Contents

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

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

Multi-Band Speech Recognition: A Summary of Recent Work at ICSI

Naghmeh Nikki Mirghafori

TR-97-051

December 1997

Abstract

Effective incorporation of different streams of information is the main goal of the multi-stream paradigm [9]. Multi-band automatic speech recognition (ASR) [18, 42, 7, 25] is a special case of the multi-stream model. In multi-band speech recognition each sub-frequency region of the speech signal is treated as a distinct source of information and the streams are combined after being processed individually. Some motivations for the multi-band paradigm are signal processing advantages, psycho-acoustic studies, robustness to noise, and taking advantage of parallel processing architectures. Multi-band paradigm has been shown to be particularly robust to noisy conditions [7, 25].

In this technical report we discuss the recent work on multi-band ASR at ICSI. This exposition consists of three themes. Our first topic is the design and implementation of a multi-band baseline system. Next, we discuss the analysis of multi-band ASR, in terms of phonetic information transmission and potential advantage of asynchronous merging of sub-band streams. The third topic is motivated by the intuition that some bands are inherently better for classifying some phones, whereas others lack sufficient information for such discrimination. We report on a multi-band system designed based on this hypothesis.

1 Introduction

1.1 Background

Speech is the most natural form of human communication. It is also the most suitable data entry medium for for many applications, such as word processing, telephone applications, hands-free operations, handicapped applications, just to name a few. The aim of ASR has been to design and implement systems capable of transcribing speech. There have been great advances in ASR technology since its inception in 1950's. 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 1920's [51]. The first "real" speech recognizer, came along next in 1952 at Bell Labs built by Davis, Biddulph, and Balashek [16]. This system was a speaker-dependent digit recognizer which achieved 2% error rate as long as the speaker did not move his head. ASR research has advanced greatly since then, creating more general systems with larger vocabulary and increased robustness to environmental and speaker variation. Word error rates between 7% and 8% are not uncommon using context-dependent phoneme level hidden Markov models (HMMs), N-gram language models and huge quantities of training data for vocabularies of up to 65,000 words [73]. In a recent DARPA evaluation, for example, a word error rate of 6% was achieved on a speaker-independent unlimited vocabulary read-speech task [62]. Although impressive, the ASR problem is not solved by any means, as the state of the art is nowhere close to human speech recognition capabilities [37, 52, 13]. The error rates of automatic speech recognizers are one or two orders of magnitude higher than those of humans for many speech recognition tasks, ranging from a 10-word digit 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 35% and around 54% if the conversation parties are familiar with one another [40]. Variables such as speaking style, speaking rate, accent, variable vocal effort, background noise, room reverberation, and channel effects degrade ASR accuracy dramatically, whereas they affect human speech recognition much less [44, 35, 31]. Clearly, ASR is not a solved problem, and with the recent debates on the incremental nature of advancement in the field [6], new and exploratory paradigms are most welcome.

1.2 Our Goal

Recently there has been much excitement generated in the ASR community on the topic of multi-band paradigm, mainly by Jont Allen's [1] cogent retelling Harvey Fletcher's [19] psycho-acoustic studies of the 1950's. Multi-band ASR method [18, 42, 7, 25] is a special case of the multi-stream paradigm. The goal of the multi-stream model is the incorporation of different information streams [9], as the streams may be audio and visual information, or different sets of features derived from speech

data. In multi-band ASR the full frequency band is divided into multiple regions, probability estimation is performed on each sub-band, and the streams of information are combined and synchronized at anchor-points (e.g., frame, phone, syllable endings). Some motivations for the multi-band paradigm are signal processing advantages, psycho-acoustic studies, robustness to noise, and taking advantage of parallel processing architectures, among others. In some experiments multi-band has had a lower word error rate in both normal and band-limited noise conditions [7, 25] compared to a traditional full-band system.

Our goal in this work is to design and implement a multi-band system, analyze and explore the advantages and disadvantages of this paradigm, and extend the state-of-the-art by implementing extensions to multi-band, more specifically, explore multi-band ASR with merged sub-band classes.

The organization of this technical report is as follows: in the proceeding subsections we offer a brief description of the HMM/ANN speech recognition system (Section 1.3) and the multi-band system (Section 1.4). In Section 2 we will discuss motivations for the multi-band approach, a merged phoneset, and the choice of test-bed. In Section 3 we discuss issues pertaining to multi-band ASR system design and implementation. Section 4 addresses the following two concerns: 1) whether the multi-band approach is inferior to 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 an advantage of multi-band paradigm. Section 7 discusses the design and implementation of a multi-band system with merged phone sets. Conclusions are in Section 8 and our future direction is discussed in Section 9. Finally, the Appendices include a detailed summary of previous work, as well as confusion matrices and merged sub-band classes.

1.3 ICSI's HMM/ANN Hybrid Speech Recognition System

This work is performed within the established framework of ICSI's hybrid hidden Markov model/artificial neural network (HMM/ANN) speech recognition system [10]. The main components of our speech recognizer are highlighted in Figure 1. The first component is the signal processing element, where each frame of speech (e.g., each 20 msec segment, overlapped every 10 msecs) 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. We usually use RASTA-PLP processing [24]. The next element in our system is the phonetic probability estimator, which is a fully connected multi-layer perceptron trained using the backpropagation algorithm [56] with softmax normalization [11] on the output layer and relative entropy error criterion [61] to estimate the probability of each phoneme corresponding to (multiple) frames of speech. Next, the phonetic probabilities, along with a grammar[1] and a

[1]We often use a bigram grammar. A bigram grammar is specified by a list of words that can follow a particular word, along with associated probabilities.

lexicon[2] are used in a dynamic programming-based Viterbi search [70], a simplified version of the Forward algorithm [2, 3], to find the best strings of words corresponding to the acoustic data.

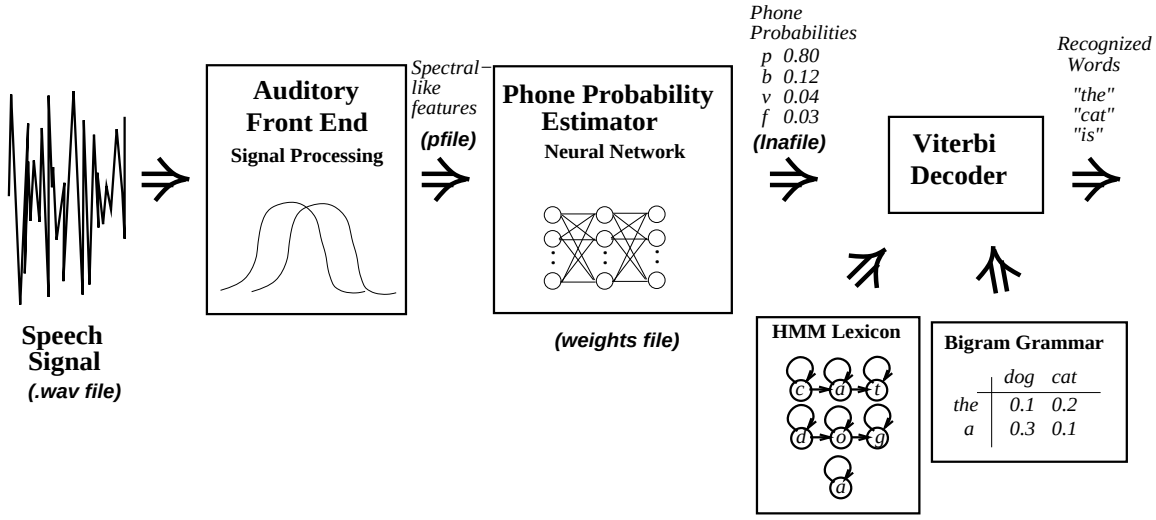

Figure 1: The ICSI hybrid HMM/ANN speech recognition architecture.

