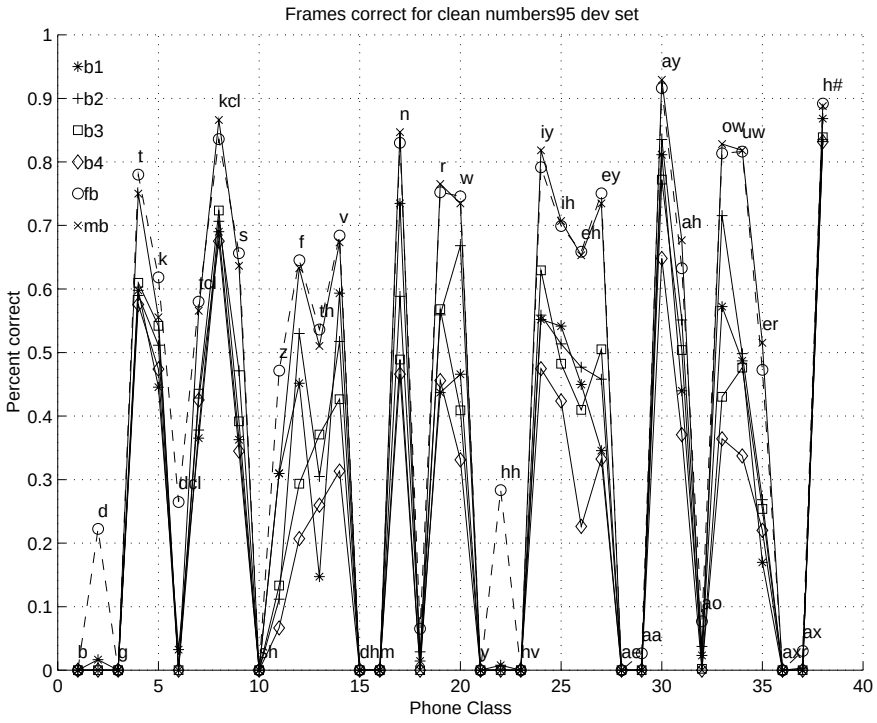

Figure 6.1: Frame accuracy for sub-band, multi-band, and full-band systems on the Numbers95 development set.

in bands 2 and 4 respectively, and is confused with many phones, such as [h#], [n], and [s]. Figures 6.2 and 6.3 expand on Figure 6.1, showing the average posterior probability estimates for the Numbers95 development set for phones [ay] and [th], respectively, for each sub-band. Figure 6.2 shows that most bands do well for [ay], though band 2 is the best. Figure 6.3 shows that band 3 is the best estimator for the phone [th] and band 1 is the worst, since it confuses [th] (phone index 21) with [t] (index 5) and [h#] (the silence phone – index 55). It is perhaps not judicious to train a system for band 1 to discriminate between [th], [t], and [h#], when they are almost indistinguishable. The goal is to find such indistinguishable classes and merge them into a super-class for each sub-band, and to develop more accurate statistical models by focusing on the most prominent features present in each band.

Note that the proposed sub-band super-classes are different from the previous work presented in Section 2.2.3 in that they are not necessarily associated with particular place or manner-of-articulation classes. The merging of phone classes into sub-band super-classes is determined in a data-driven way as opposed to using *a priori* linguistic information based on the articulatory considerations.

The approach in this chapter is also different in many aspects from the work of Bitar and Espy-Wilson [6]:

- The goal of this thesis is the recognition of continuous spontaneous speech, as opposed to broad phonetic-category recognition.
- Bitar and Espy-Wilson extract specific features (e.g., zero crossing rate, mel-cepstral)

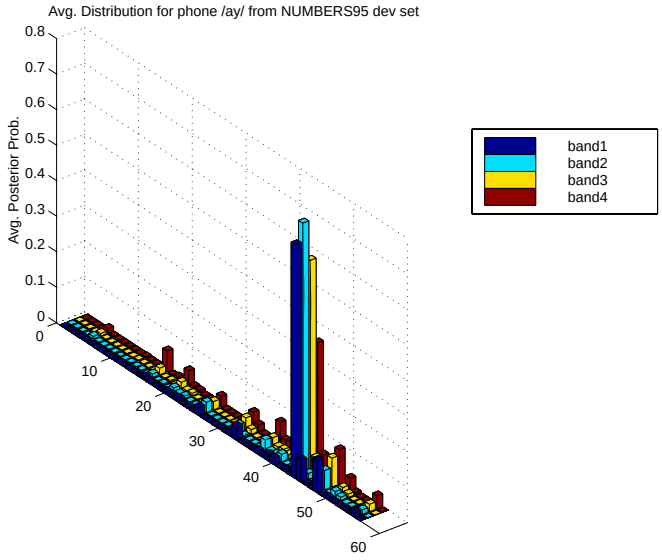

Figure 6.2: Average posterior probability estimates for all sub-bands for phone [ay] calculated on the Numbers95 development set.

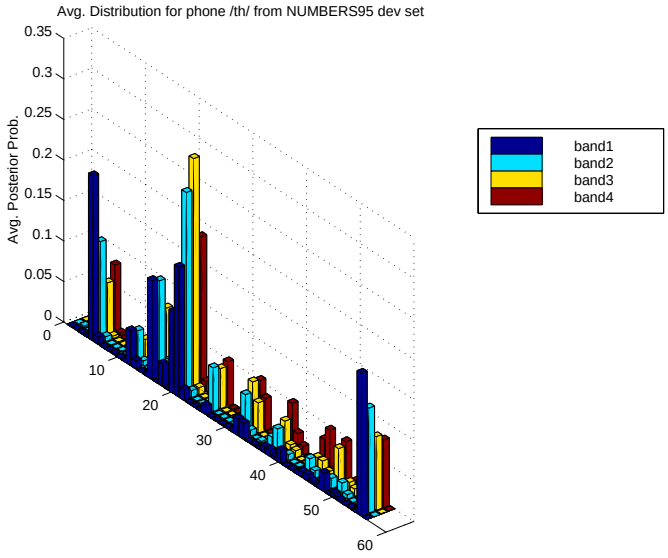

Figure 6.3: Average posterior probability estimates for all sub-bands for phone [th] calculated on the Numbers95 development set.

from multiple frequency bands to identify particular articulation features. For example, to determine *syllabic* features, they analyze the peak in 0.64-2.8 kHz energy and the peak in 2-3 kHz energy. This approach requires specialized feature processing from multiple frequency regions. Besides being computationally expensive (since one band may be processed multiple times, once for each articulation feature), there is also a potential problem with the explosion of feature dimensionality.

- The basic assumption in Bitar and Espy-Wilson's work is that articulatory features are the desired units for speech perception. In this work, by contrast, narrow-band features are the focus of study. The class boundaries and membership of the reduced sub-band classes may or may not be the same as the articulatory feature classes. For example, on average, the vowel [i] has energy maxima at roughly 280 Hz, 2250 Hz, and 2890 Hz [76], so the features for [i]-ness may be spread across multiple narrow-band classes (see Figure 6.4).

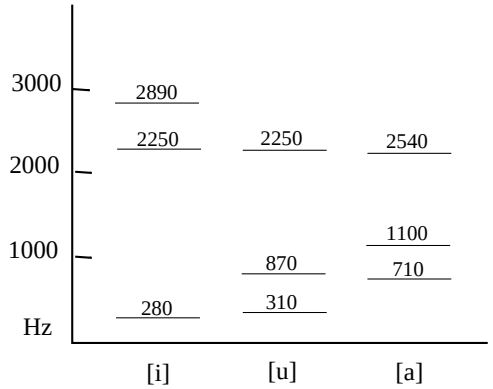

Figure 6.4: The frequencies of the first three formants in three American English vowels (from Ladefoged 1993).

6.2 Using A Mutual Information Criterion

As mentioned in the previous section, in order to determine the relative accuracy of phone discrimination for each sub-band, band-limited phone recognition was performed using the systems described in Sections 3.6 and confusion matrices created, as described in 4.1.1. Various options were considered for clustering the phone classes. Mutual information (MI) [28] between the input and the output phone classes was chosen as the function for determining the class merges. As explained in Section 3.6, MI expresses the bits of information transmitted in a system. It was verified empirically that when two highly confusable classes are combined into a super-class, MI either increases slightly or stays the same. Reducing the number of classes and yet maintaining the same amount of information transmitted in the system appeared a reasonable first option, given that MI is easy and fast to calculate. The hope was that the pattern of class merges would be different in each band so that the phone could still be distinguished even when the classes were merged.

The simple algorithm presented below calculated the change in MI for every possible

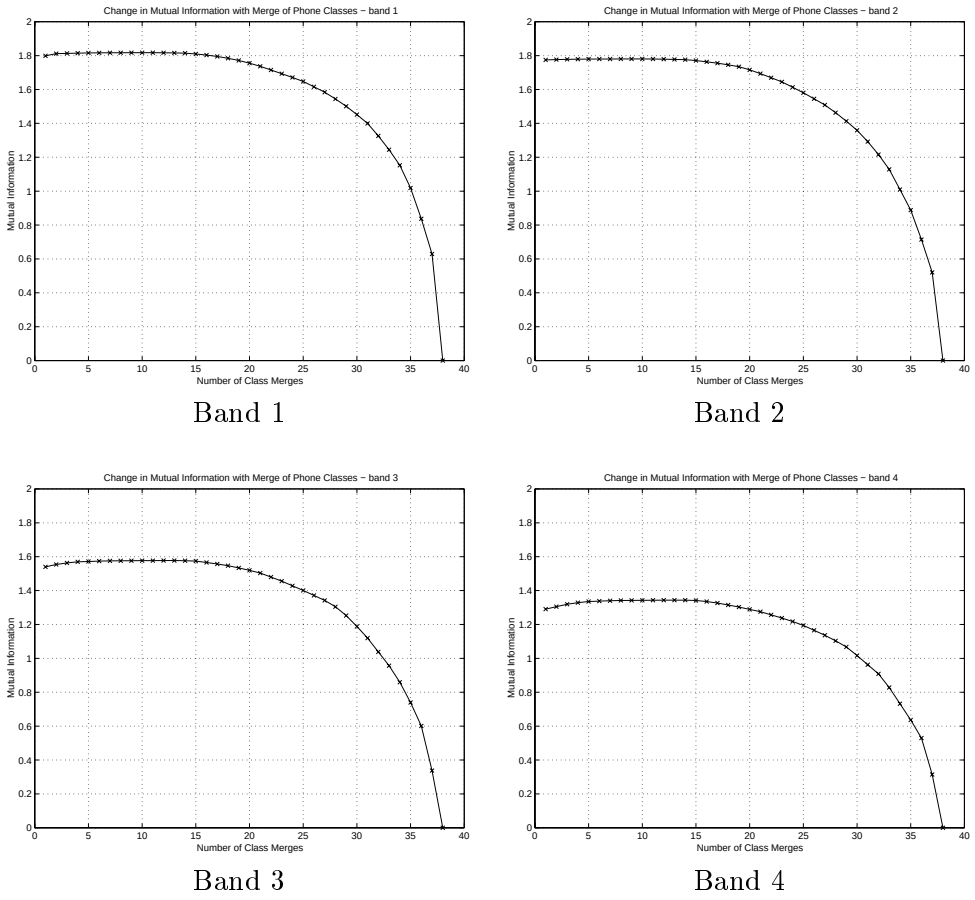

Figure 6.5: Change in mutual information for bands 1 through 4 as phone classes are merged.

pair of class merges in each sub-band and then chose the merge with the highest gain (or the least loss) in MI. This greedy algorithm was run for each sub-band independently of others:

For every pair of classes C_i and C_j
    Calculate the change in MI (Δ MI) for merging C_i and C_j
Choose C_i^* and C_j^* with the highest Δ MI
Merge class C_i^* and C_j^*

This algorithm was implemented using *MATLAB* [61] and run on the confusion matrices for each sub-band system. Figure 6.5 shows that MI stayed constant for roughly 13 to 18 class merges and then decreased. It was expected that a different set of classes would be merged in different bands. Upon examining the merged classes (shown in Table F.1), it became apparent that the same set of classes were merged in every band. More precisely, if merging was stopped just before the MI fell below the initial MI level, in every class *almost identical* classes were merged into one very large super-class. If the super-classes in all bands

are almost identical, it would not be possible to ultimately reconstruct the identity of the collapsed phones. These super-classes are shown in Table 6.1.

Band 1	aa	ae	ax	axr	b			dh	g		hv	m	sh	y		
Band 2	aa	ae	ax	axr	b	d	dcl	dh		hh	hv		sh	y		
Band 3	aa	ae		axr	b	d	dcl	dh	g	hh	hv	m	sh	y		
Band 4	aa	ae	ax	axr	b	d	dcl	dh	g	hh	hv	m	sh	y	ao	l

Table 6.1: The large super-class formed in every sub-band after merging using mutual information as the criterion.

Not only were the same classes merged in each band, but more disturbingly, they all appeared to be the classes with the smallest priors (see Table 6.2). Furthermore, the collapsing of phone classes with significant priors appeared to lower the MI significantly. It is not surprising that classes with small priors were the ones which were most easily confused, as there existed too few learning examples for adequate training. The following two questions arise: (1) would the collapsing of infrequently occurring and highly confused phones in every band decrease the word error, or would the inability to discriminate the phones grouped together in the super-class ultimately hurt word recognition?, and (2) is mutual information the right metric for choosing classes to merge?

Regarding the first concern, grouping similar phones in every sub-band class is essentially equivalent to grouping the same phones in a full-band system. Computationally, it would be simpler to answer the questions in the context of a full-band system first, and if the results were promising, repeat the experiment in a multi-band system.

A full-band system, similar to the one explained in Section 3.6, was trained, with the exception that the MLP had 400 hidden units instead of 1000. A confusion matrix was created for the full-band system, and reduced classes, according to the algorithm presented above, were chosen.

The change in MI was similar to the sub-band cases. The MI continued to stay relatively unchanged and after approximately 15–18 merges started to decrease rapidly. All class merges are reported in Table 6.3. The 16 least-frequent most-confused phones were combined into one large super-class after the first 15 merges (see Figure 6.6). This super-class was composed of the following phones: **aa, ae, ao, ax, axr, b, d, dcl, dh, g, hh, hv, l, m, sh,** and **y**. Note that these classes are identical to the ones chosen in the multi-band case, as shown in Table 6.4.

A new full-band system with the reduced set of classes was trained. The phone labels in the training set were changed so that the label for the super-class was substituted for every occurrence of the above 16 classes. An MLP was trained on the new labels and the recognition lexicon was changed similarly as well. Word recognition was performed on the merged-phone system for the Numbers95 development set. The word error rate was slightly poorer (8.9%) than the unmerged full-band (8.5%), though the difference is not statistically significant. Examining some differences in the error patterns, it is observed that, for example, the word "forty" has been mis-recognized as "thirty," since [dcl], [d], and [ao] have been merged into the super-class phone, and the pronunciations have become

Phone	Prior	Phone	Prior
dh	0.000009	kcl	0.010225
m	0.000013	th	0.019411
axr	0.000018	eh	0.022381
g	0.000032	ih	0.024639
y	0.000036	v	0.024870
b	0.000055	ey	0.028704
ae	0.000115	w	0.029674
hv	0.000180	tcl	0.033595
sh	0.000207	t	0.035225
hh	0.000586	uw	0.041834
d	0.000831	r	0.043243
dcl	0.000854	f	0.044065
l	0.001279	ah	0.044231
ax	0.001367	iy	0.052623
aa	0.001394	s	0.061791
k	0.004064	ow	0.070081
er	0.004406	n	0.079776
ao	0.006179	ay	0.087447
z	0.009874	h#	0.214665

Table 6.2: Numbers95 phonetic prior probabilities for the development set.

Full-band Merges							
1	ao	→	l	20	eh	→	ih
2	aa	→	l	21	z	→	tcl
3	ax	→	l	22	ah	→	w
4	l	→	dcl	23	ey	→	iy
5	dcl	→	d	24	f	→	s
6	hh	→	d	25	tcl	→	t
7	sh	→	d	26	ih	→	r
8	hv	→	d	27	uw	→	k
9	ae	→	d	28	ow	→	b
10	d	→	b	29	s	→	t
11	y	→	b	30	w	→	b
12	g	→	b	31	n	→	k
13	m	→	b	32	iy	→	r
14	axr	→	b	33	ay	→	b
15	dh	→	b	34	h#	→	t
16	kcl	→	k	35	r	→	k
17	er	→	ih	36	k	→	b
18	th	→	z	37	t	→	b
19	v	→	k				

Table 6.3: An ordered list of phone class merges in the full-band.