1.4 What is Multi-band Processing?

Figures 2 and 3 highlight the basic elements of a system based on the multi-band paradigm. The speech signal is divided into multiple frequency regions and acoustic features are derived for each sub-band region. These features are then used for phonetic probability estimation. The sub-band probability estimates are combined either before or during the decoding process. One main advantage of the multi-band paradigm is the ability to combine the streams asynchronously. We note that this facet of multi-band highlighted in the ASR community as a potential method for overcoming fundamental limitations of current HMM-based systems [58].

2 Motivation

In what follows, we first present motivations for the multi-band paradigm. Next, we explain why using merged linguistic classes seems promising. Finally, we will present reasons for choosing spontaneous continuous speech tasks as our test-bed.

2.1 Motivation for the Multi-Band Paradigm

The following are motivations for the multi-band approach:

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

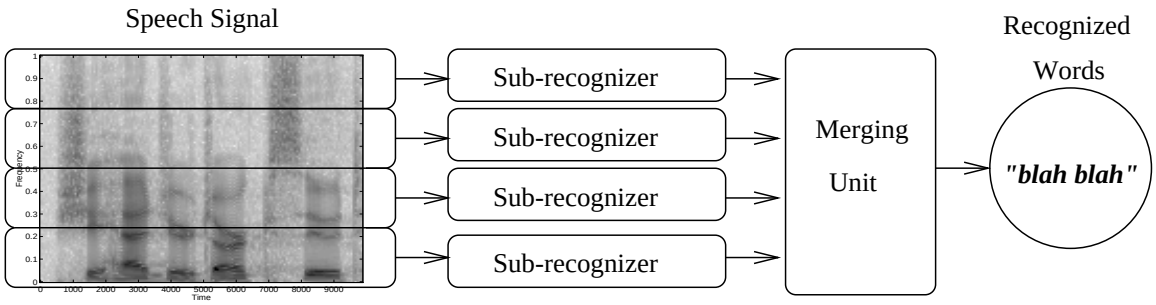

Figure 2: A simple overview of multi-band.

- Recently Rao and Pearlman [54] have proven theoretically, and shown with simulations, that auto-regressive spectral estimation from sub-bands offers a gain over full-band auto-regressive spectral estimation. Furthermore, 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.

- 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 is noise-free, but the testing data has 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. It was shown [7, 25], that the multi-band paradigm is more robust to such noise, and shows more graceful degradation.

- It has been suggested [21] that different frequency regions have different dynamic characteristics. By processing each band separately, we have the option of both developing feature extraction methods and employing variable sized temporal windows tuned to the dynamic characteristics of each frequency region.

- 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 [19, 1, 41]. We discuss the psycho-acoustic motivation in more detail in Chapter 10.2.

- Researchers have hypothesized [8, 67, 68] that phone transitions occur at different times in different bands (see Figure 4). The multi-band recognition allows us to conveniently allow for asynchronous combination of the sub-band frequency information. We will examine this supposition in Section 4.3.

- Signal processing approaches may attempt to merge the redundancy [63] in the information inherent in the speech signal. Multi-band processing is an attempt to exploit this redundancy in neighboring frequency bands.

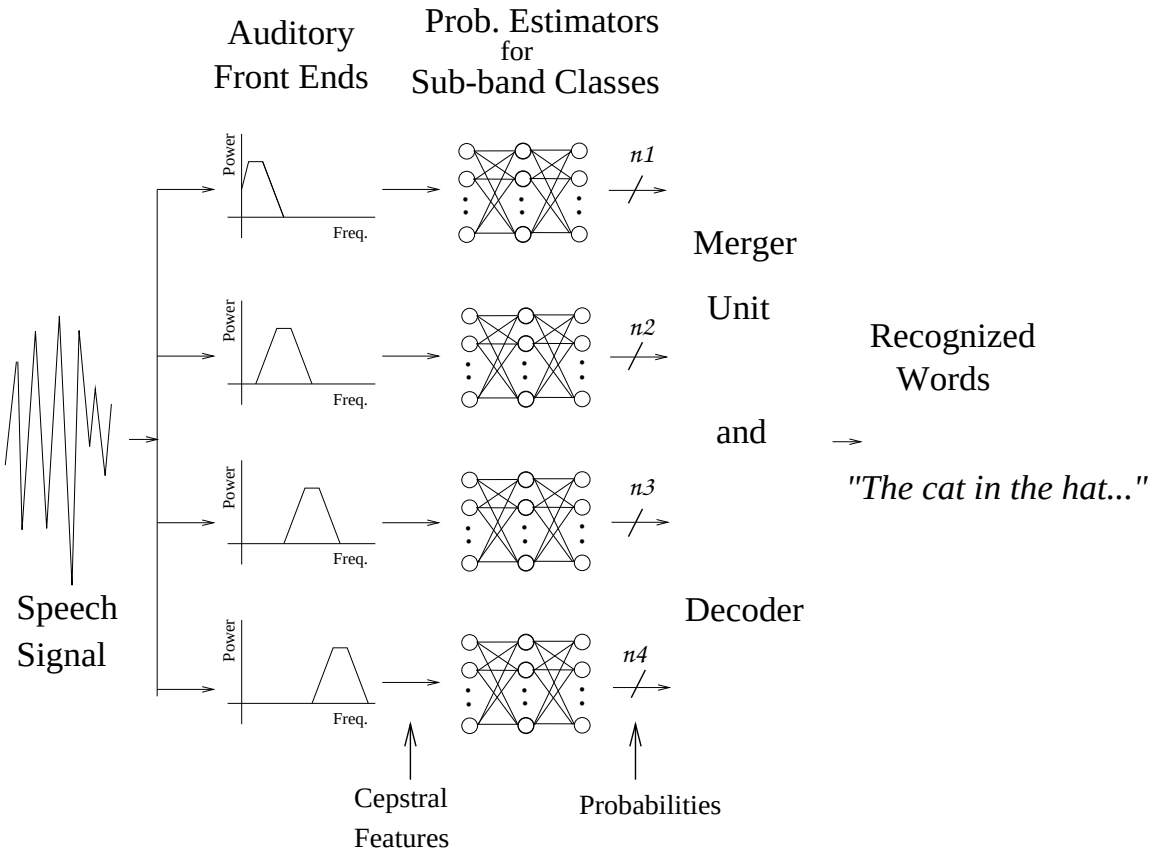

Figure 3: The structure of the multi-band system.

- By processing the bands in parallel, we are in a better position to take advantage of parallel architecture machines.

- Room reverberation causes frequency-dependent smearing of energies in the speech signal. The smearing happens in a frequency dependent way, such that typically the high frequencies get smeared less and the low frequencies get smeared more [27]. 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 we process each band independently, we can tune the feature extraction to the altered characteristics of speech in each band.

- The multi-band approach permits us to have variable sized bands, adaptable to the vocal tract parameters of the speaker. The formants of male speakers for example, are lower in frequency range than that of the female speakers. We could use this method to perform vocal tract normalization.

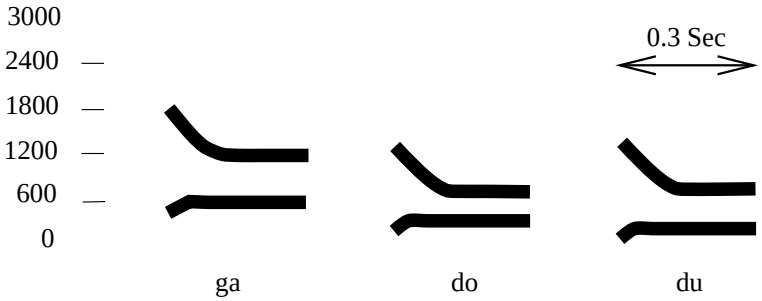

Figure 4: Synthetic spectograms using only F1 and F2 information that produced the voiced stops before various vowels. (after Delattre, Liberman, and Cooper 1955)

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

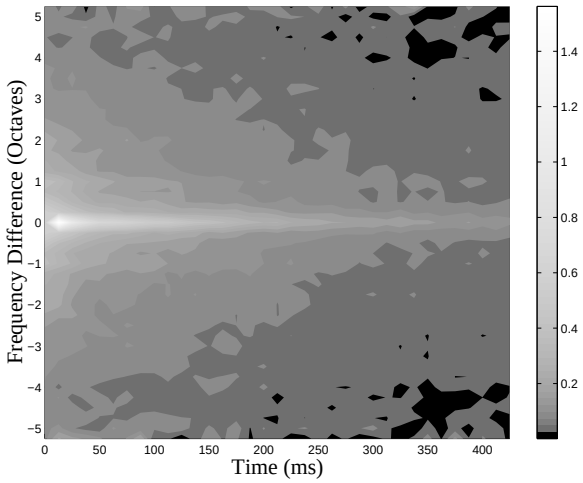

Figure 5: The time-frequency information density of a randomly selected 2-hour section of Switchboard, in bits per unit area (From Bilmes 1998).

2.2 Motivation for Merged Phonetic Categories

If phone classes are highly confused with each other in a sub-band, it might be desirable to merge them into one 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 to one class for the low frequency band, and not so for the high frequency

[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.

band. The main motivation for using such 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 between particular categories, whereas others may be devoid of sufficient information for such discrimination.

For example, figures 6 and 7 show the average posterior probability estimates for NUMBERS95 development set for phones /th/ and /ay/, respectively, for each sub-band. In Figure 6 we see that, band 3 is the best estimator for this phone, and band 1 is the worst as it confuses /th/ (phone index 21) with /t/ (index 5) and the silence phone (index 55). In Figure 7 we see that most bands, but particularly band 2 is a the best classifier for phone /ay/.

We hope to develop more accurate statistical models by focusing on the most prominent features present in each band, and by collapsing the highly confusable phone classes.

2.3 Why NUMBERS95 Corpus as a Test-Bed?

Much research in ASR has been done on read speech. Considering all the problems 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. We have chosen spontaneous speech as a test-bed for multi-band for the following reasons:

- Spontaneous speech is more 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. By dividing up the space of patterns into smaller units, we hope to increase our ability to find relevant structures in data sub-spaces, and to ultimately, increase our generalization power.

- It has been asserted [22] that spontaneous conversational speech has more temporal asynchrony than read speech. As one of the facets of multi-band 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.

3 Multi-Band ASR Research Issues

In this section we delineate various research problems in multi-band ASR. We briefly discuss the ones we choose to address, as well as the ones on which we have made design decisions.

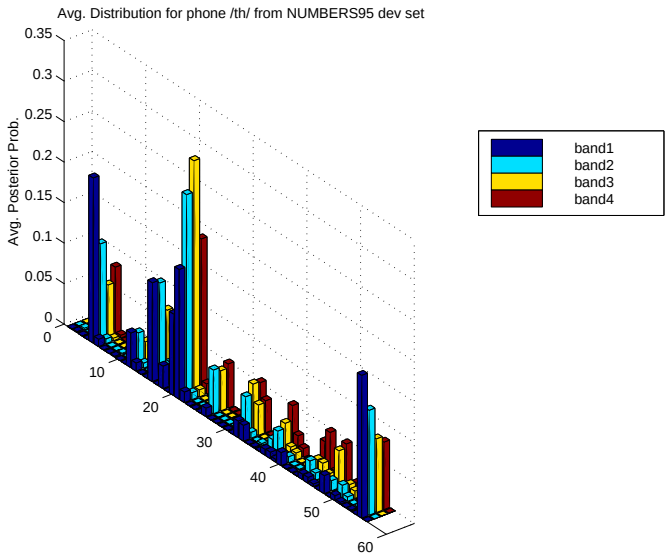

Figure 6: Average posterior probability estimates for all sub-bands for phone /th/ calculated on the NUMBERS95 development set.

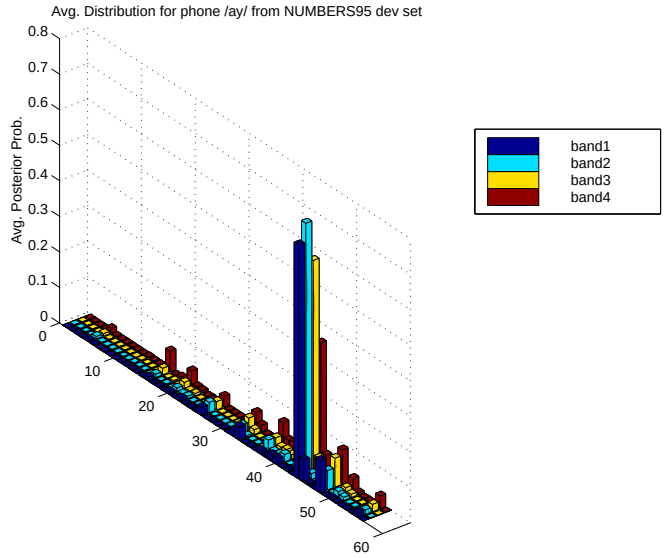

Figure 7: Average posterior probability estimates for all sub-bands for phone /ay/ calculated on the NUMBERS95 development set.