Band 1	aa	ae	ax	axr	b			dh	g		hv	m	sh	y		
Band 2	aa	ae	ax	axr	b	d	dcl	dh		hh	hv		sh	y		
Band 3	aa	ae		axr	b	d	dcl	dh	g	hh	hv	m	sh	y		
Band 4	aa	ae	ax	axr	b	d	dcl	dh	g	hh	hv	m	sh	y	ao	l
Full-band	aa	ae	ax	axr	b	d	dcl	dh	g	hh	hv	m	sh	y	ao	l

Table 6.4: The large super-class formed in the full-band, compared to the ones formed in every sub-band after phone-class reduction, using mutual information as the criterion.

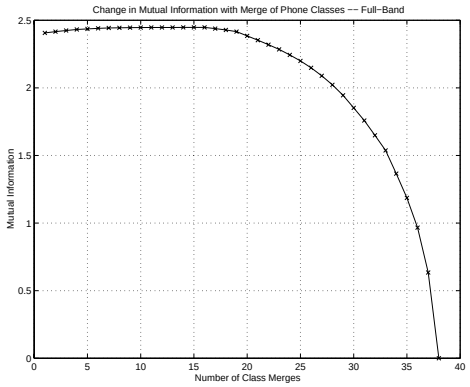

Figure 6.6: Change in mutual information for the full-band as phone classes are merged.

indistinguishable. It appears that although merging the most highly confusable classes may make the discrimination task easier, it increases the recognition word error rate since confusion among the word models in the lexicon also increases.

6.2.1 Discussion

When mutual information is used as the criterion for sub-band phone-class merging, the infrequently occurring phone classes are chosen. It appears that collapsing the infrequently occurring and highly-confused phones does not decrease the word error rate. In fact, as had been suspected, the inability to discriminate the phones grouped together in the superclass ultimately hurts the word recognition because the confusability in the lexicon increases. The pattern of phone merges is the same for the full-band and the sub-band systems.

Two questions were posed regarding MI as a criterion: (1) would the collapsing of infrequently occurring and highly confused phones actually decrease the word error? and (2) is mutual information the right metric for choosing classes to merge? The answer to the first question seems negative, because of the lexicon effects. Let us turn to answering the second question: is mutual information the right metric for choosing classes to merge?

One problem with MI may have been that the priors of each class were not represented in the decision criterion. This caused the merging to concentrate on the highly confused phones with extremely low priors, which are less important for recognition than other phones. Furthermore, if the same phones are merged in every band, as observed, reconstruction of

their identity is not possible. The most important problem, however, may have been that the decision to merge the classes according to MI was local to every sub-band. Perhaps classes should be chosen according to a global criterion that considers sub-band-phone merging with respect to the overall error reduction. The next section examines such a criterion for collapsing classes. In the absence of any data to the contrary, we are forced to conclude that MI is not a suitable criterion to use for choosing classes to merge.

6.3 Using A Global Discriminator

Using a global discrimination criterion allows the output of *all* sub-bands to influence the choice of classes to merge in *each* sub-band and also incorporates the prior probabilities of the phonetic classes. Any function that is used for merging the sub-band probability streams (as in Equation 3.6) may also be used as a global decision criterion for choosing phone classes to collapse. Collapsing sub-band classes using the multi-band merger as the global criterion may be done using a simple algorithm that tries to find the best pair of class merges in order to maximize the overall frame accuracy:

Initalization:
    Train a multi-band merging function (MM) to estimate the overall probabilities, given the sub-band probabilities
Repeat until all classes are collapsed:
    For every sub-band k
        For every pair of classes C_i^k and C_j^k
            Combine the probabilities of classes C_i^k and C_j^k for all training patterns
            Train a new multi-band merging function MM_{ij}^k
            Calculate $\Delta FrameAccuracy$ for MM_{ij}^k
    Choose C_i^{*k} and C_j^{*k} with the highest $\Delta FrameAccuracy$
    Merge class C_i^{*k} and C_j^{*k}
    Train a new multi-band merger $MM = MM_{ij}^k$

Using an MLP as the global multi-band merging function may yield the best optimization. However, insurmountable computation and time requirements make this an infeasible task, since training an MLP (on the full, unmerged, phone probabilities) on the current sub-band system takes approximately two hours, and there are roughly 350,000 configurations[2]. Due to practical considerations (e.g., hope of graduating in the current millennium), I decided to use a simpler merging function. One such simple function is the nearest-neighbor algorithm. Means and standard deviations of the features were calculated for the sub-band probabilities in the the training data. In the nearest-neighbor algorithm, class membership

[2]This approximation is calculated using a simulation of the phone-class collapsing pattern. The total number of tested configurations (i.e., tested phone pairs) is not constant, since it depends on previously chosen phone pairs. One such pattern might look like: $4\binom{40}{2} + (3\binom{40}{2} + \binom{39}{2}) + \cdots + \binom{3}{2}$. Note that 40 phones out of the 56 appear in the Numbers95 training set.

Percent Frame Error For Relative Entropy Error Criterion		
Baseline	Merged	Relative Improvement
28.2%	27.0%	4.3%

Table 6.5: The decrease in frame error as phone classes are merged according to a relative entropy error criterion using a nearest-neighbor algorithm.

for each test feature vector is decided, based on distance to the class means. This distance was calculated using the relative entropy distance metric:

$$\sum_i p(x_i) log \frac{p(x_i)}{q(x_i)}$$

The feature vectors for the global phonetic discrimination are the phonetic probability distributions of the sub-band systems. As demonstrated in Figure 6.7, class centers are the means of the sub-band phonetic probability distributions. It is important to keep in mind that the *sub-band phonetic probabilities* serve as *input features* for estimating the *overall phonetic probabilities* and merging (or collapsing) the *sub-band classes* reduces the input vector dimensionality for the global discrimination task.

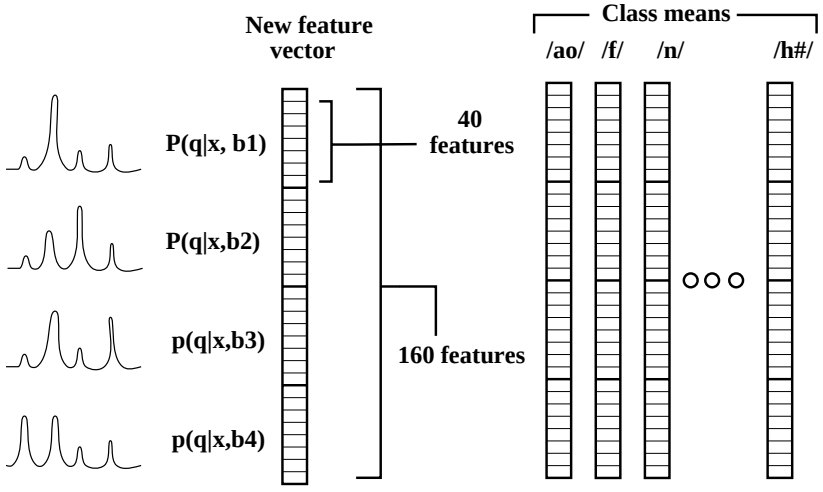

Figure 6.7: Visual demonstration of the nearest-neighbor calculations.

To reduce the turnaround time for this experiment, the amount of data was limited to half of the cross-validation set (roughly 30,000 frames) for training (finding the means). The remaining half of the cross-validation set was used to choose which sub-band classes to collapse.

The algorithm described above was run using the second half of the cross-validation data. Table 6.5 shows the final frame scores. The percentage reduction in the number of classes was 28.8%. A merged configuration at the minimum frame error point was chosen. The new number of classes for this configuration was 114, a 28.8% reduction

The Combined Sub-Band Classes			
Band 1	Band 2	Band 3	Band 4
[kcl er]	[f ay]	[v ow]	[kcl uw]
[t ax]	[n h#]	[tcl ay]	[dcl ay]
[n aa]	[eh uw]	[kcl f]	[iy aa]
[hv ah]	[v ey]	[th n]	[l ow ax]
[k aw]	[ih ow]	[ao ax]	[hh er]
[s hh l]	[kcl iy]	[dcl uw hv hh s]	[tcl hv]
[dcl w]	[k th]	[l aa k]	[t aw]
[v ay]	[ao ax]	[t aw]	[k ax]
[th h#]	[t hh]	[z iy]	[z eh]
	[s aw]		[f v]
	[tcl hv]		
	[aa ah]		

Table 6.6: The table shows the combined sub-band classes. Combining the classes was done according to the relative entropy error criterion in a nearest-neighbor classification algorithm. The classes which were not combined are not listed.

from the original 140[3]. It is interesting to note that upon collapsing classes, frame error decreases significantly ($p < 0.0005$). Since the priors are taken into account[4], infrequently occurring highly confusable classes are *not* merged first (as opposed to the MI merging criterion). Examination of the merged-phone patterns indicates that although the merges do not make sense phonetically (alas!), some appear mathematically convincing according to the confusion matrix. For example, phone classes [f] and [ay] are merged in band 1 using relative entropy distance, and as in the confusion matrix (included in Appendix E) for band 1, these two high-prior classes are quite often confused with one another. The summary of the merged classes in each band are reported below.

At the outset, the phone classes in each band, in increasing order of occurrence, were: [dh], [m], [axr], [g], [y], [b], [ae], [hv], [sh], [hh], [d], [dcl], [l], [ax], [aa], [k], [er], [ao], [z], [kcl], [th], [eh], [ih], [v], [ey], [w], [tcl], [t], [uw], [r], [f], [ah], [iy], [s], [ow], [n], [ay], [h#]. Table 6.6 lists the combined classes for each sub-band.

Word recognition was run on the Numbers95 development set. The nearest-neighbor distances calculated for each test vector were made into posterior probabilities by exponentiation (e^{-dist}). The global posterior probabilities were then decoded using the decoder Y0, as described in Section 3.6.

A comparison of the baseline multi-band results (no classes merged) with merged multi-band results are reported in Table 6.7. To summarize, almost 30% of the sub-band classes have been merged, making the global phonetic classes more easily distinguishable and re-

[3] 84 classes out of the total 224 had zero priors.

[4] The priors are represented through calculating the frame error over the test set – obviously, the less frequently occurring classes have less of an effect.

Percent Class Reduction	Baseline WER	Merged WER
28.8%	11.6%	11.5%

Table 6.7: The word recognition error rate comparison for baseline and collapsed multi-band systems on the Numbers95 development set. The frame-level merging function is the nearest-neighbor algorithm using relative entropy distance criteria.

Percent Word Error		
Condition	unmerged	merged
Volvo noise (SNR=15)	11.7	11.9
Volvo noise (SNR=5)	12.4	12.9
Pink noise (SNR=15)	38.3	38.3
Reverberation	45.7	43.1

Table 6.8: Word error rate for Numbers95 development set using relative entropy-based nearest-neighbor probability estimator in degraded speech conditions.

ducing the overall frame error rate by 4.3% relative. The word-recognition error rate on the Numbers95 development set is essentially unchanged, marginally decreasing from 11.6% for the baseline to 11.5% for the merged system.

6.3.1 Degraded Speech Conditions

In degraded speech conditions distinguishing among different phonemes becomes more difficult. Hence, if collapsing confused phone classes decreases the confusion in discriminating phone classes, this advantage should help improve performance in degraded noisy conditions as well. Reverberant speech, as well as speech with Volvo and pink noises (from the NOISEX-92 database) were tested. The system was trained on clean speech and tested on reverberant and noisy speech. The results are shown in Table 6.8.

The performance of the merged system was similar to that of the the non-merged system for Volvo and pink noise. The observed differences were not statistically significant. For reverberant speech, however, the merged system's performance was significantly better. This result may be due to frequency-dependent smearing of energy patterns in reverberated speech, making the pattern of confusion among phones particularly dependent on the sub-band frequency region.

6.3.2 Discussion

The main intuition in collapsing sub-band phone classes was that some bands do poorly in discriminating among certain phones, thereby increasing the overall randomness and error in the phonetic probability estimation. It was hoped that by making the sub-band local discrimination task easier, the overall frame accuracy would improve and, as a consequence,

word recognition error rate would thereby decrease. It was observed that as the sub-band classes are merged, the frame accuracy decreases and the word recognition accuracy improves slightly. More specifically, almost 30% of the sub-band classes were merged, making the global phonetic classes more easily distinguishable and reducing the overall frame error rate by a relative 4.3%. The word-recognition error rate on the Numbers95 development set changed marginally from 11.6% for the baseline to 11.5% for the merged system on the Numbers95 development set. The performance of the merged system was not any better for Volvo- and pink-noise conditions, but was significantly better for reverberant speech. This result may be due to frequency-dependent energy smearing in reverberated speech, making the pattern of confusion amongst phones particularly dependent on the sub-band frequency region.

6.4 Summary and Conclusions

This chapter has focused on deriving reduced phonetic classes for each sub-band. It was motivated by the observation that some frequency bands contained more information for distinguishing particular acoustic classes. I hypothesized that merging the most confusable phones in each sub-band would obviate the need to discriminate among them and therefore improve the overall accuracy of the multi-band system.

The first criterion chosen for collapsing the phone classes was mutual information, applied locally to the confusion matrix of each sub-band. This function appeared to be sub-optimal. The same phones, ones with the lowest priors, were merged into one large all-encompassing super-class in every sub-band and therefore, the discrimination of these phones was no longer possible on the over-all level. For a similar phone-class reduction for a full-band system, the word error rate increased due to the increased similarity of the word pronunciations in the lexicon.

Next, a nearest-neighbor algorithm with a relative entropy error criterion was applied as the global merging function. For clean speech, it was observed that almost 30% of the sub-band classes were merged, making the global phonetic classes more easily distinguishable and reducing the overall frame error rate. The word recognition error rate on the Numbers95 development set also decreased from 11.6% for the baseline to 11.5%, although the difference is not statistically significant. The performance of the merged system was not better for Volvo- and pink-noise conditions but was significantly better for reverberant speech.

The observed results for clean and reverberant speech seem particularly interesting. The input feature dimensionality for the global discrimination task was decreased by almost 30%, yet the discrimination accuracy improved. This validates our original hypothesis that by combining the confused phone classes, the overall phonetic discrimination task becomes easier.

Chapter 7

Combining the Multi-Band and Full-Band Streams

It has been shown that multi-band ASR performance is similar to that of the full-band paradigm on continuous speech, and significantly better in many noise conditions (especially for narrow-band noise) when the multi-band streams are combined on a frame level [135, 13]. It is less clear whether the multi-band paradigm can be used to significantly improve recognition accuracy for clean speech. Furthermore, if indeed such improvements can be confirmed, it is desirable to understand the reasons for this effect.

Research at ICSI [148, 100] and elsewhere has shown that often combining an experimental system with a conventional speech recognition system, where the two have differing error patterns, leads to improvements in the overall system performance. Other researchers have also experimented with boosting and mixture of expert algorithms [39, 26, 127] in order to develop multiple systems with different statistical characteristics and succeeded in reducing the word error rate.

Sections 7.1 and 7.2 discuss experiments with combining full-band and multi-band probability streams for clean and reverberant speech, respectively. In section 7.3, results of the error analysis for these two systems are reported. Section 7.4 discusses these experiments repeated on a system with improved phonetic labels. Section 7.5 discusses the results.

Percent Frame and Word Error for Clean Numbers							
Error	b1	b2	b3	b4	MB	FB	Mgd
Frame	40.3	37.9	43.2	50.4	23.1	23.4	30.1
Word	33.7	24.9	34.4	47.8	8.3	7.9	6.3

Table 7.1: Frame and word error, in percent, for sub-bands 1 through 4 (b1 through b4), multi-band (MB), full-band (FB), and the merged (Mgd) systems for clean natural numbers.

Percent Word Error for Reverberant Numbers						
CW	b1	b2	b3	b4	MB	Mgd
9	68.1	61.2	68.7	76.2	39.9	30.3
17-11	66.2	60.5	67.9	75.9	38.2	29.5
17	65.7	59.0	67.4	75.7	42.8	31.6

Table 7.2: Percent word error for bands 1 through 4, multi-band (MB), and merged (Mgd) systems for reverberant natural numbers for different sizes of feature-input context-windows (CW). The baseline FB system has a word error rate of 32.2%.

7.1 Experiments with Clean Speech

The baseline full-band and multi-band systems used for the experiments in this section are discussed in Chapter 3. As already observed, the word error rate of the multi-band and the full-band system are similar. The question remains whether multi-band information can be used to enhance baseline full-band performance.