3.1 Selecting Sub-Bands

What is the optimal number of sub-bands and what should the cutoffs be? Duchnowski [18] chose four non-overlapping bands [100-700Hz], [700-1500Hz], [1500-3000Hz], [3000-4500Hz] roughly based on the formant regions. Bourlard and Dupont [7] 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 [17-778Hz], [707-1631Hz], [1506-2709Hz], and [2121-3769Hz] 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 psychoacoustic experiments on amplitude modulated noise in four frequency bands [60]. Tibrewala and Hermansky [67] observed that the error rates for a two- and four-band system were lower than that of a seven-band system.

One study which may be of interest in the sub-band boundary decision is that of Houtgast and Verhave [28]. 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 on a linear domain. On a logarithmic scale, however, there is less overlap between channels in the 1-4 kHz frequency range, suggesting that we should have more bands in 1-4 kHz because of the higher information content. Recall that for our experiments, we are using telephone quality speech which was low-passed around 4kHz.

The smaller the frequency region which a sub-band covers, the higher is 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.

Determining the optimal number of bands is not the focus of the work reported here. Based on previous experiments and psycho-acoustic observations, we choose four bands, with cutoffs based on the RASTA-PLP filter cutoffs (see Table 1), similar to that of [7], except that we discard the first two critical bands because telephone speech is band-passed above roughly 300Hz. Our experiments show that the first two bands do not help word recognition. Our four chosen bands are [216-778Hz], [707-1631Hz], [1506-2709Hz], and [2121-3769Hz].

3.2 Acoustic Features

Designing an optimal feature set for each sub-band is another research issue that warrants investigation. This is one of the topics addressed in Tibrewala's work [64], where she uses high and low modulation spectrum for capturing the long term variations (e.g., syllable level) and short term variations (e.g., phone level). Though still in an inchoate stage, spectral sub-band centroids as features have also recently been proposed in [50].

In our previous work [42], we used power spectrum values obtained after PLP critical band filter analysis, cube-root compression, and equal loudness equalization [23]. These critical band features performed more poorly than LPC-cepstra computed

Filter Number	Low Cutoff	Mid Freq	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 1: The half-power low and high frequency cutoffs for the RASTA-PLP filters when the sampling frequency is 8kHz.

on band-limited critical band values, as also observed by [7, 67]. In this work, we use RASTA-PLP processing [24] in each sub-band. Since RASTA-PLP processing emphasizes transitions (and because it has channel robustness characteristics), it is a particularly good candidate for this task.

3.3 DSP Window Size

One can imagine using a different feature extraction 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, where the energy smearing is less. For normal conversational speech, different window sizes may allow capturing different dynamics of speech such as short term and long term variations [64].

For our experiments, we choose equal window sizes for all bands: 25 ms for the signal processing window (computed every 10 msecs).

3.4 Level of Combination

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. Bourlard and Dupont [7]

have run preliminary experiments with state, phone, and syllable combination levels and their results were inconclusive. Cooke et. al. [14] have claimed that asynchrony is helpful, though open questions regarding the generalizability of their experiments remain.

Merging the sub-band information on the state level is fairly simple, and it involves estimating the overall phonetic probabilities for each frame given the posterior probability estimates of all narrow-band estimators. Merging on phone or syllable levels pose a bigger challenge, as it may be implemented using an algorithm such as HMM decomposition [69] or two-level dynamic programming [53]. Increased space and search time requirements may be a problem with these algorithms. We have started to look into their application in asynchronous merging of the bands.

3.5 Method of Combination

An important research issue 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.

Traditionally, non-linear methods have had higher accuracy compared to linear combination strategies. Non-linear merging (using an MLP) produce lower word error rates than linear merging (e.g., multiplying the probabilities) [7, 67] when multi-band information is merged on the frame level. We explore different ways of combining the probabilities (e.g., adding or multiplying the probabilities, adding their logs, etc.). Another question is whether the merging should be done in a posterior or likelihood domain. All previous experiments cited performed stream merges in the likelihood domain. We explore the above areas in this work.

Another question is in what domain should the merging occur: phonetic posteriors, phonetic likelihoods, or as recently explored in [49], feature vectors? Although there is some advantages to merging the feature vectors (e.g., the simplicity of having just one probability estimator), one can not take advantage of asynchronous merging on that level. Merging in the posterior and likelihood domains has been compared mathematically, not practically, in [9]. In this work, we will experiment with both methods.

3.6 Merged Phone Classes

Part of this work focuses on defining merged phonetic classes for each sub-band. The main hypothesis is that some frequency bands contain more information for distinguishing particular acoustic classes. The corollary of this hypothesis is that some bands have less or no information for distinguishing between some classes. Using only relevant information to distinguish among *limited* classes may be better than trying to categorize *all* classes using either limited information, or all (including noisy, perhaps) information. The former method (i.e., using limited narrow band information for discriminating between all classes for each sub-band) has been used in multi-band

approaches [42, 7, 25, 18], and the latter (using all available features for distinguishing between all classes) is the norm in the traditional ASR systems.

In order to determine the merged phonetic classes for each band, we performed band-limited phone recognition and created confusion matrices. We merged the highly-confusable categories in each band according to winner-takes-all, mutual information [15], and a simple greedy estimator. The details are discussed in Section 7.

3.7 Using Transition-based Classes

One of the most immediate extensions of the multi-band paradigm is the use of transition based phone classes, similar to those of Stochastic Perceptual Auditory-event Models (SPAM) [47]. The main idea is that as we divide the full-band into sub-bands, we better isolate the spectral transition regions. The mathematical modeling power of the recognizer should be more focused on distinguishing between states representing change than on states corresponding to little change (steady states).

First we define new classes based on transitions instead of phones, as a di-phone pair made up of the phone classes. Given n phone classes, we can have at most n^2 transition classes. Based on our observation from the training set, we can prune down the number of transition classes. The remaining research issues and design decisions are similar to that of the phone-based system.

3.8 Choosing a Database

NUMBERS [12] is a continuous speech database recorded over the telephone and sampled at 8kHz. It has been prepared at Oregon Graduate Institute from census information (phone numbers, birth-dates, zip-codes, etc.), and includes noise, non-speech sounds, and cut-off speech. A portion of the database is phonetically hand-transcribed. Table 2 shows a summary of the specs for this databases. For our experiments, we use a 6000 utterance subset (roughly 2.5 hours of speech) of the NUMBERS95 database which we will refer to as the core subset. Two hours of the data is used for training and cross validation, and about forty minutes for testing. All utterances in the core subset are chosen to have phonetic hand transcriptions and to exclude cut-off speech segments. Our baseline error rate for a full-band system is 7.9% for NUMBERS95 subset. Note that this baseline result is obtained using a simple single pronunciation lexicon with a uniform two state per phone duration, without the benefit of embedded training.

We have chosen the core subset of NUMBERS95, although we note that NUMBERS95 does not offer complete phonetic coverage, i.e., there are no voiced stops (phone /b/) and limited representation of the laterals (e.g., /l/ in eleven).

	numbers95 Subset
Number of words	32
Total Utterances	6K
Duration	4 hrs
Examples	'seventy', 'thousand'

Table 2: Some features of the NUMBERS95 subset.