The probability streams of the full-band and the multi-band systems were merged by simply multiplying the likelihoods from each system. This simple merging mechanism was chosen because it has been observed at ICSI [148, 64, 71] that when the probability streams from the two systems are highly accurate, multiplying the probabilities often works better than training a merger MLP. Additionally, multiplication is a simple operation and does not increase the number of parameters of the overall system.

The resulting probability stream was decoded using Y0. The word error rate of the combined system decreased to 6.3%. In other words, errors were reduced by 20%.

Note that the combined system has roughly twice as many parameters as the other systems, so it is possible that doubling the number of parameters in the full-band system might produce a similar improvement. A full-band system with twice as many parameters as the baseline (with 2400 hidden units in the MLP instead of 1000) was trained. The word error rate of this system was 8.9%. It is unlikely that the improvement in the combined system was due merely to an increase in the number of parameters.

Thus, it appears that combining multi-band and full-band systems significantly[1] reduces the word error rate for the test set over either system alone, or when compared to a version of the full-band system with an extended parameter set.

Section 7.3 reports on the analysis of the error patterns of the two streams, in an attempt to understand how they might counteract each other.

7.2 Experiments with Reverberant Speech

Similar experiments were performed on digitally reverberated versions of the data. The reverberant data set was generated as described in Section 7.2.

[1]For this size test set, an absolute difference of more than 1.1% is considered statistically significant (using z-scores on a binomial distribution).

Natural reverberation usually affects low frequencies more than high frequencies, since most common room boundary materials are less absorptive at low frequencies, leading to longer reverberation times and more smearing of the spectral information at those frequencies. The baseline system had a feature input window of nine (four frames of context in the past and the future) for all frequency bands. It was decided to increase the size of the feature-input window for the low-frequency sub-bands. More specifically, the input window size for the lowest band was doubled, which made it roughly equal to the length of a syllable (200 ms). The size of the neighboring higher frequency windows were decreased by two frames, so the "pyramid" system had 17, 15, 13, and 11 frames of input for bands 1 through 4, respectively. To study the effects of overall window size increase, four sub-band systems, each with 17 frames of input, were trained. The size of the overall system was kept constant by decreasing the number of hidden units. More precisely, the number of hidden units were changed from [497, 497, 372, 372] for the nine-frame sub-band systems to [272, 297, 398, 428] for the pyramid system, and [272, 272, 331, 331] for the 17-frame sub-band systems. The word error rate for each sub-band, the multi-band, and the merged systems are reported in Table 7.2. The word error rate for the full-band system is 32.2%, which is significantly better than each of the multi-band systems (38.2% – 42.7%). However, merging the less accurate multi-band stream with the full-band stream still *improves* the overall word error rate (29.5% – 31.6%). Although the word error rate of each 17-context-frame sub-band system was less than that of the pyramid system, this was not true when the sub-bands were merged together or with the full-band system. In short, adding the multi-band pyramid system information to that of the full-band system reduces the word error rate from 32.2% to 29.5%. This is an error reduction of 8%.

For the sake of completeness, the pyramid experiments were also repeated for clean speech. Neither of the conditions significantly changed the word error rate from the baseline setup. For instance, word error rate for the pyramid windows was 6.5%, in comparison with the 6.3% for the nine-frame window. Thus, it appears that using the extended windows, particularly the pyramid case, improves word error rate for reverberant speech without substantially hurting performance for clean speech.

7.3 Analysis of Multi-Band and Full-Band Error Patterns

In addition to simply observing a reduction in word error rate, it is also important to try to understand why such a reduction occurs. One explanation, inspired by the combination-of-experts community, is that the error rate decreases when experts with different characteristics (preferably orthogonal) are combined [63]. It is desirable to understand how the full-band and multi-band recognizers are different, and how this difference affects performance [148]. If possible, it is desirable to associate these differences with phonetic content: are there particular phones or features that one system is better at discriminating than the other?

Phone recognition[2] on both the full-band and the multi-band systems were performed. Confusion matrices for phone classes were generated, both for the phone recognition results and for the frame-by-frame comparison of the phone decoding path. The main difference is

[2]The phone recognition system was built together with Brian Kingsbury at ICSI.

that the latter gives more weight to long phones, since the classification for every frame is counted.

Detailed statistics for every phone token were also generated – whether both systems were right or wrong and how this affected the merged system's classification. The summarized results of this analysis are in Table 7.3.

		Mgd $\surd$	Mgd $\times$
FB $\surd$	MB $\surd$	86.8%	0.2%
	MB $\times$	2.2%	1.3%
FB $\times$	MB $\surd$	1.9%	1.2%
	MB $\times$	0.3%	6.1%

Table 7.3: A summary of the analysis on the recognized phone string for the full-band (FB), multi-band (MB), and the merged (Mgd) system as compared to the correct results. $\surd$ means the phone classification of that system was correct, $\times$ means that the phone classification was incorrect.

The following were observed:

- It rarely occurs that the classification of the merged stream is incorrect when both full-band (FB) and multi-band (MB) streams have the correct phone classification (only for $0.2\% = \frac{0.2}{87.0}$ of such phone tokens). Conversely, it is also unusual for the co-occurring errors of the two streams to be corrected by merging (only for $4.7\% = \frac{0.3}{6.4}$ of such phone tokens).

- Nearly all of the correctly classified phones in the merged stream were actually correct in both streams (95.1% of the correctly classified tokens). Of the phones correctly classified by the merged system, which were also correct in one stream only, roughly half were correct in each stream (2.5% and 2.1% for MB and FB respectively).

- Most of the phones that were incorrectly classified in the merged stream were incorrect in both steams (69.6% of the incorrectly classified tokens). Of the phones incorrectly classified by the merged system, which were also incorrect in one stream only, roughly half were incorrect in each stream (13.9% and 14.3% for MB and FB respectively).

Neither system seems to be significantly better nor worse than the other. In fact, the pattern of errors seems very similar, except for a small minority (6.6%) of frames. When the two systems are combined, the correct classification of a stream prevails more often than its misclassification. Not shown in the table are the following observations:

- Most of the MB and FB phone errors are identical (76% of the misclassified tokens).

- Examining the errors for each phone class, it is observed that for [sil] and [tcl] (t-closure), the MB system is correct significantly more often than the FB system. The reverse is true for the vowel [ao].

- Examining the frame-based confusion matrices, it is observed that the MB system is significantly more accurate in classifying [sil], [r], [w], and [tcl] phones. The FB system, on the other hand, is significantly more accurate in classifying [ao], [n], [iy], [ah], [f], and [s]. Research on the acoustic cues for the perception of liquids and glides has shown that the duration of the formant transitions provides the essential cue for these speech sounds [106]. For discrimination of vowels, however, simultaneous identification of the location of the first two formants is necessary. Perhaps the divide-and-conquer MB strategy makes it difficult for the fine across-sub-band information analysis necessary for accurate discrimination of vowels, whereas the transition pattern becomes more apparent, explaining better liquid and glide discrimination.

7.4 Improving the Word Error Rate

In Section 3.1.5, it was mentioned that all the systems in this thesis are trained on phonetic labels obtained from hand-segmentation, and therefore both the baseline and the experimental system performance are about 1-2% lower than other published works on Numbers95 from ICSI.

Since combining the full-band and the multi-band systems showed the most pronounced improvement in the word recognition error rate, I decided to repeat this experiment with improved phonetic labels to demonstrate that the multi-band system is competitive with other methods. This section reports on this repeated experiment.

7.4.1 RASTA-PLP Features

The baseline full-band system was improved first. The baseline system setup was identical to the one described in Section 3.6, except that the phone labels were re-estimated by applying two iterations of embedded alignment. The MLP was then trained on these improved labels. Also, an improved lexicon (after one iteration of forced alignment, as described in section 3.3.7) was used both for the embedded alignment and recognition. The word recognition error rate of this full-band system, using the improved lexicon, on the Numbers95 development set was 6.1%, which is significantly better than the baseline of 7.9% for the system with the hand-segmented phone labels.

The multi-band system parameters were chosen to be the same as the one described in Section 3.6. The MLPs for the sub-bands were trained given the sub-band RASTA-PLP features as input and the improved phonetic labels (from the full-band system) as training targets. A merger MLP (with 300 hidden units and a context window of one) was trained to combine the posterior probabilities of the sub-band systems. The combined probability stream was decoded using the Y0 decoder and the improved lexicon. The word error rate of the multi-band system on the Numbers95 development set using the MLP merger trained was 7.6%

Next, the probability estimates of the full-band system were combined with the multi-band system by multiplying the scaled likelihoods. The resulting stream was decoded with Y0 and the new lexicon. The word error rate on the Numbers95 development set was 5.5%. It appears that even though the aligned multi-band system performs more poorly than the

full-band system alone, the overall performance improves significantly in combination. The word error rate has been reduced by 10%. The results are summarized in Table 7.4.

RASTA-PLP Features			
	Full-Band	Multi-Band	Combined
Number of Parameters	209,000	293,000	502,000
Word Error Rate	6.1%	7.6%	5.5%
Percent Improvement	10%	28%	–

Table 7.4: The word recognition error rate and the percent improvement of the Combined system reported for the Numbers95 development set for the improved label full-band, multi-band, and combined systems with RASTA-PLP features.

7.4.2 PLP Features

Most recently, experiments by Hagen and Bourlard [49] have suggested that sub-band PLP features may be better suited than sub-band RASTA-PLP features for multi-band recognition, at least for some bands. A potential explanation for this result is that RASTA-PLP processing attempts to make up for adverse channel effects by filtering the signal and emphasizing the transitions. This filtering may cause some loss and smearing of speech. When the channel variabilities are pronounced, this loss is outweighed by the channel robustness gains. However, when considering sub-bands, channel variabilities may be small and the additional filtering in RASTA-PLP processing may not be advantageous. At least for the Numbers95 corpus, PLP processing seems to produce somewhat fewer errors compared to RASTA-PLP processing for each sub-band system, as shown in Table 7.5.

Full- and multi-band systems, identical to the ones described in Section 7.4.1, were trained on *full-band PLP* and *sub-band PLP* features, respectively. The phone labels in the full-band system were re-estimated by applying two iterations of embedded alignment. The full-band and sub-band MLPs were then trained on these improved labels. The word recognition error rate of this PLP-based full-band system, using the improved lexicon, for the Numbers95 development set was 6.0%. The word error of the multi-band system (sub-bands merged with a 300-hidden-unit-merger MLP) was 6.6%. Combining the streams of the two systems by simply multiplying the scaled likelihoods and decoding the combined stream resulted in a word error rate of 5.3%. Similar to the improved RASTA-PLP results in Section 7.4.1, the combination of the two systems resulted in a significant (12%) reduction of the word error rate. The results are summarized in Table 7.6.

7.4.3 PLP and RASTA-PLP features

According to word error rate on the Numbers95 database, it appears that PLP features are significantly better suited than RASTA-PLP features to a multi-band paradigm (6.6% vs. 7.6%) and yield comparable results for a full-band one (6.0% vs. 6.1%). The question

Frame Error			
Band	RASTA-PLP	PLP	Percent Improvement
b1	40.20	38.82	3.43
b2	37.40	35.00	6.42
b3	42.50	42.23	0.64
b4	49.50	48.59	1.84
Word Error			
Band	RASTA-PLP	PLP	Percent Improvement
b1	33.70	30.90	8.31
b2	24.90	23.00	7.63
b3	34.40	33.80	1.74
b4	47.80	45.50	4.81

Table 7.5: The word and frame error for sub-band systems trained using either sub-band PLP or RASTA-PLP features, on phonetically hand-transcribed labels. For an explanation of the system parameters see Section 3.6.

PLP Features			
	Full-Band	Multi-Band	Combined
Number of Parameters	209,000	293,000	502,000
Word Error Rate	6.0%	6.6%	5.3%
Percent Improvement	12%	20%	–

Table 7.6: The word recognition error rate and the percent improvement of the Combined system reported for the Numbers95 development set for the improved label full-band, multi-band, and combined systems with PLP features.

arises as to whether a combination of a multi-band PLP-based system with a full-band RASTA-based system would result in further improvements?

The results of this experiment are reported in Table 7.7. The RASTA-PLP-based full-band system described in Section 7.4.1 was combined with the PLP-based system described in Section 7.4.1 by multiplying the scaled likelihood streams and decoding the resulting stream. The error rate of 4.7% for the development test set is the lowest that we are aware of. For comparative purposes, Table 7.7 also reports experiments using the evaluation data set of the Numbers95 corpus. The reduction in word error rate of the combined system, compared to the full-band system, was 23% to 29%.

7.4.4 Fairness Comparisons

As observed in the previous section, significant improvements were observed when the two multi- and full-band systems were combined. Two questions arise: (1) Is the improvement

MB PLP Combined with FB RASTA-PLP				
Number of Parameters		Full-Band 209,000	Multi-Band 293,000	Combined 502,000
Development Set	Word Error Rate	6.1%	6.6%	**4.7**%
	Percent Improvement	23%	29%	–
Evaluation Set	Word Error Rate	6.6%	6.3%	**4.7**%
	Percent Improvement	29%	25%	–

Table 7.7: The word recognition error rate and the percent improvement of the Combined system reported for the Numbers95 development and evaluation set for the improved label full-band, multi-band, and combined systems.

solely due to the increase in the number of parameters? and (2) Is the improvement observed in Section 7.4.3 solely due to combination of two different feature sets?

The first concern was examined in Section 7.1 and is revisited here with respect to the systems with improved results. To examine whether the improvement was due to the increased number of parameters alone, a full-band system with 2400 hidden units (about 502,000 parameters) was trained with two iterations of forced alignment. The word error rate of this system was 5.9%. Again, the simple increase of number of parameters does not seem to be responsible for the significant improvement in word error rate.

Regarding the second concern, we combined the full-band RASTA-PLP-based system (described in Section 7.4.1) with a full-band PLP-based system (described in Section 7.4.1) by multiplying the scaled likelihoods and decoding the resulting stream. The resulting word error rate (5.1%) was higher than that obtained by combining the multi- and full-band systems (4.7%). Therefore, the improvement observed in Section 7.4.3 does not seem to have been solely due to combination of different feature sets, but rather, appears to have been a combination of both this effect and the differences between the multi- and full-band systems. The results are reported in Table 7.8.

Combining FB RASTA-PLP and FB PLP Features			
	RASTA-PLP FB	PLP FB	Combined
Word Error Rate	6.1%	6.0%	5.1%
Percent Improvement	16%	15%	–

Table 7.8: The word recognition error rate and the percent improvement of the Combined system reported for the Numbers95 development set for the improved label RASTA-PLP-based full-band and PLP-based full-band system combination.

7.5 Discussion and Conclusions

It was shown that multi-band ASR could be used to improve the speech recognition accuracy of natural numbers for clean speech when a multi-band information stream is used in addition to the full-band one. Specifically, this combination reduced the word error rate by 20%. These experiments were repeated with improved phone labels and a similar 23% reduction in word error rate compared to the improved full-band system was observed. The word error rate of 4.7% is the lowest value observed for the Numbers95 development set at ICSI and elsewhere.

It was observed that a similar combination method significantly reduced the error rate on reverberant speech. Extending the input window to the neural network probability estimators, particularly for the low-frequency bands, improved recognition for reverberant speech without substantially changing the performance for the clean case.

This combination approach may not simply be dismissed as a natural advantage of combining different experts. As the recent work of Wu [148] has also suggested, improving ASR systems by combination works best if each of the systems focuses on a particular aspect of speech. Such a combination method may not be as successful (considering the number of alloted parameters) if it does not take advantage of the inherent characters of speech processing [127].

Additionally, the error patterns of the full-band and multi-band paradigms were analyzed in an attempt to understand why the combination of the two streams is effective. It appears that in most cases, both systems either classify the phone correctly or incorrectly. However, in many instances, one system is correct while the other is wrong. In 62% of these instances the correct classification prevails. Finally, in about 5% of the instances in which both systems are incorrect, the merged system (miraculously!) performs the classification correctly, whereas in the 0.2% of the instances in which both systems' classification is correct, the merged system makes an error.

Besides the overall advantages, it was also observed that the MB and FB system are different in their level of accuracy for various phone classes. Most notably, the MB system is inferior to the FB system in classifying some fricatives and vowels, while the MB system excels in classifying silence and some liquids and glides.

Chapter 8

Epilogue

In this dissertation, a multi-band approach to automatic speech recognition was explored. There has been an explosion of interest in this topic recently, and the generated interest may be sustained since many aspects of this promising area remain unexplored.

8.1 Summary and Conclusions

Multi-band ASR is a special case of the multi-stream model, where each sub-frequency region of the speech signal is treated as a distinct source of information and the streams are combined after being processed independently. Some motivations for the multi-band paradigm are advantages of processing narrow-band signals, results from psycho-acoustic studies, robustness to noise, and the potential for taking advantage of parallel processing architectures. The multi-band paradigm has shown graceful degradation in noisy conditions.