4 An Analysis of the (Dis)Advantages of Multi-Band

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 section 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 between bands may be lost. Our experience and that of our colleagues has been that recognition performance has not been hurt by this approach, but in the work reported here we examine the estimator performance in a more detailed fashion. In particular, we analyze the phonetic feature transmission pattern in each sub-band, the merged multi-band, and full-band probability streams. As discussed in Section 4.2, we use methods similar to those of Miller & Nicely [41] and calculate confusion matrices for phone and feature classes, and use mutual information as a measure of information transmission in a channel.

In Section 4.3 we focus our attention on the following: some multi-band researchers [68, 8, 67, 43, 5] 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 analyzed; 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, we could not justify consideration of longer-term merging units for multi-band ASR. In Section 4.3, we examine this assumption by analyzing the transition lags in each sub-band to see if sub-band transitions occur asynchronously.

4.1 System Description

First, a few words on our HMM/MLP based [10] full-band baseline system: We train the MLP phonetic probability estimator on a 9 frame window of 8th-order

RASTA-PLP [24], energy, and delta-RASTA-PLP features for every 25 ms window, stepped every 10 ms. The MLP is fully connected and has 153 inputs (9 frames with 17 features per frame), 1000 hidden units, and 56 outputs (one output for each phone[4]), and is trained using backpropagation [56], with softmax normalization [11], and relative entropy error criterion [61] at the output layer. The system is trained on hand-transcribed phone labels (without embedded realignment). Using a multiple pronunciation lexicon (derived from hand transcriptions), and a bigram language model, the WERR of this baseline system on the test set is 7.9%.

For our multi-band system, we divide the frequency range into four bands of [216-778Hz], [707-1631Hz], [1506-2709Hz], and [2121-3769Hz], as discussed in 3.1. From the sub-bands, we derive [3rd, 3rd, 2nd, 2nd] order RASTA-PLP features, respectively, as well as energy and corresponding deltas. We train four MLPs on these acoustic features, that is, one on each sub-band. The input layer to each MLP has a context window of nine frames, for total input layer sizes of [72, 72, 54, 54] respectively. We choose hidden layer sizes of [497, 497, 372, 372], respectively, 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[1]. The frame-by-frame information from the four sub-band streams is combined using a *merger* MLP, which takes the output of the sub-band MLPs as input, has 300 hidden units, and an output of 56 phones[1]. The WERR on the test set for this merged multi-band system is 8.2%. The performance difference between the baseline and multi-band systems is not statistically significant.

4.2 Is Phonetic Information Lost?

4.2.1 Experimental Setup

The first question we want to answer is whether any phonetic feature information is lost in multi-band ASR. For this analysis we use phone and broad category confusion matrices, as in the seminal studies of Miller and Nicely [41] on human speech recognition.

A confusion matrix (CM) is simply an extended matrix of *hits* and *misses* for all classes, as in Table 3. The column headings represent the features we intend to *transmit*, and the row headings correspond to the *received* features. In Table 3, for example, 93 instances of /s/ are perceived as /eh/. We use frame level phonetic classification on the test set for generating phone CMs. To better observe the patterns in the data, we merge the phone CMs according to membership in broad category feature classes (as in Table 4), and generate feature confusion matrices (example in Table 5). We classify phonetic classes according to six broad categories: *CV* (consonant, vowel, silence), *duration* (short, long, mid), *frontness* (front, back, neither), *manner* (vowel, diphthong, liquid, glide, stop, closure, nasal, fricative, silence), *place*

[4]Note that some of the 56 total phones do not occur in the NUMBERS database and have zero priors.

(high, low, mid, labial, dental, coronal, palatal, retroflex, velar, glottal, silence), and *voicing* (voiced, unvoiced).

	t	s	eh	sil	...
t	5722	252	31	316	...
s	258	8495	110	1159	...
eh	11	93	3118	37	...
sil	436	2733	68	40237	...
...	...	...	...	...	...

Table 3: An example of a phone-based confusion matrix.

	vowel	consonant	silence
t	-	+	-
s	-	+	-
eh	+	-	-
sil	-	-	+
...	...	...	...

Table 4: An example of binary acoustic features for CV classification.

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

Table 5: An example of a feature-based confusion matrix.

To summarize the confusion matrix, we calculate mutual information (MI) for each CM [41] as $\sum_{i,j} p_{ij} log \frac{p_{ij}}{p_i p_j}$, where i is the feature we would like to transmit, and j is the feature that is perceived. We estimate the probabilities p_{ij}, p_i, and p_j 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.

We can further calculate the transmission of each phonetic sub-feature (e.g., sub-feature fricative $\in$ manner), by reducing the full CMs to a 2x2 CM for each

$sub\_feature$ and $\overline{sub\_feature}$) (the results in Figure 9). Finally, the maximum possible feature transmission for the idealized condition is the MI of a matrix of the same dimensions and with the class priors on its diagonal.

4.2.2 Observations on Feature Transmission

Figure 8 shows all features, and Figure 9 shows sub-features of *manner* transmitted as a percentage of the maximum. We observe the following:

1. Multi-band feature transmission is always as good as or better than the comparable full-band system, except for *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 54000 acoustic frames.

2. The results are consistent with our knowledge of acoustic phonetics, as, for example, we would expect the low frequency band to contain the most information about *voicing*. Comparing our results with [41], we observe similar patterns also for *fricatives* and *nasals*.

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, as the sum of information transmission over all bands far exceeds 100%. Lippmann [38] has highlighted this redundancy as a source of human robustness to speech degradations.

   In the next section, we examine the transition asynchrony hypothesis.

4.3 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. In this section we study this hypothesis.

4.3.1 Experimental Setup

In order to obtain the phone transition boundaries, we perform forced alignment on each sub-band independently. Furthermore, to allow maximum freedom of shifting in transition boundaries, we perform embedded realignment (i.e., Viterbi realignment and retraining the MLP in each iteration) for six iterations. The WERR on the NUMBERS95 cross-validation set is our stopping criterion, and it reaches a minimum value after the second iteration of realignment.

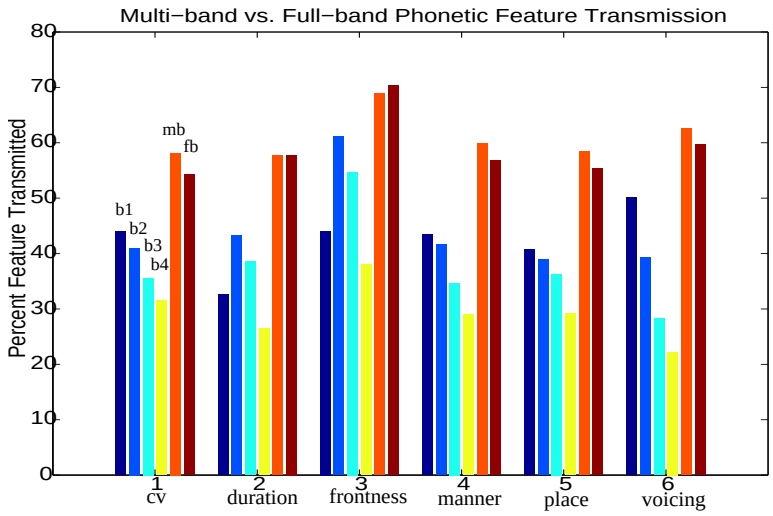

Figure 8: Phonetic features transmitted as a percentage of maximum possible, measured by mutual information.