My efforts were concentrated on five aspects of the multi-band approach. The first was the design choices and implementation of a baseline multi-band system. In the process of this design, I experimented with various design parameters. Next, I analyzed the performance of the baseline multi-band system to see whether the division of the bands caused phonetic information to be transmitted more poorly, and whether sub-band transitions occurred asynchronously, as surmised by many researchers. Based on the observations of asynchrony in my analysis, I investigated two algorithms for asynchronous merging of the sub-band streams. Later, I focused on deriving reduced sub-band phone classes, motivated by the observations that some bands seemed to be better than others for discriminating between particular categories. Finally, I experimented with using full-band and multi-band information streams in combination for improving the word error rate.

The multi-band paradigm has shown graceful degradation and robustness in noisy situations. I made a conscious choice not to concentrate on noisy speech, since (1) the multi-band paradigm appeared to match the characteristics of narrow-band noisy speech very well, whereas, the potential for improving clean speech recognition seemed less certain, and therefore, a greater and more attractive challenge, and (2) much of the efforts of other researchers and our collaborators focuses on this problem. I believed that multi-band processing could be used to improve the recognition accuracy of clean speech, and based on the results of this thesis, my conjecture has been confirmed.

The summary of my results are as follows:

There were many parameters to adjust and spaces to search in designing the baseline multi-band system. In the parameter search, I observed that (Chapter 3):

- Narrow-band PLP features, followed by RASTA-PLP features, were better than spectral features.
- An MLP merger was better than simple linear and log-linear merging functions for frame-by-frame merging of the sub-band probabilities.
- When using an MLP merging function, using posteriors instead of the scaled likelihoods of the sub-band streams was slightly superior.
- In allocating parameters in the merger MLP, it seemed best to increase the size of the hidden layer rather than the input window size.
- Embedded alignment of each sub-band MLP independently followed by the alignment of the merger MLP did not improve the recognition results. However, word recognition error rate improvements were observed if the merger MLP alone was iteratively aligned.
- The baseline multi-band system alone did not show inherent robustness to fast and reverberant speech.

In Chapter 4, I analyzed two common assumptions about multi-band ASR. Contrary to the objections of some multi-band ASR critics, I did *not* observe multi-band ASR to be inferior to a full-band approach because of phonetic information lost by the division of the frequency space into sub-bands. In fact, when using an MLP merger for the sub-band probability streams, phonetic features were transmitted better (47.15% for our database) than the comparable full-band system (45.48%).

Also reported in Chapter 4, I observed that some phone transitions tended to be delayed or advanced in a frequency-dependent manner. The overall transition lags also seemed to depend on both frequency and contrast conditions (speaking rate and reverberation). In particular, roughly one-third of the sub-band transitions in the control condition did not occur within 50 ms of each other. Furthermore, the high-frequency band timings had a spread that strongly depended on speaking rate. Thus, there was an indication for potential improvements in acoustic modeling if longer time-scale information stream merging (i.e., phone, syllable, or word) was used.

The observations in Chapter 4 led me to experiment with asynchronous merging of the sub-band streams (in Chapter 5). However, no improvements using asynchronous merging were observed as compared to merging the streams synchronously on the frame level.

Motivated by the observations that some bands seemed to be better than others for discriminating between particular categories, I tried to derive reduced sub-band phone sets in Chapter 6. It was observed that as the sub-band classes were merged, the frame error rate showed a decreasing trend, and the word recognition accuracy improved slightly. More specifically, almost 30% of the sub-band classes were merged, thereby making the global phonetic classes more easily distinguishable and reducing the overall frame error rate by a

relative 4.3%. The performance of the system with reduced phone classes was not better for Volvo and pink noise conditions, but was significantly better for reverberant speech. This result may be due to frequency-dependent energy-pattern smearing in reverberated speech, making the pattern of confusion among phones particularly dependent on the sub-band frequency region.

In Chapter 7, the full-band and multi-band systems were combined by multiplying the scaled likelihood streams on a frame-by-frame basis. A word recognition error rate of 4.7% on the Numbers95 development set was achieved, which is the the lowest recognition word error rate that we know of to date on this corpus. An analysis of the error patterns of the systems showed systematic differences in the two systems. Also, in Chapter 7, a multi-band pyramid scheme, with a larger input layer for low frequencies, was implemented for reverberant speech. The combination of this system with the full-band system resulted in a significant decrease in the word error rate.

8.2 Discussion and Contributions

General contributions of this work to ASR may be new explorations in an area of great interest to the research community. The design and implementation has shown that multi-band is a feasible method for ASR. It is hoped that this investigation aids better definition and refinement of future research in this area.

Most notably, the analysis of feature transmission in multi-band showed that it is a viable technique, and that information is not lost due to the division of the frequency bands. This is an important finding, since many researchers had asserted that multi-band was inherently inferior to full-band and dismissed this approach. This finding provided evidence for the negation of such assertions.

Another contribution of this work is the observation that the phone transitions in sub-bands do occur asynchronously, confirming the conjectures of many researchers in this area. It was observed, however, that by warping the bands in order to align these transitions we do not improve recognition performance. In fact, it may be that some phone-identity information inherent to the nature of this misalignment is lost. Asynchronous merging of sub-bands had been highlighted as a promising area of research by many, although no significant improvements had been observed. In the absence of any data to support otherwise, I have concluded that relaxing the synchrony requirement for all phone to phone transitions is not advantageous, and even suggest that some potentially useful transition information may be lost.

I had surmised that combining the sub-band phonetic categories that are confused with each other into a phonetic super-class for a given sub-band would improve the word error rate. Using relative entropy as the distance measure in a nearest-neighbor algorithm, the number of sub-band classes was reduced by almost 30%. Even though the size of the *input space* for the discrimination task was decreased by almost 30%, the overall phone discrimination task accuracy improved by about 4%. The word recognition error rate also decreased, although marginally, confirming the original hypothesis.

This work showed that significant improvements of the recognition accuracy may be witnessed if the multi-band stream is used in conjunction with the full-band stream. The analysis showed that some patterns of errors in the two systems, specifically for some phone

classes, are different, which may partly explain why the two systems complement each other's shortcomings. Combining the multi-band and the full-band system on a frame-by-frame basis resulted in the most significant improvements for clean speech that have yet been observed in the multi-band approach.

8.3 Future Work

It is not uncommon for dissertations to answer some questions and pose far more. This one is no exception. In what follows, I try to delineate some future research directions in multi-band ASR.

The topic of asynchronous merging of the sub-bands seemed particularly attractive because of the disagreement between sub-bands about where a phoneme starts and ends. Since the asynchronous merging approach used in this thesis did not provide any improvement, perhaps as an alternative method diphone units should be used instead of mono-phones. The window span of the input to the merger unit may have to be increased to contain the distributed phoneme transition in all sub-bands. Transition-Based (TB) Phone Modeling, a simplified version of the Stochastic Perceptual Auditory-event Model (SPAM) [99], which is similar to diphones, emphasizes the transition regions, and may be an interesting future direction. The main idea of TB models is that the modeling power of the recognizer should be more focused on distinguishing between states representing change than on states corresponding to little change. The intuitive argument for applying TB models to multi-band is that since transitions in each of the sub-bands occur asynchronously, TB models can determine regions of significant change more accurately. Yet when combining the sub-band streams, this asynchrony could be used as additional information, instead of disposed, as it was previously when the synchrony constraints were relaxed.

In Chapter 6, it was shown that although the size of the input space to the multi-band merger was reduced by almost 30% (by combining the confusable sub-band classes), the overall phonetic classification improved, and the word recognition improved marginally. Perhaps a different global discrimination function, such as an MLP, a non-greedy algorithm, or a set of phonetic constraints would prove even more effective in choosing classes to merge. Furthermore, based on the experience of Chapter 7, perhaps a more significant improvement could be observed if the reduced phone-set system is combined with a full-band one.

Another area worthy of future study is further experimentation with frame-by-frame merging criteria. In this thesis, combining the sub-band probability streams using an MLP proved to be the best merging mechanism for clean speech. Experimenting further with traditional statistical methods such as linear and non-linear opinion pools and supra Bayesian methods [63, 67] may provide further improvements.

It has been shown here that some sub-band features (PLP followed by RASTA-PLP) are better than others (sub-band critical-band features) for multi-band processing. Devising feature extraction techniques, especially for sub-band processing, may improve the sub-band recognition performance, and therefore improve the overall multi-band system further.

8.4 Final Thoughts

What is the relationship of this work to the greater body of speech knowledge? The results seem consistent with many psycho-acoustic experiments and models, and perhaps intimate a combination (full-band and multi-band) model for human speech perception.

Similar to the psycho-acoustic observations of [89, 38, 43, 142, 81, 46] automatic speech recognition of narrow-band features was surprisingly accurate. Furthermore, Greenberg et al.'s [46] observation of human sensitivity to spectral asynchrony in narrow-slit experiments are not inconsistent with my experimental observations in Chapter 6, which suggested that relaxing the synchrony requirement may dispose of potentially useful recognition cues.

Fletcher's work [37] espoused a sub-band-based model of human speech processing. Although not all aspects of that model were represented in this work (e.g., merging the sub-band streams according to the relative reliability of the channels), the centerpiece of the model, independent processing of sub-bands and merging of the information on a higher level, was represented. The results are not inconsistent with the Fletcherian model.

The best speech recognition system in this work was a hybrid one: the combination of the full-band and the multi-band processing systems. There is already evidence suggesting that human speech recognition benefits from combination of different streams of information, specifically, visual and auditory cues seem to be processed independently and combined [86]. It is imaginable that human speech processing may also be a combination of both full-band and sub-band information processing, where, in addition to processing the full frequency range, independent sub-band frequency processors extract specialized information according the the sub-band characteristics and the signal-to-noise ratio. In short, human speech processing may prove to be best represented by a combination of the Fletcherian and more traditional models of speech processing.

Appendix A

"How Do Humans Process and Recognize Speech?"

Jont Allen's paper entitled *"How Do Humans Process and Recognize Speech?"* [1] was the original motivation for this dissertation. This section includes an extended summary of Allen's paper, which reviews the work of Harvey Fletcher [37]. Harvey Fletcher started the articulation index (AI) proprietary project at AT&T Bell Labs around 1918.

There are many aspects to the following model. The centerpiece, however, is the proposal that human speech recognition is performed based on narrow-band processing of the signal and that these information streams are merged later in the recognition process. This main facet of the model is the one on which my work has focused and expanded on.

A.1 Motivation

The motivation of Fletcher's work was to quantify the quality of speech sounds on the telephone to improve speech intelligibility and preference. Today, this study seems particularly relevant to the relatively new science of automatic speech recognition.

A.2 Highlights

- Speech is recognized based on a cascade of layers.
- Context effects are strong and confound the study of recognition.
- Removing context factors from the human speech recognition problem increases testing efficiency and reduces test variability.
- The phone (one allophone of a phoneme) seems to be the basic unit of speech.
- The phone is derived from independent auditory and visual channels.
- Each channel depends on the SNR, not the energy.
- To account for independent articulation channels, humans must be extracting local features which are integrated in the brain.

- Using context-free databases in ASR should be fruitful.

A.3 The Model

The proposed model for human speech processing (see Fig. 6 in [1]) has five layers, all of which are below the language level. There is no feedback in this simplified model. The first layer is the cochlear filter bank, which affects the SNR. The next level is the *feature extraction* layer, where the input is SNR_k and output is a partial-phone recognition, $1 - e_k$. The third layer maps the features to phones. In the last two layers, syllables and words are found.

A.4 Controlling Context Entropy

The contextual entropy of a CVC (consonant-vowel-consonant, e.g., *fit*) word is lower than that of a nonsense CVC (e.g. *fif*); in other words, context is a strong cue for speech understanding. Fletcher controlled this variable in the articulation index experiments by using nonsense CVCs. If we use the phone frequencies given by Fletcher, the average phone entropy is 4.3 bits/phone. This entropy would drop as context effects are included.

Note that Fletcher used the word *articulation* to mean the probability of correctly identifying nonsense speech sounds, while *intelligibility* is the probability of correctly identifying meaningful speech sounds, such as words.

A.5 The Articulation Experiment

Fletcher used balanced nonsense syllables (CV, VC, and CVC) to minimize context effects. He held the speech corpus constant for each experiment in order to maintain constant source entropy. He computed an average over many speaker-listener pairs. The articulation was varied by: (1) changing the SNR, and (2) low-pass and high-pass filtering the speech.

The syllable lists were spoken and the listeners recorded what they heard (see Figure 1 in [1]). Then error probabilities $1 - c$ and $1 - v$ for the consonant and vowel sound-units, respectively, were computed. To simplify computations, an average {C,V} speech-unit articulation probability s was computed from the composition of {C,V} units in the database (i.e., $s = \frac{(2c+v)}{3}$ for CVCs). For perfect conditions (i.e., high SNR, no HP/LP filtering) the recognition accuracy was 98.5%.

A.6 The Results of the Experiment

Fletcher empirically found that the CVC syllable articulation $S(\alpha)$, where α is the gain applied to the speech, is accurately predicted from the phone articulations $c(\alpha)$ and $v(\alpha)$ by the relation $S(\alpha) = c(\alpha)^2 v(\alpha) \approx s(\alpha)^3$; this implies that the three sound-units are analyzed as independent sounds. In other words, the probability of correct recognition for the average *phone* determines human nonsense syllable recognition. Note, however, that this is only true for nonsense words, since for meaningful words context will decrease the entropy.

After having found the fundamental importance of the phone articulation, s, to human speech recognition, Fletcher further studied s for various channel frequency responses and channel noise, using LP and HP filters on the speech. He found that the *partial articulations*, s_L and s_H, for low and high bands, respectively, did not sum to the wide-band articulation s. Hence, he looked for a non-linear transformation of the partial articulations that would make them additive, as in

$$A(s_L(f_c, \alpha)) + A(s_H(f_c, \alpha)) = A(s(\alpha)) \tag{A.1}$$

where f_c is the high- and low-pass cut-off frequency. Fortunately Fletcher found $A(s)$, which he called *articulation index*, empirically:

$$A(s) = \frac{log_{10}(1-s)}{log_{10}(1-s_{max})} \tag{A.2}$$

where constant s_{max} is the maximum articulation and $e_{min} = 1 - s_{max} = 0.015$ is the corresponding minimum articulation error. Solving equation A.2 for s, we get $s(A) = 1 - e_{min}^{A}$ which could be written in terms of $e(A)$ simply as $e(A) = e_{min}^{A}$.

The nonlinearly transformed articulation defines an *articulation index density* $D(f)$ over frequency f. Integration or summation over this density gives the *articulation index* A. A can be viewed as a fundamental internal variable of speech recognition. $A(s)$ could be calculated using the articulation index density:

$$A(s_L(f_c)) = \int_0^{f_c} D(f)df \tag{A.3}$$

$$A(s_H(f_c)) = \int_{f_c}^{\infty} D(f)df \tag{A.4}$$

$$A(s) = \int_0^{\infty} D(f)df \tag{A.5}$$

where the density $D(f)$ is uniquely determined from:

$$D(f_c) = \frac{\partial}{\partial f_c} A(s_L(f_c)) \tag{A.6}$$

Fletcher derived the *density over frequency of the phone articulation index* $D(f)$ from his experiments.

A.7 Independent Articulation Bands

Fletcher then showed that phones are processed in independent *articulation bands* (frequency channels), and that these independent estimates are "optimally" merged. For example if we have 100 spoken sounds and 10 errors are made while listening to the low band, and 20 errors are made while listening to the high band, then $e = 0.1 * 0.2 = 0.02$, that is, only 2 errors will be made when listening to the full band and $s = 1 - 0.02 = 0.98$. This result was showed through the following observations: from equations A.1 and A.2 we get:

$$log(1-s) = log(1-s_L) + log(1-s_H) \tag{A.7}$$

which becomes

$$1 - s = (1-s_L)(1-s_H) \tag{A.8}$$

and in terms of the articulation error $e = 1 - s$

$$e = e_L e_H \tag{A.9}$$

This equation is true for any value of f_c. It says that we are listening to *independent* sets of phone features in the two bands and processing them independently, up to the point where they are fused to produce the phone estimates. Fletcher expanded the two-band example, above, to the multi-channel model, where $K = 20$ bands and $1 - s = e_1 e_2 e_3 ... e_k ... e_K$.