Instead of using our usual multiple pronunciation lexicon, we use whole-sentence models in the forced alignment to insure that identical phone sequences are taken in each sub-band. We generate whole-sentence models using the phonetic hand-transcriptions and the corresponding average phone durations.

We also generate transition statistics on digitally-reverberated versions of the data, as well as on fast and slow speech. 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. 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 determined from a count of manually transcribed phones over non-silence regions.

For any given phone transition, we calculate the transition lags in each sub-band as compared to 1) the full-band, and 2) other sub-bands. Figure 10 show the histograms of average transition lags of the four sub-bands with respect to the full-band for broad phonetic categories, where each plot in row *feat1* and column *feat2* corresponds to a *feat1* $\longrightarrow$ *feat2* transition.

4.3.2 Observations on Asynchrony of Transitions

Examining the generated statistics, we observe that sub-band transitions do indeed occur asynchronously. More precisely:

1. Transition lags (with respect to the full-band transition boundaries) 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,

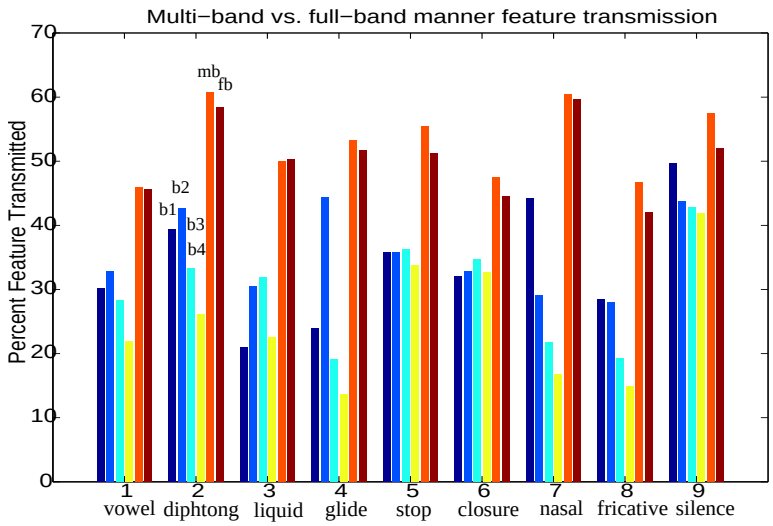

Figure 9: Manner of articulation features transmitted as a percentage of maximum possible, measured by mutual information.

5.0, 5.6] frames for the sub-bands, respectively. The higher the frequency range, the more shifted are the transition boundaries compared to the full-band.

2. More distant sub-bands have less agreement in transition boundaries, as the 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.

3. 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.

4. Some broad category transitions are sharp (e.g., sil $\rightarrow$ stop), and some have a relatively flat distribution (e.g., vowel $\rightarrow$ liquid) (see Figure 10 for more examples).

For contrast conditions of speaking rate and room reverberation, we also find strong changes in transition timing, as reflected in a modified variance rather than a systematic difference in the means. Table 6 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 most sensitive to speaking rate variations.

Table 6 further confirms our 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.

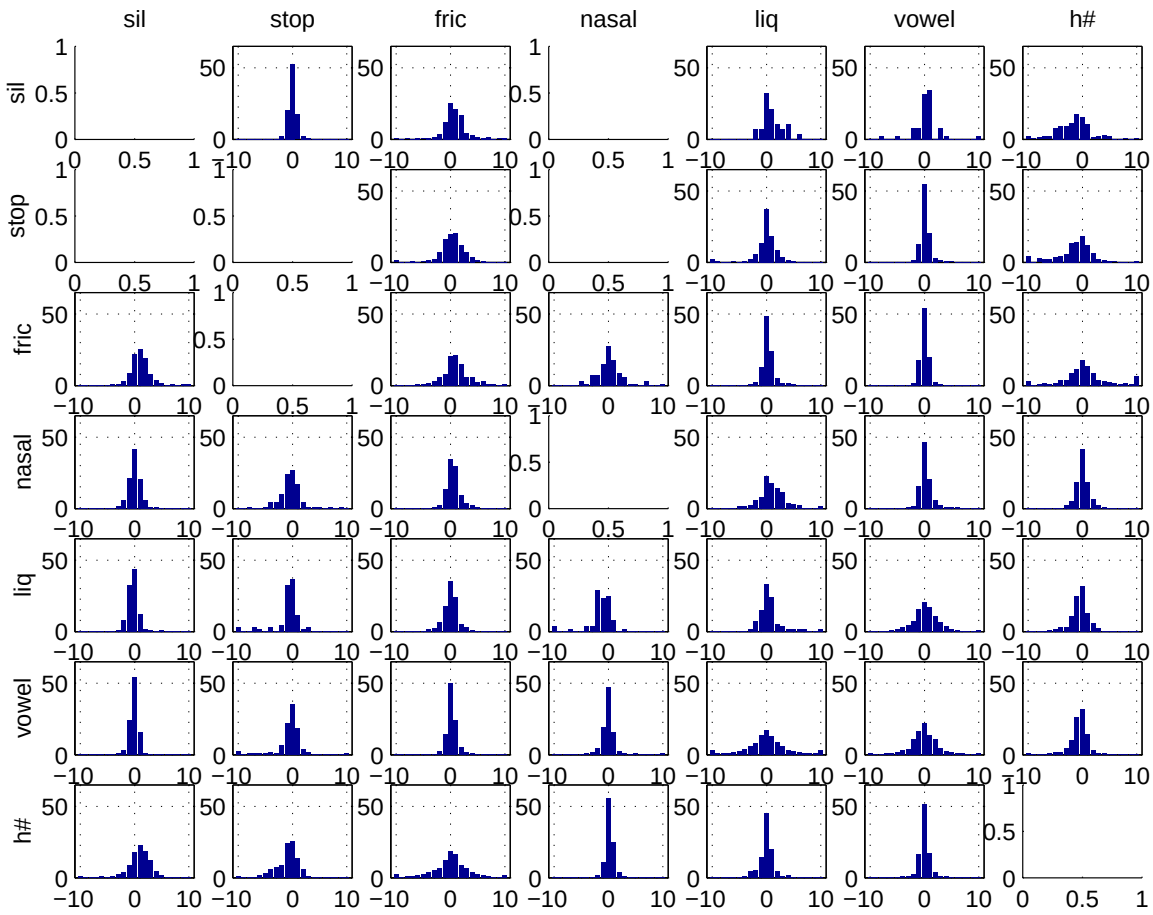

Figure 10: Histogram of average transition lags for broad phonetic categories for the four sub-bands. Each frame corresponds to 10 ms. /h#/ is the end/beginning of sentence silence.

4.4 Conclusions

We have tested two common assumptions on multi-band ASR: 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, we calculated phonetic feature *transmission* for sub-bands. Not only did we fail to substantiate the above objection, but we observed the contrary. We confirmed the second hypothesis by analyzing the *transition* lags in each sub-band.