Taking $K = 20$ makes each band correspond to 1 mm along the basilar membrane. Since Fletcher identified a critical-band to be about 0.5 mm[1] along the basilar membrane, one articulation band represents two critical-bands. Figure 5 in [1] shows the ratio of $D(f)/kf$ where $D(f)$ is the articulation density and $k(f)$ is the critical ratio (a measure of the relative bandwidths of our cochlear filters). This plot of $D(f)/kf$ represents the measure of relative speech articulation per critical-band and it is approximately uniform over the speech band.

A.8 Band Error e_k and SNR

Fletcher and colleagues found that when the SNR is varied, the articulation band error is given by the relation:

$$e_k = (e_{min})^{SNR_k/30K} \tag{A.10}$$

where SNR_k is the signal-to-noise ratio in dB for each band k. Recall that each band k corresponds to 1 mm along the basilar membrane and that e_{min} is 0.015. This relationship tells us that it is the speech SNR *and not the energy* which determines the phone articulation s.

Jont Allen noted a serious error in the model in the limit. For very low SNR_k, we expect that articulation should go to chance: $e_k \longrightarrow 1 - 1/M \Longrightarrow s_k \longrightarrow 1/M$ where M is determined by the number of bits/band. However, according to the model, it goes to zero: $e_k \longrightarrow 1 \Longrightarrow s_k \longrightarrow 0$.

A.9 The Recognition Chain

The recognition chain in the model is the following:

- The signal enters the cochlea and is spectrally decomposed into critical-bands which define the signal-to-noise ratios SNR_k, where k labels the cochlear frequency channel.

[1] A more recent estimate of the critical-band length along the basilar membrane is 0.9 mm [48].

- The first layer processes the output from the cochlea and defines the phone features represented by partial articulation errors, e_k, using equation A.10.
- Then, the independent-band phone articulation model $s = e_1 e_2 ... e_k .... e_K$ determines s.
- The phones are transformed into (nonsense) CVC syllable units with articulation $S = s^3$.
- Words are determined with intelligibility $W(A) = 1 - (1 - S(A))^j$, where the constant $j > 1$ depends on the entropy of the word corpus and may be empirically determined.

Allen argues that any feedback mechanism could create serious "real-time" problems, so it is unlikely that feedback is common or significant between the deeper and outer layers.

A.10 The Phone Feature Space?

Allen uses the term *feature* to refer to 1 bit of partial recognition as a binary concept. Recall that for $K = 20$, each band corresponds to 1 mm along the basilar membrane. Given that the distance between 300 Hz and 8 kHz on the basilar membrane is 20 mm[2] and the phone entropy is 4.3 bits (see above), there are $20/4.3 = 4.65$ mm/feature. Furthermore, since 1 octave corresponds to about 5 mm along the basilar membrane, the average phone feature density is about 1 feature/octave and the average feature length is 1 octave/feature.

Allen argues against a template-based (*across-frequency* processing) approach to phonetic feature extraction and espouses locally based *across-time* processing: First, the cochlea breaks the external world down into a tonotopic array of critical-bands. The CNS "reconstructs" the scene from these pieces by making a huge cascade of binary decisions. The local feature extractors define the feature length along the tonotopic axis. The number of correlated feature regions determines the dimension of the space and the coordinates (e.g., the probability of the feature being present). Phone recognition is not the synchronous timing of feature events, but rather constitutes some more abstract relation between the presence and absence of features and their geometrical relations in a multi-dimensional feature space.

[2]More recently, this estimate has been placed around 28 mm.

Appendix B

Phone Symbols

The following are the set of symbols used in this thesis. The ASCII version of the set is based on the symbol set of TIMIT [136], a hand-labeled read-speech database commonly used in the ASR community.

ASR Phone Symbols					
ICSI56 phoneset	IPA	Example	ICSI56 phoneset	IPA	Example
pcl	p^o	(p closure)	bcl	b^o	(b closure)
tcl	t^o	(t closure)	dcl	d^o	(d closure)
kcl	k^o	(k closure)	gcl	g^o	(g closure)
p	p	**p**ea	b	b	**b**ee
t	t	**t**ea	d	d	**d**ay
k	k	**k**ey	g	g	**g**ay
q		ba**t**	dx		dir**t**y
ch	$\widehat{t}$	**ch**oke	jh	$\widehat{d}$	**j**oke
f	f	**f**ish	v	v	**v**ote
th		**th**in	dh		**th**en
s	s	**s**ound	z	z	**z**oo
sh		**sh**out	zh		a**z**ure
m	m	**m**oon	n	n	**n**oon
em	m	bott**om**	en	n	butt**on**
ng		si**ng**	h#		(silence)
nx	~	wi**nn**er	el	l	bott**le**
l	l	**l**ike	r	r	**r**ight
w	w	**w**ire	y	j	**y**es
hh	h	**h**ay	hv		a**h**ead
er		b**ir**d	axr		butt**er**
iy	i	b**ee**t	ih		b**i**t
ey	e	b**ai**t	eh		b**e**t
ae	æ	b**a**t	aa		f**a**ther
ao		b**ough**t	ah		b**u**t
ow	o	b**oa**t	uh		b**oo**k
uw	u	b**oo**t	oy	^{y}	b**oy**
aw	^{w}	ab**ou**t	ay	^{y}	b**i**te
ax		**a**bout	ix		deb**i**t

Table B.1: The ICSI56 phoneset (table constructed by Eric Fosler-Lussier and Chuck Wooters at ICSI).

Appendix C

Broad Category Memberships

The following table includes the phone broad category memberships that have been used in this thesis. This division has been done by Gary Tajchman at ICSI.

Broad Category Membership			
ICSI56 phoneset	Broad Phonetic Category	ICSI56 phoneset	Broad Phonetic Category
pcl	silence	bcl	silence
tcl	silence	dcl	silence
kcl	silence	gcl	silence
p	stop	b	stop
t	stop	d	stop
k	stop	g	stop
q	silence	dx	stop
ch	fricative	jh	fricative
f	fricative	v	fricative
th	fricative	dh	fricative
s	fricative	z	fricative
sh	fricative	zh	fricative
m	nasal	n	nasal
em	nasal	en	nasal
ng	nasal	h#	silence
nx	nasal	el	liquid
l	liquid	r	liquid
w	liquid	y	liquid
hh	fricative	hv	fricative
er	vowel	axr	vowel
iy	vowel	ih	vowel
ey	vowel	eh	vowel
ae	vowel	aa	vowel
ao	vowel	ah	vowel
ow	vowel	uh	vowel
uw	vowel	oy	vowel
aw	vowel	ay	vowel
ax	vowel	ix	vowel

Table C.1: The ICSI56 phoneset broad category memberships.

Appendix D

Feature Category Membership

Tables D.1 and D.2 include feature category memberships used in Chapter 4. The assignments for the ICSI phoneset were made by Gary Tajchman based on [143].

	vowel	consonant	diphthong	liquid	glide	stop	closure	nasal	fricative	lateral	approximant	voice	high
p	-	+	-	-	-	+	-	-	-	-	-	-	-
b	-	+	-	-	-	+	-	-	-	-	-	+	-
t	-	+	-	-	-	+	-	-	-	-	-	-	-
d	-	+	-	-	-	+	-	-	-	-	-	+	-
k	-	+	-	-	-	+	-	-	-	-	-	-	-
g	-	+	-	-	-	+	-	-	-	-	-	+	-
q	-	+	-	-	-	+	-	-	-	-	-	-	-
f	-	+	-	-	-	-	-	-	+	-	-	-	-
v	-	+	-	-	-	-	-	-	+	-	-	+	-
th	-	+	-	-	-	-	-	-	+	-	-	-	-
dh	-	+	-	-	-	-	-	-	+	-	-	+	-
s	-	+	-	-	-	-	-	-	+	-	-	-	-
z	-	+	-	-	-	-	-	-	+	-	-	+	-
sh	-	+	-	-	-	-	-	-	+	-	-	-	-
zh	-	+	-	-	-	-	-	-	+	-	-	+	-
hh	-	+	-	-	-	-	-	-	+	-	-	-	-
hv	-	+	-	-	-	-	-	-	+	-	-	+	-
m	-	+	-	-	-	-	-	+	-	-	-	+	-
n	-	+	-	-	-	-	-	+	-	-	-	+	-
ng	-	+	-	-	-	-	-	+	-	-	-	+	-
l	-	+	-	+	-	-	-	-	-	+	-	+	-
r	-	+	-	+	-	-	-	-	-	-	-	+	-
w	-	+	-	-	+	-	-	-	-	-	+	+	-
y	-	+	-	-	+	-	-	-	-	-	+	+	+
ih	+	-	-	-	-	-	-	-	-	-	-	+	+
eh	+	-	-	-	-	-	-	-	-	-	-	+	-
ae	+	-	-	-	-	-	-	-	-	-	-	+	-
ah	+	-	-	-	-	-	-	-	-	-	-	+	-
aa	+	-	-	-	-	-	-	-	-	-	-	+	-
ao	+	-	-	-	-	-	-	-	-	-	-	+	-
uh	+	-	-	-	-	-	-	-	-	-	-	+	+
ax	+	-	-	-	-	-	-	-	-	-	-	+	-
iy	+	-	+	-	-	-	-	-	-	-	-	+	+
ey	+	-	+	-	-	-	-	-	-	-	-	+	-/+
ay	+	-	+	-	-	-	-	-	-	-	-	+	-/+
oy	+	-	+	-	-	-	-	-	-	-	-	+	-/+
aw	+	-	+	-	-	-	-	-	-	-	-	+	-/+
ow	+	-	+	-	-	-	-	-	-	-	-	+	-/+
uw	+	-	+	-	-	-	-	-	-	-	-	+	+
er	+	-	-	-	-	-	-	-	-	-	+	+	-
axr	+	-	-	-	-	-	-	-	-	-	+	+	-
h#	-	-	-	-	-	-	-	-	-	-	-	-	-
pcl	-	-	-	-	-	-	+	-	-	-	-	-	-
bcl	-	-	-	-	-	-	+	-	-	-	-	-	-
tcl	-	-	-	-	-	-	+	-	-	-	-	-	-
dcl	-	-	-	-	-	-	+	-	-	-	-	-	-
kcl	-	-	-	-	-	-	+	-	-	-	-	-	-
gcl	-	-	-	-	-	-	+	-	-	-	-	-	-

Table D.1: Phonetic-feature-class assignments for a subset of the ICSI56 phoneset, derived by Gary Tajchman based on [Withgott and Chen 1993].

	low	front	back	round	short	long	labial	dental	coronal	palatal	retroflex	velar	glottal
p	-	-	-	-	-	-	+	-	-	-	-	-	-
b	-	-	-	-	-	-	+	-	-	-	-	-	-
t	-	-	-	-	-	-	-	-	+	-	-	-	-
d	-	-	-	-	-	-	-	-	+	-	-	-	-
k	-	-	-	-	-	-	-	-	-	-	-	+	-
g	-	-	-	-	-	-	-	-	-	-	-	+	-
q	-	-	-	-	-	-	-	-	-	-	-	-	+
f	-	-	-	-	-	-	+	-	-	-	-	-	-
v	-	-	-	-	-	-	+	-	-	-	-	-	-
th	-	-	-	-	-	-	-	+	-	-	-	-	-
dh	-	-	-	-	-	-	-	+	-	-	-	-	-
s	-	-	-	-	-	-	-	-	+	-	-	-	-
z	-	-	-	-	-	-	-	-	+	-	-	-	-
sh	-	-	-	-	-	-	-	-	-	+	-	-	-
zh	-	-	-	-	-	-	-	-	-	+	-	-	-
hh	-	-	-	-	-	-	-	-	-	-	-	-	+
hv	-	-	-	-	-	-	-	-	-	-	-	-	+
m	-	-	-	-	-	-	+	-	-	-	-	-	-
n	-	-	-	-	-	-	-	-	+	-	-	-	-
ng	-	-	-	-	-	-	-	-	-	-	-	+	-
l	-	-	-	-	-	-	-	-	+	-	-	-	-
r	-	-	-	-	-	-	-	-	-	-	+	-	-
w	-	-	-	-	-	-	+	-	-	-	-	-	-
y	-	+	-	-	-	-	-	-	-	-	-	-	-
ih	-	+	-	-	-	-	-	-	-	-	-	-	-
eh	-	+	-	-	-	-	-	-	-	-	-	-	-
ae	+	+	-	-	-	-	-	-	-	-	-	-	-
ah	-	-	-	-	-	-	-	-	-	-	-	-	-
aa	+	-	-	-	-	-	-	-	-	-	-	-	-
ao	+	-	+	-	-	-	-	-	-	-	-	-	-
uh	-	-	+	-	-	-	-	-	-	-	-	-	-
ax	-	-	-	-	+	-	-	-	-	-	-	-	-
iy	-	+	-	-	-	+	-	-	-	-	-	-	-
ey	-	+	-	-	-	+	-	-	-	-	-	-	-
ay	+/-	-/+	-	-	-	+	-	-	-	-	-	-	-
oy	+/-	-/+	+/-	-/+	-	+	-	-	-	-	-	-	-
aw	+/-	-	-/+	-/+	-	+	-	-	-	-	-	-	-
ow	-	-	+/-	-/+	-	+	-	-	-	-	-	-	-
uw	-	-/+	+/-	-/+	-	+	-	-	-	-	-	-	-
er	-	-	-	-	-	-	-	-	-	-	+	-	-
axr	-	-	-	-	-	-	-	-	-	-	+	-	-
h#	-	-	-	-	-	-	-	-	-	-	-	-	-
pcl	-	-	-	-	-	-	-	-	-	-	-	-	-
bcl	-	-	-	-	-	-	-	-	-	-	-	-	-
tcl	-	-	-	-	-	-	-	-	-	-	-	-	-
dcl	-	-	-	-	-	-	-	-	-	-	-	-	-
kcl	-	-	-	-	-	-	-	-	-	-	-	-	-
gcl	-	-	-	-	-	-	-	-	-	-	-	-	-

Table D.2: Phonetic-feature-class assignments for a subset of the ICSI56 phoneset, derived by Gary Tajchman based on [Withgott and Chen 1993].

Appendix E

Confusion Matrices

Figures E.1, E.2, E.3, E.4, E.5, and E.6 show confusion matrices generated for sub-bands 1 through 4, the full-band, and multi-band, respectively. The data have been generated on Numbers95 development set. Detailed descriptions of the systems used are in Chapter 3.

Confusion matrix for band 1

	b	d	g	t	k	dcl	tcl	kcl	s	sh	z	f	th	v	dh	m	n	l	r	w	y	hh	hv	iy	ih	eh	ey	ae	aa	ay	ah	ao	ow	uw	er	axr	ax	h#
b	-	-	-	1	-	2	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
t	-	17	-	4560	113	8	227	8	486	5	77	114	1051	30	-	-	37	2	90	55	-	2	-	43	23	51	72	-	-	15	28	1	69	65	11	-	5	306
k	-	4	-	74	392	-	12	55	132	-	2	17	16	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1	3	-	1	-	-	-	-	12
dcl	-	-	-	-	-	6	-	-	-	-	-	2	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
tcl	-	-	-	95	17	4	2656	150	155	-	42	170	139	79	-	-	82	-	8	6	-	-	2	37	9	-	21	-	-	2	4	3	28	22	5	-	-	595
kcl	-	-	-	11	69	20	223	1528	40	-	2	148	41	24	-	-	4	-	9	-	-	-	-	5	50	2	3	-	-	-	4	-	8	3	-	-	6	86
s	-	20	2	550	166	5	242	52	4853	14	201	1234	516	98	-	-	68	2	42	83	-	9	4	49	44	46	17	-	-	65	84	12	110	44	-	-	6	1025
z	-	23	-	68	-	16	61	-	149	-	662	45	65	28	-	-	60	3	5	35	-	17	-	47	42	4	15	-	-	-	18	-	10	14	2	-	-	88
f	-	-	5	281	31	11	467	91	1920	9	88	4309	358	77	-	-	58	1	34	105	-	20	7	27	18	29	28	-	1	74	16	16	157	25	6	-	3	1063
th	-	-	-	290	5	-	56	4	185	1	89	118	619	8	-	-	4	-	45	12	-	1	-	4	17	7	40	-	-	-	10	1	21	3	4	-	-	152
v	6	20	-	84	4	41	243	32	105	-	16	137	52	3195	-	-	270	4	28	24	-	1	1	25	21	94	28	-	3	118	94	-	51	-	12	-	12	401
n	-	13	-	91	-	23	202	17	207	-	162	111	31	456	-	-	12691	23	172	626	-	16	3	1196	96	40	320	-	3	365	614	6	454	226	3	-	47	676
l	-	-	-	2	-	-	-	-	1	-	-	-	-	-	-	-	-	4	2	2	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r	-	3	-	81	1	-	21	6	61	-	61	35	98	26	-	-	130	14	4091	143	4	4	-	524	413	117	277	-	3	69	177	240	973	527	108	-	10	160
w	-	3	-	106	1	3	25	-	209	1	84	102	64	30	-	-	722	36	98	2993	-	37	9	105	93	181	64	5	-	465	483	17	266	84	6	-	5	163
hh	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	3	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
iy	-	26	-	94	1	4	60	8	105	-	60	53	20	39	-	-	922	6	675	94	4	-	1	6290	266	35	1045	-	-	135	155	17	332	2480	101	-	18	331
ih	-	2	-	46	1	5	7	68	144	-	54	37	25	25	-	-	31	4	275	85	-	1	-	190	2889	241	64	-	4	22	108	54	274	222	44	3	12	70
eh	-	-	-	15	-	-	3	-	46	-	2	12	8	74	-	-	13	-	50	105	-	-	-	11	153	2179	7	-	7	132	304	11	104	-	15	-	1	34
ey	-	4	-	52	4	3	21	-	41	-	4	5	43	4	1	-	132	13	191	33	-	-	2	391	70	74	2148	-	3	88	81	9	692	74	133	-	9	60
ay	6	1	-	58	-	2	1	-	193	1	5	96	23	246	-	3	531	65	530	699	-	-	5	262	118	543	454	20	183	15364	1815	125	1841	83	39	-	5	168
ah	-	8	-	50	6	-	2	2	66	-	3	16	17	113	-	-	223	4	177	415	-	4	-	114	229	735	166	-	44	589	4211	16	317	22	19	-	99	95
ao	-	-	-	4	-	-	-	-	1	-	-	-	-	-	-	-	1	-	13	9	-	-	-	4	5	-	2	-	-	-	6	31	113	2	13	-	-	3
ow	-	2	-	274	5	3	71	26	292	-	32	178	99	99	-	-	391	89	1944	552	-	2	-	430	371	369	1156	-	51	1217	876	720	8684	525	247	-	9	468
uw	-	3	-	78	2	-	29	7	72	-	30	49	14	2	-	-	186	3	599	81	-	-	2	1263	246	28	111	-	-	36	28	20	203	4410	12	-	9	163
er	-	-	-	4	-	1	7	-	-	-	-	1	-	3	-	-	5	-	33	3	-	-	-	6	28	26	41	-	-	-	14	15	39	2	162	-	1	4
ax	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	1	-	-	-	2	-	-	-	-	-	-	-	-
h#	-	28	-	657	62	28	2638	160	3916	14	461	2552	903	725	1	-	709	4	252	265	-	13	3	369	134	44	136	-	-	170	442	24	427	225	12	1	45	40356

Figure E.1: Confusion matrix for band 1.

Confusion matrix for band 2

	b	d	g	t	k	dcl	tcl	kcl	s	sh	z	f	th	v	dh	m	n	l	r	w	y	hh	hv	iy	ih	eh	ey	ae	aa	ay	ah	ao	ow	uw	er	axr	ax	h#
d	-	-	-	2	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
t	-	37	-	4498	114	6	176	14	482	8	58	132	525	52	1	-	135	5	112	60	-	2	-	185	89	49	92	-	1	61	70	-	64	166	27	-	7	292
k	-	3	-	60	450	-	5	41	133	-	-	10	13	6	-	-	4	-	-	-	-	-	-	3	3	3	-	-	-	19	-	-	2	-	-	-	2	14
dcl	-	-	-	-	-	-	1	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6
tcl	1	2	-	123	15	38	2749	154	100	-	42	90	124	127	-	-	233	4	18	10	-	4	-	29	8	2	51	-	-	5	9	2	23	10	-	-	-	584
kcl	-	1	-	25	58	20	217	1564	17	-	4	36	11	98	-	-	78	-	3	-	-	1	-	4	46	3	1	-	-	3	7	-	1	1	2	-	-	65
s	-	17	-	587	151	-	287	17	6305	16	541	974	583	88	-	-	526	3	117	38	-	35	3	257	183	112	141	-	5	77	99	2	103	204	8	3	6	1056
z	-	4	-	32	-	-	28	-	127	1	239	26	41	8	-	-	51	-	15	6	-	1	1	7	42	5	17	-	-	3	5	-	8	7	-	-	-	68
f	-	2	6	149	19	2	176	36	787	1	81	5056	176	126	-	-	237	6	50	114	-	40	10	45	20	21	30	-	-	107	26	15	132	64	2	-	-	963
th	-	7	-	276	5	-	93	8	177	-	97	90	1281	13	1	-	107	1	117	17	-	8	-	32	83	9	19	-	-	32	14	-	20	11	12	-	5	171
v	2	14	-	45	4	15	107	88	49	-	6	120	23	2786	-	-	206	2	20	22	-	-	-	23	15	115	19	-	1	87	103	-	45	32	16	-	10	427
n	-	26	-	216	14	60	599	147	606	1	170	383	189	433	-	1	10162	9	227	106	-	10	7	908	103	56	182	-	6	433	409	17	306	189	6	-	31	1653
l	-	-	-	1	-	-	-	-	-	-	-	-	-	3	-	-	8	3	3	-	-	-	-	1	-	-	-	-	-	-	1	-	8	-	-	-	-	1
r	-	5	-	74	2	3	25	3	95	-	34	62	209	16	-	-	155	16	5237	70	1	8	2	360	221	67	118	-	33	256	764	206	427	366	104	-	8	160
w	-	-	-	101	1	-	14	-	29	-	9	86	20	7	-	-	160	12	64	4293	-	4	1	78	16	18	51	4	-	37	690	37	374	119	4	-	7	247
hh	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	3	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-
iy	-	9	-	118	3	-	58	8	324	-	34	73	51	26	-	-	845	5	396	110	1	-	-	6366	193	98	893	-	2	395	141	37	350	423	6	-	4	492
ih	-	3	-	77	3	6	19	55	171	5	79	39	97	28	-	-	69	1	168	15	-	3	-	242	2740	452	198	-	1	106	226	6	77	271	67	-	22	107
eh	-	2	-	22	4	-	8	-	99	-	6	29	10	60	-	-	29	-	17	14	-	-	-	35	320	2311	128	-	-	154	237	3	99	54	54	-	2	40
ey	-	2	-	68	-	1	56	2	134	-	25	48	14	30	-	-	129	-	28	33	-	-	-	392	199	242	2847	-	6	233	57	5	191	257	33	-	31	1653
aa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-
ay	-	4	-	274	2	-	17	2	294	-	18	185	73	225	-	-	656	7	483	71	-	3	4	654	170	284	539	21	150	15015	364	43	972	356	240	-	3	270
ah	-	7	-	47	20	1	6	2	76	-	4	22	13	74	-	1	249	3	598	583	-	1	1	91	242	324	56	-	70	127	5280	39	186	65	26	-	86	91
ao	-	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-	1	-	29	2	-	-	-	-	-	-	-	-	-	-	6	50	63	-	-	-	-	-
ow	-	4	-	85	1	-	36	12	136	-	15	145	34	90	-	-	254	165	954	355	-	4	2	316	93	325	256	-	26	644	414	851	10858	1342	33	-	7	444
uw	4	1	-	126	2	-	12	3	154	-	22	99	43	35	-	-	108	12	401	75	3	-	7	268	234	241	231	-	-	163	108	8	561	4513	46	-	10	383
er	-	-	-	14	-	1	7	-	2	-	2	8	6	1	-	-	7	-	31	-	-	-	-	-	33	17	11	-	-	29	7	-	5	10	256	-	2	-
ax	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	2	-	-	-	-	-	1	4
h#	5	30	1	601	12	32	2578	54	3080	13	650	1826	667	1053	-	1	2869	17	273	428	3	4	1	1098	282	91	335	-	1	165	519	17	299	598	10	1	83	38812

Figure E.2: Confusion matrix for band 2.

Confusion matrix for band 3

	b	d	g	t	k	dcl	tcl	kcl	s	sh	z	f	th	v	dh	m	n	l	r	w	y	hh	hv	iy	ih	eh	ey	ae	aa	ay	ah	ao	ow	uw	er	axr	ax	h#
d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
t	-	24	-	4651	41	6	208	9	272	-	53	175	276	40	-	-	123	6	96	72	-	1	-	246	115	83	309	-	2	94	86	5	143	161	22	-	5	281
k	-	1	-	19	477	1	2	56	113	-	6	6	7	5	-	-	2	-	4	-	-	1	-	-	5	9	-	-	1	17	-	13	-	-	-	3	6
tcl	-	5	-	127	5	37	3166	116	39	-	51	64	73	77	-	-	278	-	17	41	-	-	-	36	12	5	52	-	-	3	7	-	39	24	3	-	1	773
kcl	-	2	-	20	58	9	147	1602	14	-	-	10	5	76	-	-	70	-	10	17	-	-	-	5	56	6	7	-	-	2	2	-	2	5	-	-	15	95
s	-	19	-	339	170	1	136	15	5242	17	473	1152	356	104	-	-	702	24	111	326	-	29	5	199	192	130	158	-	2	172	230	70	679	165	7	-	15	839
z	-	3	-	20	-	-	24	-	130	-	285	72	41	1	-	-	24	-	16	24	-	-	1	13	23	5	29	-	11	-	27	12	40	2	-	-	1	85
f	-	3	-	387	11	3	107	21	1171	3	95	2801	379	116	-	-	515	3	103	378	-	5	-	77	160	112	97	-	11	208	227	79	423	237	15	2	12	742
th	-	7	-	172	6	3	53	3	66	-	47	58	1559	9	2	-	55	2	185	46	-	1	1	32	94	6	8	-	-	20	27	15	68	6	36	-	3	115
v	8	16	-	34	8	37	148	64	49	-	16	129	20	2297	-	-	438	3	12	49	-	-	-	26	24	84	9	-	1	77	118	5	71	47	1	-	8	288
n	-	13	-	150	6	28	542	114	732	-	126	700	162	844	-	-	8446	21	163	417	-	7	5	514	80	82	44	1	3	278	293	26	626	275	16	-	26	1229
l	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
r	-	9	-	117	4	3	24	8	140	-	10	86	224	30	-	-	141	15	5320	59	-	16	2	337	243	136	50	-	22	395	228	148	450	249	195	-	5	185
w	-	-	-	50	7	1	33	16	404	-	178	601	157	46	-	-	385	26	103	2628	-	11	-	24	26	31	11	-	-	125	637	51	395	38	4	-	-	556
iy	-	9	-	212	2	5	59	1	278	-	35	64	36	13	-	-	351	16	383	55	4	-	1	7171	196	52	1160	-	4	332	65	8	314	220	5	-	2	261
ih	-	2	-	75	4	4	19	71	160	-	44	53	87	36	-	-	37	-	233	18	-	6	-	241	2573	243	102	-	1	108	117	12	80	128	46	2	2	57
eh	-	2	-	83	11	1	7	2	112	1	7	30	34	61	-	-	43	1	63	17	-	-	-	36	237	1985	15	-	13	148	222	7	156	144	27	-	1	39
ey	-	12	-	182	3	1	59	3	121	-	25	39	14	6	-	-	32	5	29	4	4	-	-	552	166	107	3139	-	4	89	18	-	210	104	34	-	1	61
ay	4	3	-	148	2	2	19	21	289	1	40	155	63	154	-	3	552	4	695	139	-	13	15	817	280	318	395	24	170	14620	933	88	1009	322	201	-	13	317
ah	-	6	-	75	18	-	14	6	247	-	31	133	27	86	-	-	307	1	439	448	-	5	3	82	369	760	22	-	23	636	4827	78	484	93	53	-	97	215
ao	-	-	-	3	-	-	-	-	6	-	-	1	1	-	-	-	1	-	3	1	-	-	-	-	-	-	-	-	-	1	-	3	10	3	-	-	-
ow	-	19	7	177	10	2	58	2	1200	5	189	513	171	156	-	-	893	112	709	267	-	18	6	374	152	312	243	-	29	1130	623	392	6531	1530	39	-	18	1022
uw	-	2	-	201	8	1	30	2	310	-	16	292	32	65	-	-	253	15	314	31	-	-	-	177	166	277	236	-	6	142	107	17	807	4314	5	-	-	327
er	-	-	-	25	-	-	6	-	-	-	-	1	16	-	-	-	4	-	-	6	-	-	-	-	26	7	2	-	-	36	19	-	5	7	242	-	-	1
ax	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-
h#	-	24	-	360	29	40	2413	82	2284	18	411	2406	463	1163	-	-	3621	23	310	1381	-	14	-	435	139	95	127	-	11	317	747	322	2618	983	3	-	83	38985

Figure E.3: Confusion matrix for band 3.

Confusion matrix for band 4

	b	d	g	t	k	dcl	tcl	kcl	s	sh	z	f	th	v	dh	m	n	l	r	w	y	hh	hv	iy	ih	eh	ey	ae	aa	ay	ah	ao	ow	uw	er	axr	ax	h#
t	-	37	-	4390	25	9	179	11	222	-	25	117	353	48	-	-	158	1	105	98	-	-	-	191	75	84	120	2	3	146	116	4	151	157	57	-	14	280
k	-	1	-	38	417	1	8	43	105	-	2	14	10	21	-	-	8	-	9	4	-	4	-	4	3	6	5	-	-	7	22	-	39	-	-	-	2	9
dcl	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
tcl	-	3	-	137	5	39	3094	73	47	-	18	54	82	100	-	-	262	1	14	65	-	-	-	37	9	2	38	-	-	11	6	-	24	15	2	-	2	686
kcl	-	-	-	11	57	6	164	1495	3	-	-	36	29	185	-	-	109	4	20	6	-	3	-	7	45	5	2	-	-	8	3	-	7	3	6	-	-	157
s	-	9	2	402	178	1	122	8	4615	16	397	925	224	126	-	2	365	23	238	427	-	44	3	297	122	109	88	1	14	631	409	86	814	164	23	-	11	647
z	-	-	-	1	1	-	7	1	62	-	142	8	32	1	-	-	16	-	6	35	-	7	-	8	17	3	4	-	6	1	19	-	20	1	-	-	1	23
f	-	9	2	204	28	1	91	13	748	6	163	1975	327	154	-	-	540	11	124	391	-	6	-	134	92	114	150	-	6	349	247	51	529	100	15	-	18	568
th	-	16	-	210	4	6	60	5	49	-	58	84	1091	11	2	-	77	1	171	41	-	-	6	30	157	34	8	-	-	17	91	2	66	5	43	1	8	116
v	2	2	-	87	20	13	208	136	34	-	5	130	40	1688	-	-	435	1	58	51	-	-	-	52	18	63	23	-	2	70	97	6	56	36	1	-	2	229
n	-	26	1	355	25	36	627	166	650	-	124	1066	210	1178	-	-	8050	33	662	430	-	5	4	555	181	187	133	-	8	434	479	57	802	982	20	-	37	1815
r	-	9	-	107	13	3	34	27	168	-	33	73	228	118	-	-	409	8	4269	117	-	14	-	357	178	132	49	-	9	463	288	86	673	176	196	-	4	164
w	-	1	2	71	11	-	92	25	556	-	375	480	162	58	-	-	289	22	107	2128	-	10	5	60	48	27	45	-	-	203	428	38	427	39	7	-	6	521
iy	-	10	-	172	9	1	77	6	384	-	24	100	68	41	-	-	274	16	400	55	2	1	13	5407	212	200	934	-	27	691	241	18	735	627	11	-	-	250
ih	-	1	-	106	-	4	15	53	155	-	26	26	188	30	-	-	71	-	175	15	-	-	-	257	2260	553	70	-	1	176	241	12	70	71	67	1	-	61
eh	-	5	-	66	2	-	4	1	65	-	2	64	36	45	-	-	45	1	33	14	-	-	-	244	256	1096	76	-	7	126	221	-	54	92	11	1	2	24
ey	4	3	-	96	2	1	57	5	60	9	9	123	17	5	-	-	44	-	20	20	6	10	6	385	81	139	2065	1	6	451	53	1	235	97	12	-	2	92
ay	4	4	-	373	5	4	39	11	1366	9	98	899	181	168	-	-	686	19	935	394	6	10	6	1663	653	1036	1755	16	124	1269	1489	231	1876	1211	166	-	20	845
ah	-	11	-	166	12	1	15	2	508	-	104	182	247	125	-	-	231	1	246	316	-	5	2	313	499	565	98	-	25	559	3549	55	477	96	22	1	69	231
ao	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	3	-	5	-	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-
ow	-	12	-	124	19	3	44	8	1205	-	88	732	186	87	-	1	698	97	967	337	-	15	-	647	173	160	315	5	42	1420	647	458	5529	882	45	-	20	871
uw	-	-	-	139	6	-	7	9	146	-	5	202	20	39	-	-	545	16	205	52	-	-	-	78	201	116	-	-	7	294	106	4	560	3055	31	-	6	229
er	-	2	-	23	-	-	-	-	-	-	1	1	19	1	-	-	3	-	46	2	-	-	-	8	47	4	-	-	-	29	18	2	9	22	210	-	2	-
ax	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
h#	6	19	-	348	41	56	2320	115	2230	14	438	2250	453	1156	-	-	3955	22	548	1427	-	3	-	288	131	126	121	-	21	578	807	227	2020	1227	9	-	70	38661

Figure E.4: Confusion matrix for band 4.