Our 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

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

database) than the comparable full-band system (59.06%) for roughly 54K frames.

For the second question, we found that there is no consistent delay or advance of phone transitions in a frequency-dependent manner, as the per-band transition lags had a mean close to zero. However, the spread of these transition lags were both dependent on frequency and on contrast conditions (speaking rate and reverberation). In particular, 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, we would expect that there is a potential for improvements in acoustic modeling if longer time-scale information stream merging (i.e., phone or syllable) is used.

5 Designing a Multi-band System

As will be described later in Section 10.1, we performed some preliminary experiments [42] with the Bellcore Digits database in 1995, which have served as proof of concept experiments for multi-band ASR. As discussed in Section 3.8, in the work reported here we have used NUMBERS95, a continuous spontaneous speech database. In this section we discuss the experiments performed to choose parameters for our baseline multi-band system.

For our multi-band system, we divide the frequency range into four bands of [216-778Hz], [707-1631Hz], [1506-2709Hz], and [2121-3769Hz], as discussed in 3.1. As we noted, finding the optimal frequency cutoffs and number of bands is an open research issue. Based on previous experimental results, our current choice seems appropriate. This division results in [4,4,3,3] critical band filter values to cover the four regions (from low frequency to high frequency), respectively. To introduce some smoothing, each region is represented with one PLP order less than the number of points in that region. Hence, we derive [3rd, 3rd, 2nd, 2nd] order RASTA-PLP features, respectively, as well as energy and corresponding deltas for each narrow-band region. We trained

four MLPs on these acoustic features, that is, one on each sub-band. The input layer to each MLP has a context window of nine frames. The total number of inputs is [72, 72, 54, 54] for each MLP, respectively. We choose hidden layer sizes of [497, 497, 372, 372], so that the total number of parameters in the four MLPs and the full-band system are roughly equal (each has roughly 200K parameters), and the number of hidden units in each net are proportional to the input layer size. There are 56 output units, one for every phone, as in the full-band MLP[1].

5.1 Experimenting with Acoustic Features

As mentioned above, we derive [3rd, 3rd, 2nd, 2nd] order RASTA-PLP features, respectively, as well as energy and corresponding deltas from the narrow-band regions. In order to ensure that the narrow-band features we derive contain the information necessary for ASR, we train a *full-band* system on the narrow-band features. The WERR on this systems was 7.9%, compared to the WERR for the baseline full-band system trained on RASTA-PLP order 8 features (plus deltas and delta energy) of also 7.9%.

In our preliminary experiments in 1995 [42], we had expected spectral features to be good for multi-band recognition. However, we observed that they were outperformed by narrow-band cepstral features. For sake of completeness, we extracted spectral features and trained a full-band recognition on these 15 spectral features and their deltas. The WERR of this system was 7.9% on NUMBERS95 development set, which is similar to that observed using both full-band and narrow-band cepstral features for a full-band system. As we will see later in Section 5.3, however, our earlier observations were re-confirmed as cepstral features outperformed spectral features in our multi-band paradigm.

5.2 Varying the Sizes of Sub-band MLPs

To see if we could sustain the same level of accuracy by having fewer parameters, we reduced the number of MLP hidden units in the sub-band probability estimators. 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, we set the size of the hidden units to 200. As we see in Table 7, the frame error as well as word recognition error rate became worse. This implies that, for this experiment, we need about the same number of parameters in the multi-band system to have a performance comparable to that of the full-band system. Since we didn't wish to have a multi-band system with many more parameters than the full-band system, we did not increase the number of hidden units.

	Larger MLPs		Smaller MLPs	
	Frame Err.	Word Err.	Frame Err.	Word Err.
Band1	40.2	33.7	40.8	34.2
Band2	37.4	24.9	38.9	27.8
Band3	42.5	34.4	44.6	36.6
Band4	49.5	47.8	49.5	48.1
Multi-Band	20.9	8.3	20.6	-

Table 7: The word and frame error for multi-band systems with different number of parameters. The frame error is measured on the NUMBERS95 cross validation set, and the word error 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.

5.3 Merging the Sub-Band Results on the Frame-Level

Merging the sub-band results on the frame-level corresponds to ignoring the asynchrony and merging the streams on a frame by frame basis. One advantage of this is that it is easy to perform, simply by estimating the overall phonetic probability given the sub-band probability distributions. We chose to experiment with linear and non-linear merging. For linear merging, we tested sum of likelihoods ($\sum likelihood$) and product of likelihoods ($\prod likelihood$). The WERRs on NUMBERS95 were 15.9% and 13.3% respectively. Product of likelihoods seemed to be the best simple linear merging scheme. For sake of completeness, we developed a similar multi-band system using 15 spectral features and their deltas (see Section 5.1). The WERR on NUMBERS95 development set using product of likelihoods was 14.1%.

For non-linear merging, we experimented with various sizes of a two layer MLP. We varied two parameters: the size of the hidden layer, and the size of the input window. The size of the input window determines how many frames of past and future probability distributions we consider for classification of the current frame.

As we see in Table 8, the number of parameters in the merger MLP increase dramatically as the input window size is increased. It appears that increasing the number of hidden units and keeping the window size at one is a more efficient way of allocating extra parameters. For example, an MLP merger with 50 hidden units and window size of three has roughly the same number of parameters as an MLP merger with 300 hidden units and a window size of one; however, the word error rate of the former (9.6%) is significantly higher than that of the latter (8.3%). We chose 300 hidden units and a window of one as the parameters of our merger MLP.

As a side-note, it has been suggested by Fred Jelinek [30] that training the merger MLP on the same training data which the sub-band MLPs were trained may lead to over-training. As an alternative, we trained a 50 hidden unit merger MLP with a

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

Table 8: The word error rate on the NUMBERS95 development set as the number of hidden units and input window size in the merger MLP are varied. The merger MLP was trained on the NUMBERS95 training set.

window of nine frames on the cross validation data. The word error on the development set was 12.6%, which is significantly worse than what is reported in Table 8 for the same size MLP trained on the training data (9.7% error rate). The reason may be that the cross validation set is too small and does not provide enough variation of training samples. As NUMBERS95 is a small database and training data is scarce, we cannot adequately test the above hypothesis. Training the merger on an independent and large training set should be addressed on a larger database.

Another 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 which may improve the probability distributions. To test this, we trained another MLP (with 300 hidden units and a window of one) on the training data, and stopped the training according to the cross validation frame error. The input to this net was the output of the full-band MLP, and the output is the posterior probability of the 56 phones. To our surprise, the word error of this full-band system increased to 9.0% (up from 7.9%). Perhaps training on the same data will result in over-training. We note that Tibrewala [64] observed an improvement when she trained a similar probability estimator on an data set different from the training set.

Another parameter we tested was the difference between using likelihoods versus posterior probabilities. Up to now our colleagues have merged the likelihoods only. Again, we trained a merger MLP with 300 hidden units and a window of one, either

on the likelihoods of the sub-band MLPs or the posterior probabilities. The error rate on the NUMBERS95 development set was 8.3% for merging posterior-s and 8.6% for merging likelihoods. Although the differences are not significant for this dataset, for a larger database we may actually see significant differences. Merging posteriors actually may make more sense, as dividing posteriors by priors may exclude prior information.