Confusion matrix for full-band

	b	d	g	t	k	dcl	tcl	kcl	s	sh	z	f	th	v	dh	m	n	l	r	w	y	hh	hv	iy	ih	eh	ey	ae	aa	ay	ah	ao	ow	uw	er	ax	h#
t	-	31	-	5722	23	4	198	2	252	1	88	121	262	28	-	-	24	-	25	28	-	-	-	74	26	31	35	-	-	20	29	-	32	146	18	-	2	316
k	-	1	-	11	486	-	15	78	111	2	-	7	4	6	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	2	1	-	1	-	-	-	-	26
tcl	-	5	-	224	15	43	4102	47	79	-	61	56	116	44	-	-	95	-	14	10	-	4	-	55	7	2	66	-	-	6	-	2	10	12	3	-	-	740
kcl	-	1	-	5	82	17	116	1862	9	-	-	46	6	13	-	-	8	-	6	3	-	-	-	3	63	1	5	-	-	3	5	2	4	9	2	-	-	35
s	-	15	1	258	194	6	159	2	8495	31	508	794	543	137	-	-	77	-	44	17	-	15	5	79	114	110	60	-	-	57	81	5	90	50	1	-	1	1159
z	-	5	-	32	1	1	34	-	82	-	668	16	25	16	1	-	56	-	1	16	-	2	8	31	41	4	10	-	1	63	17	23	125	32	1	-	-	73
f	-	3	6	156	17	2	70	9	572	3	31	5915	136	107	-	-	56	4	18	31	-	10	-	16	7	6	3	-	1	63	17	2	9	-	24	-	-	907
th	-	16	-	156	4	1	72	1	101	-	49	42	2111	21	-	-	8	-	112	-	-	12	-	4	30	5	1	-	-	4	2	-	24	9	-	1	-	135
v	8	7	-	29	1	21	74	7	64	-	17	155	12	3566	-	-	153	9	15	25	-	1	-	6	20	118	12	-	-	157	101	1	33	7	6	-	7	308
n	-	13	-	59	2	36	201	9	180	-	119	85	30	297	-	1	14413	17	71	177	-	15	7	346	73	67	111	-	2	295	509	9	219	236	2	-	31	653
r	2	10	-	55	-	1	41	3	36	-	12	31	146	19	-	-	99	15	6990	60	-	5	-	336	187	77	64	-	16	61	112	220	453	80	153	-	11	211
w	-	-	-	44	1	-	15	6	18	-	10	52	15	23	-	-	171	18	25	4630	-	-	1	61	8	18	9	2	-	28	567	28	208	101	1	-	4	199
iy	-	24	-	124	3	9	69	5	101	-	48	59	14	10	1	-	354	5	383	59	4	-	1	9253	186	26	713	-	4	97	53	8	38	114	11	-	10	245
ih	-	1	-	32	-	-	3	5	126	-	67	35	37	10	-	-	31	-	200	1	-	-	-	169	3744	193	98	-	1	10	67	4	40	165	118	4	15	80
eh	-	-	-	11	-	-	2	-	93	-	-	1	4	65	-	-	9	1	17	7	-	-	-	1	97	3118	23	-	1	56	211	3	80	37	14	-	-	37
ey	-	8	-	50	2	-	69	2	41	-	11	23	10	8	-	-	35	-	36	9	4	-	-	334	107	117	4464	-	-	78	27	-	85	63	15	-	2	118
ay	-	8	-	48	2	-	5	1	113	-	13	104	11	173	-	-	349	1	87	75	-	19	3	169	52	196	214	23	149	17377	399	2	376	39	26	-	110	208
ah	-	8	-	13	6	-	3	-	19	-	2	14	2	72	-	2	241	2	220	699	-	-	-	24	214	377	22	-	96	157	6446	41	161	11	19	-	-	89
ao	-	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	5	-	-	-	-	-	-	-	-	-	1	36	6	36	2	-	-	-	4
ow	2	1	-	59	3	-	36	10	118	-	26	115	85	87	-	-	141	186	709	225	-	5	-	58	54	210	70	-	-	224	373	961	12447	368	23	-	16	466
uw	-	-	-	85	-	4	16	9	36	1	16	42	9	25	-	-	243	17	147	78	-	4	105	155	99	83	-	-	27	35	62	4	228	7378	15	-	3	226
er	-	-	-	14	-	-	9	-	-	-	-	4	11	2	-	-	2	-	31	-	-	-	-	-	31	3	6	-	-	30	12	1	4	-	485	-	4	7
h#	-	31	-	436	40	33	1963	86	2733	7	392	1824	669	656	-	-	708	2	203	270	-	39	10	270	119	68	146	-	1	177	483	16	488	203	13	-	77	40237

Figure E.5: Confusion matrix for full-band.

Confusion matrix for multi-band

	d	g	t	k	dcl	tcl	kcl	s	sh	z	f	th	v	dh	m	n	l	r	w	y	hh	hv	iy	ih	eh	ey	ae	aa	ay	ah	ao	ow	uw	er	axr	ax	h#
d	40	-	35	-	6	5	-	2	-	1	2	-	5	-	-	9	-	3	-	-	-	-	-	9	-	-	-	-	-	3	-	-	-	2	-	1	3
t	36	-	5948	22	10	226	3	311	1	49	125	313	12	-	-	44	1	36	34	-	-	-	70	15	35	59	-	1	10	15	-	22	82	4	-	2	245
k	-	2	-	11	544	-	1	49	145	-	1	9	-	1	-	1	-	-	-	-	-	-	-	-	-	3	-	-	-	2	-	-	-	-	-	-	6
dcl	-	1	-	10	1	49	33	-	-	-	7	-	4	3	-	-	11	-	1	1	-	-	-	-	-	1	-	-	-	-	1	-	-	-	-	-	-	7
tcl	-	-	-	178	6	24	4218	55	58	-	25	43	97	14	-	-	66	-	13	2	-	-	-	40	7	-	55	-	-	3	6	2	7	15	8	-	-	578
kcl	-	-	-	3	77	8	39	1851	8	-	1	23	2	13	-	-	-	-	1	2	-	-	-	3	33	1	-	-	-	1	-	-	1	1	-	-	-	57
s	-	11	1	199	177	2	90	4	8782	26	349	811	430	77	-	-	61	-	16	12	-	18	1	73	89	66	49	-	1	77	40	3	86	50	-	1	2	850
z	-	2	-	36	-	1	33	-	216	-	1008	15	52	20	-	-	65	-	11	9	-	8	-	45	76	6	19	-	-	3	31	-	5	10	-	-	3	108
f	-	-	6	132	14	-	109	16	849	6	48	6155	240	104	-	-	47	1	8	30	-	22	3	18	17	19	19	-	-	76	15	13	98	22	2	-	3	931
th	-	9	-	192	5	4	74	8	141	-	85	69	2253	26	2	-	13	1	144	-	-	3	2	2	34	5	14	-	-	1	6	-	2	4	26	3	2	177
v	6	3	-	12	-	-	12	6	75	-	14	131	25	3683	-	-	144	1	12	27	-	-	-	4	19	111	9	-	3	125	105	-	50	1	4	-	11	376
n	-	12	-	70	-	28	213	17	130	-	103	81	32	219	-	-	14336	26	70	235	-	8	6	357	74	43	61	-	5	205	477	1	180	118	1	38	493
l	-	-	-	3	-	-	-	-	1	-	-	-	-	6	-	-	5	18	-	6	-	-	-	-	-	-	-	-	-	-	-	-	11	-	2	-	-	-
r	1	10	-	25	-	1	27	8	38	-	19	23	130	24	-	-	102	12	7044	53	-	6	-	423	264	71	65	-	8	58	195	265	538	93	221	-	7	108
w	-	-	-	41	-	-	9	-	19	-	1	36	3	19	-	-	218	18	37	4792	-	2	-	63	11	15	11	3	-	32	788	20	250	109	1	-	9	136
hh	-	-	-	-	-	-	-	-	10	-	2	2	-	-	-	-	6	-	-	-	-	36	-	-	-	-	-	-	-	-	3	-	-	9	-	-	-	1
hv	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
iy	-	9	-	111	-	2	50	1	80	-	39	26	16	-	-	-	240	1	407	46	6	1	6	9021	212	29	634	-	-	81	31	3	62	274	19	-	1	207
ih	-	3	-	12	-	7	6	92	131	-	61	27	29	19	-	-	23	-	181	7	1	-	-	179	3730	214	137	-	2	22	146	1	34	100	95	-	15	38
eh	-	-	-	11	-	-	-	-	74	1	2	17	1	92	-	-	19	1	13	16	1	-	-	23	109	3192	29	-	12	114	366	4	94	26	12	-	2	30
ey	-	5	-	55	-	-	54	-	33	-	14	10	8	7	-	-	37	2	17	5	1	-	1	422	104	142	4664	-	7	109	25	-	105	82	21	-	-	80
ae	-	-	-	-	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19	1	-	-	23	11	-	-	1	-	-	-	3
aa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	15	15	-	-	-	-	-	-	-
ay	4	-	-	30	-	1	4	1	67	-	-	65	1	229	-	2	376	4	108	50	-	7	5	171	51	172	157	22	153	17346	353	9	404	19	28	-	3	58
ah	-	12	-	11	2	-	1	-	13	-	4	11	3	57	-	1	258	3	118	530	-	4	2	11	197	346	14	-	83	103	6058	12	87	10	1	-	112	59
ao	-	-	-	1	-	-	-	-	-	-	-	4	-	2	-	-	3	1	78	8	-	-	-	-	-	2	-	-	-	-	19	103	245	3	1	-	3	1
ow	1	5	-	39	1	1	18	10	112	-	20	146	19	76	-	-	210	175	678	249	-	4	3	101	73	215	115	-	16	434	396	874	12343	451	27	-	8	331
uw	-	-	-	81	1	-	17	1	24	-	24	25	-	3	-	-	150	5	160	53	-	1	1	132	130	105	50	-	2	14	39	2	209	7395	18	-	6	145
er	-	-	-	10	-	2	2	-	1	-	-	-	6	-	-	-	1	-	35	1	-	-	1	-	22	6	4	-	-	10	23	-	6	3	451	-	1	2
ax	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	2	-	-	-	-	-	-	-	2	6	-	-	-	20	-	-	-	-	-	-	9	3
h#	-	20	-	370	30	27	1983	92	2055	11	261	1685	538	679	-	-	825	7	172	257	-	4	-	227	66	26	65	-	1	71	388	26	324	190	10	-	58	41446

Figure E.6: Confusion matrix for multi-band.

Appendix F

Merged Classes using Mutual Information Criterion

This section includes a list of merged classes in each band. The unmerged classes are not listed. The details of the collapsing process can be found in Chapter 6.

Table F.1 shows the phone class merges in each sub-band. Note that, for example, in band one, merges 2 through 4 merge [aa], [hv], and [ae] into [sh]. The fifth merge combines [sh] (and therefore the three previous merges, as well as [ax] from the 1st merge) into class [b]. Merges 6 through 10 combine more phone classes with small priors with class [b], thereby making a very large super-class. As explained in 6.2, similar behavior is observed in other bands.

	Band1			Band2			Band3			Band4		
1	ax	→	aa	sh	→	d	aa	→	d	ax	→	aa
2	aa	→	sh	aa	→	hv	dcl	→	d	aa	→	l
3	hv	→	sh	ae	→	hv	hh	→	d	l	→	d
4	ae	→	sh	d	→	b	sh	→	d	hh	→	d
5	sh	→	b	y	→	b	hv	→	d	sh	→	d
6	y	→	b	hv	→	g	ae	→	d	hv	→	d
7	g	→	b	axr	→	b	d	→	b	ae	→	d
8	axr	→	b	m	→	g	y	→	b	d	→	b
9	m	→	b	dh	→	b	g	→	b	y	→	b
10	dh	→	b	ao	→	l	axr	→	b	g	→	b
11	dcl	→	d	hh	→	dcl	m	→	b	axr	→	b
12	hh	→	d	ax	→	b	dh	→	b	m	→	b
13	l	→	d	dcl	→	b	ao	→	l	dh	→	b
14	er	→	ey	g	→	b	ax	→	l	ao	→	dcl
15	ao	→	d	er	→	eh	er	→	r	er	→	th
16	kcl	→	k	th	→	z	kcl	→	k	kcl	→	k
17	th	→	z	kcl	→	k	th	→	z	eh	→	ih
18	d	→	b	l	→	b	l	→	b	dcl	→	b
19	f	→	s	eh	→	ih	f	→	s	w	→	f
20	ah	→	eh	z	→	s	eh	→	ih	th	→	z
21	uw	→	iy	v	→	k	n	→	v	f	→	s
22	z	→	t	ey	→	iy	ey	→	iy	v	→	k
23	ey	→	r	tcl	→	k	tcl	→	k	ey	→	iy
24	tcl	→	k	ah	→	w	z	→	s	ah	→	ih
25	eh	→	w	uw	→	r	uw	→	ow	tcl	→	k
26	ih	→	r	f	→	s	ah	→	w	uw	→	ow
27	v	→	k	s	→	t	ih	→	r	z	→	t
28	r	→	b	r	→	b	k	→	t	ow	→	r
29	s	→	t	ih	→	b	ow	→	w	s	→	t
30	ow	→	b	n	→	k	v	→	s	ih	→	iy
31	k	→	t	ow	→	b	s	→	b	n	→	k
32	ay	→	w	iy	→	w	iy	→	r	iy	→	r
33	iy	→	n	k	→	t	t	→	b	ay	→	r
34	w	→	b	w	→	b	ay	→	r	t	→	b
35	h#	→	t	ay	→	b	w	→	b	k	→	b
36	n	→	b	h#	→	t	h#	→	b	r	→	b
37	t	→	b	t	→	b	r	→	b	h#	→	b

Table F.1: An ordered list of phone class merges in each band.

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] Takayuki Arai and Steven Greenberg. Speech intelligibility in the presence of cross-channel spectral asynchrony. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 2, pages 933–936, Seattle, WA, May 1998.

[3] L. E. Baum. An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[4] R. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. *Bulletin of the American Mathematical Society*, 73:360–363, 1967.

[5] Jeff Bilmes. Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 469–473, Seattle, WA, May 1998.

[6] Nabil N. Bitar and Carol Y. Espy-Wilson. Knowledge-based parameters for HMM speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 29–32, Atlanta, GA, USA, May 1996.

[7] Nabil N. Bitar and Carol Y. Espy-Wilson. The design of acoustic parameters for speaker-independent speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1239–1243, Rhodes, Greece, September 1997.

[8] M. Blomberg. Modelling articulatory inter-timing variation in a speech recognition system based on synthetic references. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 789–792, Genova, Italy, September 1991.

[9] Hervé Bourlard. Towards increasing speech recognition error rates. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 883–894, Madrid, Spain, 1995. Keynote Paper.

[10] Hervé Bourlard. Personal communication, 1998.

[11] Hervé Bourlard, Bart Dh̀oore, and Jean-Marc Boite. Optimizing recognition and rejection performance in wordspotting systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 373–376, Adelaide, South Australia, April 1994.

[12] Hervé Bourlard and Stéphane Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proceedings of the International Conference on Spoken Language Processing*, pages 426 – 429, Philadelphia, PA, USA, October 1996.

[13] Hervé Bourlard and Stéphane Dupont. Subband-based speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 2, pages 125–128, May 1997.

[14] Hervé Bourlard, Stéphane Dupont, and Christophe Ris. Multi-stream speech recognition. Technical Report IDIAP-RR 96-07, IDIAP, Martigny, Valais, Switzerland, December 1996.

[15] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Press, 1994.

[16] Leo Breiman. Stacked regressions. Technical report, University of California at Berkeley, Department of Statistics, 1992. Technical Report 367.

[17] Leo Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[18] Leo Breiman. Bias, variance, and arcing classifiers. Technical report, University of California at Berkeley, Department of Statistics, 1996. Technical Report 460.

[19] John Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman Soulié and J. Hérault, editors, *Neurocomputing: Algorithms, Architectures, and Applications*, pages 227–236. NATO ASI Series, 1990.

[20] J.S. Bridle, M.D. Brown, and R.M Chamberlain. An algorithm for connected word recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, pages 899–902, Paris, May 1982.

[21] Numbers corpus, release 1.0, 1995.

[22] Christophe Cerisara and Jean-Paul Haton. Multi-band continuous speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1235–1239, Rhodes, Greece, September 1997.

[23] Jordan R. Cohen. The summers of our discontent. In *Proceedings of the International Conference on Spoken Language Processing*, pages 9–10, 1996. Proceedings Addendum.

[24] Richard Comerford, John Makhoul, and Richard Schwartz. The voice of the computer is heard in the land and it listens too! *IEEE Spectrum*, 34(12):39–47, December 1997.

[25] Switchboard Corpus: Recorded Telephone Conversations. Produced by NIST, Sponsored by DARPA, October 1992.

[26] G.D. Cook, S.R. Waterhouse, and A.J. Robinson. Ensemble methods for connectionist acoustic modelling. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 4, pages 1959–1962, Rhodes, Greece, September 1997.

[27] Martin Cooke, Andrew Morris, and Phil Green. Missing data techniques for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 2, pages 863–866, Munich, Germany, April 1997.

[28] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 1991.

[29] K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24(6):637–642, November 1952.

[30] P. C. Delattre, A. M. Liberman, and F. S. Cooper. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27:769–773, 1955.

[31] John R. Deller, Jr., John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Company, New York, 1993.

[32] A. P. Dempster, M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–88, 1977.

[33] Li Deng and Don X. Sun. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, 95(5):2702–2719, May 1994.

[34] Harris Drucker, Corinna Cortes, L. D. Jackel, Yann LeCun, and Vladimir Vapnik. Boosting and other ensemble methods. *Neural Computation*, 6(6):1289–1301, 1994.

[35] Paul Duchnowski. *A New Structure for Automatic Speech Recognition.* PhD thesis, Massachusetts Institute of Technology, September 1993.

[36] Stéphane Dupont. Personal communication, 1996.

[37] Harvey Fletcher. *Speech and Hearing in Communication.* Krieger, New York, 1953.

[38] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 19(1):90–119, January 1947.

[39] Jürgen Fritsch and Michael Finke. Improving performance on switchboard by combining hybrid HME/HMM and mixture of Gaussians acoustic models. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 4, pages 1963–1966, Rhodes, Greece, September 1997.

[40] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, February 1986.

[41] Sadaoki Furui and Chin-Hui Lee. Robust speech recognition– An overview. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, page 93, Snowbird, Utah, December 1995. IEEE.

[42] M. J. F. Gales and S. J. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, pages 233–236, San Francisco, CA, March 1992.

[43] Oded Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):115–132, January 1994.

[44] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 517–520, 1992.

[45] David Graff. The 1996 Broadcast News Speech and Language-Model Corpus. In *DARPA Speech Recognition Workshop*, Westfields Internatinal Conference Center, Chantilly, Virginia, February 1997. DARPA.

[46] Steven Greenberg, Takayuki Arai, and Rosaria Silipo. Speech intelligibility derived from exceedingly sparse spectral information. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, December 1998. In press.

[47] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, volume 3, pages 1647–1650, Munich, Germany, April 1997. IEEE.

[48] Donald D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *Journal of the Acoustical Society of America*, 33:1344–1356, 1961.

[49] Astrid Hagen. Personal communication, 1998.

[50] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 1990.

[51] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[52] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[53] Hynek Hermansky and Sangita Sharma. TRAPS – classifiers of temporal patterns. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, December 1998. In press.

[54] Hynek Hermansky, Sangita Tibrewala, and Misha Pavel. Towards ASR on partially corrupted speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 462 – 465, Philadelphia, PA, USA, October 1996.

[55] H. G. Hirsch. Estimation of noise spectrum and its applications to SNR estimation and speech enhancement. Technical Report TR-93-012, International Computer Science Institute, Berkeley, CA, 1993.

[56] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, January 1994.

[57] Tammo Houtgast and Herman J. M. Steeneken. The modulation transfer function in room acoustics. *Bruel and Kjaer Technical Review*, 3:3–12, 1985.

[58] Tammo Houtgast and Jan A. Verhave. A physical approach to speech quality assessment: Correlation patterns in the speech spectrogram. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, pages 285–288. European Speech Communication Association, Istituto Int. Comunicazioni, 1991.

[59] X. D. Huang. Phoneme classification using semicontinuous hidden Markov models. *IEEE Transactions on Signal Processing*, 40(5):1062–1067, May 1992.

[60] L.E. Humes, D.D. Kirks, T.S. Bell, and C. Ahlstrom. Application of the Articulation Index and the Speech Transmission Index to the recognition of speech by normal-hearing and hearing-impaired listeners. *Journal of Speech Hearing Research*, 29:447–462, 1986.

[61] MathWorks Inc., 1998.

[62] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[63] Robert A. Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5):867–888, September 1995.

[64] Adam Janin. Personal communication, 1998.

[65] Fred Jelinek. Personal communication, Johns Hopkins Summer Workshop on ASR, August 1996.

[66] Fred Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1997.

[67] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, March 1994.

[68] Brian E. D. Kingsbury. *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*. PhD thesis, University of California, Berkeley, California, 1998. To be published.

[69] Brian E. D. Kingsbury and Nelson Morgan. Recognizing reverberant speech with RASTA-PLP. In *ICASSP*, volume 2, pages 1259–1262, Munich, Germany, April 1997. IEEE.

[70] Brian E.D. Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectogram. *Speech Communication*, 25(1–3):117–132, August 1998.

[71] Katrin Kirchhoff and Jeff A. Bilmes. Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, Phoenix, AZ, USA, March 1999. Submitted.

[72] Dennis H. Klatt. Review of the ARPA speech understanding project. *The Journal of the Acoustical Society of America*, 62(6):1324–1366, December 1977. Reprinted in (Waibel and Lee, 1990).

[73] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Neural Information Processing Systems*, volume 7, pages 231–238, 1995.

[74] Karl D. Kryter. Methods for the calculation and use of the articulation index. *Journal of the Acoustical Society of America*, 34(11):1689–1697, November 1962.

[75] Karl D. Kryter. Validation of the articulation index. *Journal of the Acoustical Society of America*, 34(11):1698–1702, November 1962.

[76] Peter Ladefoged. *A Course in Phonetics*. Harcourt Brace & Company, third edition, 1993.

[77] Kai-Fu Lee. *Automatic Speech Recognition*. Kluwer Academic Publishers, Norwell, MA, 1989.

[78] Kai-Fu Lee and Hsiao-Wuen Hon. Speaker independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, November 1989.

[79] A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

[80] Philip Lieberman and Sheila Blumstein. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, 1988.

[81] Richard P. Lippmann. Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech and Audio Processing*, 4(1):66–69, 1996.

[82] Richard P. Lippmann. Speech perception by humans and machines. In *Proceedings of the Workshop on the Auditory Basis of Speech Perception*, pages 309–316, Keele University, UK, July 1996.

[83] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[84] Richard P. Lippmann and Beth A. Carlson. Robust speech recognition with time-varying filtering, interruptions, and noise. In Sadaoki Furui, B.-H. Juang, and Wu Chou, editors, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 365–372, Santa Barbara, CA, December 1997.

[85] Alvin Martin, Jon Fiscus, Mark Przybocki, and Bill Fisher. The evaluation: Word error rates & confidence analysis. In *9th Hub-5 Conversational Speech Recognition Workshop*, page Section 12, Linthicum Heights, Maryland, September 1998.

[86] D.W. Massaro, M.M. Cohen, and P.M.T. Smeele. Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, 100:1777–1786, 1996.

[87] Paul McCourt, Saeed Vaseghi, and Naomi Harte. Multi-resolution cepstral features for phoneme recognition across speech sub-bands. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 557–560, Seattle, WA, May 1998.

[88] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience Puplication, 1997.

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

[90] Naghmeh Nikki Mirghafori. On robustness to fast speech in automatic speech recognition. Master's thesis, University of California at Berkeley, Berkeley, CA, December 1995.

[91] Nikki Mirghafori. Automatic speech recognition using multiple frequency bands. (http://www.icsi.berkeley.edu/ ~nikki/papers/Multiband_asr_1995.ps), May 1995.

[92] Nikki Mirghafori. An alternative approach to automatic speech recognition using sub-band linguistic categories. Thesis Proposal (http://www.icsi.berkeley.edu/ ~nikki/papers/thesis_prop.ps), December 1996.

[93] Nikki Mirghafori, Eric Fosler, and Nelson Morgan. Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, pages 491–494, Madrid, Spain, September 1995.

[94] Nikki Mirghafori, Eric Fosler, and Nelson Morgan. Why is ASR harder for fast speech and what can we do about it? In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 179–183, Snowbird, Utah, December 1995.

[95] Nikki Mirghafori, Eric Fosler, and Nelson Morgan. Towards robustness to fast speech in ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 335–338, Atlanta, Georgia, May 1996.

[96] Nelson Morgan. *Room Acoustics Simulation with Discrete-Time Hardware*. PhD thesis, UC Berkeley, 1980.

[97] Nelson Morgan and Hervé Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995.

[98] Nelson Morgan and Hervé Bourlard. Neural networks for statistical recognition of continuous speech. *Proceedings of the IEEE*, 83(5):742–770, May 1995.

[99] Nelson Morgan, Hervé Bourlard, Steven Greenberg, and Hynek Hermansky. Stochastic perceptual auditory-event-based models for speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1943–1946, Yokohama, Japan, September 1994.

[100] Nelson Morgan, Su-Lin Wu, and Hervé Bourlard. Digit recognition with stochastic perceptual speech models. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 771–774, Madrid, Spain, 1995.

[101] A. C. Morris and J. M. Pardo. Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 115–118, Madrid, Spain, 1995.

[102] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder. Improvements in beam search for 10,000-word continuous speech recognition. In *ICASSP*, volume 1, pages 9–12, San Francisco, California, March 1992. IEEE.

[103] Hermann Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(2):263–271, April 1984.

[104] Hermann Ney and Xavier Aubert. A word graph algorithm for large vocabulary, continuous speech recognition. In *ICSLP*, pages 1355–1358, Yokohama, Japan, September 1994.

[105] NIST. Continuous speech recognition corpus, September 1993. National Institute of Standards and Technology Speech.

[106] J. D. O'Connor, L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F.S. Cooper. Acoustic cues for the perception of initial /w,y,r,l/ in English. *Word*, 13:24–43, 1957.

[107] Martin Oerder and Hermann Ney. Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP*, volume 2, pages 119–122, Minneapolis, Minnesota, April 1993. IEEE.

[108] S. Okawa, E. Bocchieri, and A. Potamianos. Multi-band speech recognition in noisy environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 2, pages 641–444, Seattle, WA, May 1998.

[109] Kuldip K. Paliwal. Spectral subband centroids as features for speech recognition. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 124–130, Santa Barbara, CA, December 1997.

[110] M.P. Perrone and L.N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R.J. Mammone, editor, *Neural Networks for Speech and Image Processing*. Chapmann-Hall, 1993.

[111] John R. Pierce. Whither speech recognition. *Journal of the Acoustical Society of America*, 46:1049–1051, 1969.

[112] Louis C. W. Pols. Flexile human speech recognition. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–283, Santa Barbara, CA, December 1997.

[113] Patti Price, William M. Fisher, Jared Bernstein, and David S. Pallett. The DARPA 1000-word Resource Management database for continuous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 651–654, New York, New York, April 1988.

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

[115] Lawrence R. Rabiner. Applications of speech recognition in the area of telecommunications. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 501–510, Santa Barbara, CA, December 1997.

[116] Sudhakar Rao and Wiliam A. Pearlman. Analysis of linear prediction, coding, and spectral estimation from subbands. *IEEE Transactions on Information Theory*, 42(4):1160–1178, July 1996.

[117] Christopher Ris. Four-band multi-band results on Switchboard database. Reported at the Johns Hopkins 96 Workshop (http://www.clsp.jhu.edu/ws96/ris/ results-report.html), August 1996.

[118] T. Robinson, L. Almeida, J.M. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J.P. Neto, S. Renals, M. Saerens, and C. Wooters. A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE project. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1941–1944, Berlin, Germany, September 1993.

[119] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[120] M. J. Russell, K. M. Ponting, S. M. Peeling, S. R. Browning, J. S. Bridle, and R. K. Moore. The ARM continuous speech recognition system. In *ICASSP*, Albuquerque, April 1990.

[121] Martin Russell. Progress towards speech models that model speech. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 115–122, Santa Barbara, CA, December 1997.

[122] Hiroaki Sakoe. Two-level DP-matching– a dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(6):588–595, December 1979.

[123] David Sankoff and Joseph B. Kruskal. *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, MA, 1983.

[124] Bülent Sankur, Yasemin P. Kahya, E. Çagatay Güler, and Tanju Engin. Feature extraction and classification of nonstationary signals based on the multiresolution. In *International Conference on Pattern Recognition*, 1994.

[125] Ruhi Sarikaya and John N. Gowdy. Subband based classification of speech under stress. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 1, pages 569–572, Seattle, WA, May 1998.

[126] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, pages 875–878, 1997.

[127] Holger Schwenk. Using boosting to improve a hybrid HMM/neural network speech recognizer. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, Phoenix, AZ, March 1999. Submitted.

[128] Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, October 1995.

[129] S. A. Solla, E. Levin, and M. Fleisher. Accelerated learning in layered neural networks. *Complex Systems*, 2:625–640, 1988.

[130] Richard M. Stern. Specification of the 1995 ARPA HUB 3 evaluation: Unlimited vocabulary NAB news baseline. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 5–7, February 1996.

[131] K. Stevens, S. Keyser, and H. Kawasaki. Toward a phonetic and phonological theory of redundant features. In J. Perkell and D. Klatt, editors, *Invariance & Variability in Speech Processes.* Erlbaum, Hillsdale, N.J., 1986.

[132] Sangita Tibrewala. Personal communication, October 1996.

[133] Sangita Tibrewala. Seven-band multi-band results on Switchboard database. Reported at the Johns Hopkins 96 Workshop (http://www.clsp.jhu.edu/ws96/ris/results-report.html), August 1996.

[134] Sangita Tibrewala and Hynek Hermansky. Multi-band and adaptation approaches to robust speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2619–2622, Rhodes, Greece, September 1997.

[135] Sangita Tibrewala and Hynek Hermansky. Sub-band based recognition of noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 2, pages 1255–1258, May 1997.

[136] TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech Disc 1-1.1, October 1990. NTIS Order No. PB91-505065.

[137] Mikio Tohyama. Response statistics of rooms. In Malcolm J. Crocker, editor, *Encyclopedia of Acoustics*, volume 2, chapter 77, pages 913–923. John Wiley and Sons, Inc., New York, New York, 1997.

[138] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland, and M. A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, volume 2, pages 1247–1250, April 1997.

[139] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, pages 845–848, 1990.

[140] T.K. Vintsyuk. Element-wise recognition of continuous speech consisting of words from a specified vocabulary. *Kibernetika (Cybernetics)*, 2:133–143, March-April 1979.

[141] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[142] Richard M. Warren, Keri R. Riener, James A. Bashford, Jr., and Bradley S. Brubaker. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception and Psychophysics*, 57(2):175–182, 1995.

[143] M. Margaret Withgott and Francine R. Chen. *Computational Models of American Speech*. Center for the Study of Language and Information (CSLI), Menlo Park, CA, 1993.

[144] D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[145] Kevin Woods, W. Philip Kegelmeyer Jr., and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 19(4):405–410, April 1997.

[146] Charles Clayton Wooters. *Lexical Modelling in a Speaker Independent Speech Understanding System*. PhD thesis, UC Berkeley, November 1993. ICSI Technical Report TR-93-068.

[147] Su-Lin Wu. *Incorporating Information From Syllable-length Time Scales into Automatic Speech Recognition*. PhD thesis, University of California at Berkeley, May 1998.

[148] Su-Lin Wu, Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, pages 721–724, Seattle, WA, May 1998.

[149] Steve Young. Large vocabulary continuous speech recognition: A review. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 29–44, Snowbird, Utah, December 1995.

[150] Victor Zue and et. al. Recent progress on SUMMIT system. In *Proceedings of the Third DARPA Workshop on Speech and Natural Language*, pages 380–384, June 1990.