5.4 Performing Embedded Alignment

We wondered about the effects of embedded training [71] on our multi-band system. Specifically, would the alignment of the phone labels, and therefore the overall accuracy, improve if we align the phone labels in each band iteratively and then align the phone labels in the merger MLP iteratively? We ran embedded training on each band without retraining the lexicon using the new labels, and stopped after the third iteration (according to average WERR on NUMBERS95 cross validation set). Then, using the feed-forward probabilities form aligned sub-band systems, we performed embedded alignment on the merger MLP, stopping after two iteration (according to the same criterion). The final WERR for NUMBERS95 development set was 8.5%, which is not statistically different from the unaligned system with WERR of 8.3%. One reason may be that the phone labels in high frequency sub-bands (which are 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 will avoid this potential problem. We are look into this further.

5.5 Merging Probability Distributions or the Best Phone Label?

As mentioned in Section 10.1, one main difference between the work of Duchnowski and ours is that he only retained the identity of the best phone in each sub-band instead of the full probability distribution. We experimented with discarding the full distribution (by substituting zeros for the non-max probabilities) and only retaining the phone with the highest probability in the sub-band streams. The WERR of the multi-band system increased to 15.2% (from the baseline 8.3%) on NUMBERS95 development set.

5.6 Combining the Probabilities of Full- and Multi-band Systems

Research at ICSI on merging streams of information [72, 48] has shown that often merging the probabilities of the conventional full-band system with another experimental system with different error characteristics leads to improvements in the overall system performance. When linearly merging by simply multiplying the posterior probabilities from the full band and multi-band streams, we observed a WERR 7.1% on NUMBERS95 development set (better than each stream independently.

As non-linear merging is usually superior in performance to linear merging, we tried to combine the probability estimates from the full- and the multi-band systems by training a 50 hidden unit merger on the output probability distributions. The word error rate of this combined system on NUMBERS95 development set was 10.5%, which is significantly higher than that of either system alone. This poor performance might be due to the MLP merger having too few parameters to adequately capture the distribution.

6 Testing Frame-Level Merging on Degraded Speech

We tested our baseline multi-band system on reverberant, fast, and noisy speech to observe any potential robustness to these degradations.

We generated a digitally-reverberated versions of the data 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. We trained the system on clean speech and tested on reverberant speech. The word error on NUMBERS95 development set was 32.2% for the full-band system and 39.9% for the base-line multi-band system (i.e., MLP merger with 300 hidden units and an input window of one). Linearly merging the probability streams by simple frame by frame multiplication, the WERR was 44.4%.

To determine fast speech, we set the threshold to one standard deviation above the mean rate of the training set. The speaking rates were determined from a count of manually transcribed phones over non-silence regions. Based on this criterion, 206 out of 1206 sentences in the development set were considered fast. The word error rate on the fast sentences for the full-band system was 13.9%, and for the multi-band system was 14.2%. The difference is not statistically significant.

Based on the above limited experiments it appears that frame-level merged multi-band does not seem to have an inherent advantage over a full-band system for reverberant and fast speech. We will look into robustness to other types of noise in the future.

7 Merged Phone Sets Experiments

In this section we focus on defining merged phonetic classes for each sub-band. Our intuition is that some frequency bands contain more information for distinguishing particular acoustic classes. The corollary of this hypothesis is that some bands have less or no information for distinguishing between some classes. Therefore, using only relevant information to distinguish among *limited* classes may be better than trying to categorize *all* classes using either limited information, or all (including noisy, perhaps) information.

For example, as we see in Figure 11, phone /ow/ is recognized correctly 72% and 36% of the time in bands two and four, respectively, as is mostly confused with phones

/h#/ and /ay/. Phone /f/, on the other hand, is recognized correctly 53% and 21% of the time in bands two and four respectively, and is confused with many phones such as /h#/, /n/, and /s/. Now, the question is, would collapsing confused classes decrease the confusion?

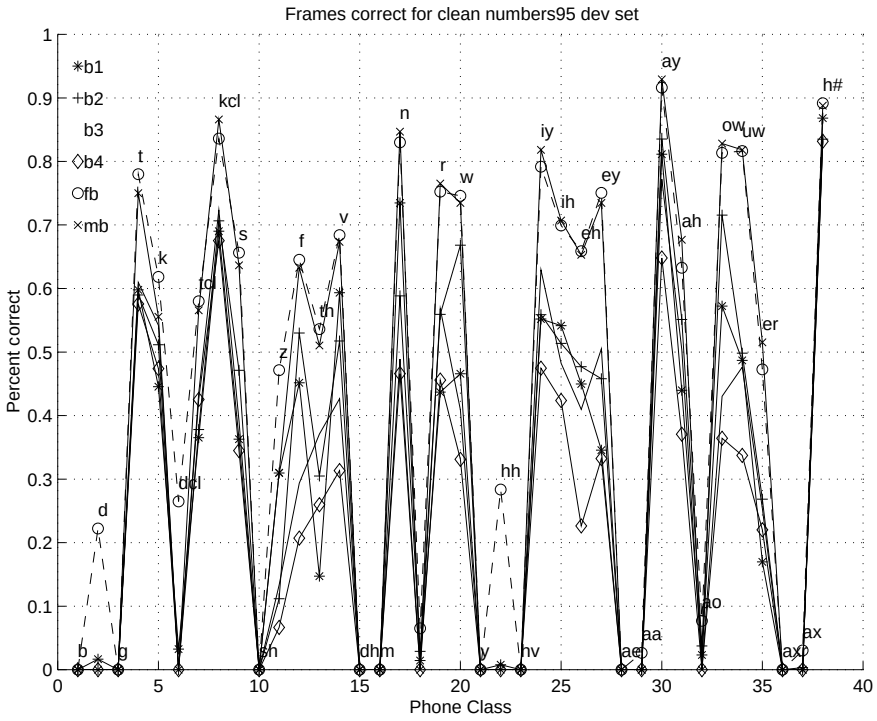

Figure 11: Frame accuracy for sub-band, multi-band, and full-band systems on the NUMBERS95 development set.

In order to determine the merged phonetic classes for each band, we performed band-limited phone recognition and created confusion matrices for each sub-band (included in the Appendix). A confusion matrix (CM) is simply an extended matrix of *hits* and *misses* for all classes, as in Table 3. The column headings represent the features we intend to *transmit*, and the row headings correspond to the *received*[5] features. In Table 3, for example, 93 instances of /s/ are perceived as /eh/. We used frame level phonetic classification[6] on the test set for generating phone CMs. We merged the highly-confusable categories in each band according to winner-takes-all, mutual information [15], and a simple greedy estimator. Our hope is that the pattern of class merges will be different in each band so that we will be able to distinguish the identity of the phone class when the sub-band information is merged. The following sections discuss these experiments.

[5]The received phone is the one with the maximum posterior probability in a given frame.

[6]The phone classification for each frame is determined based on a winner-takes-all strategy from the phonetic distributions for each frame.

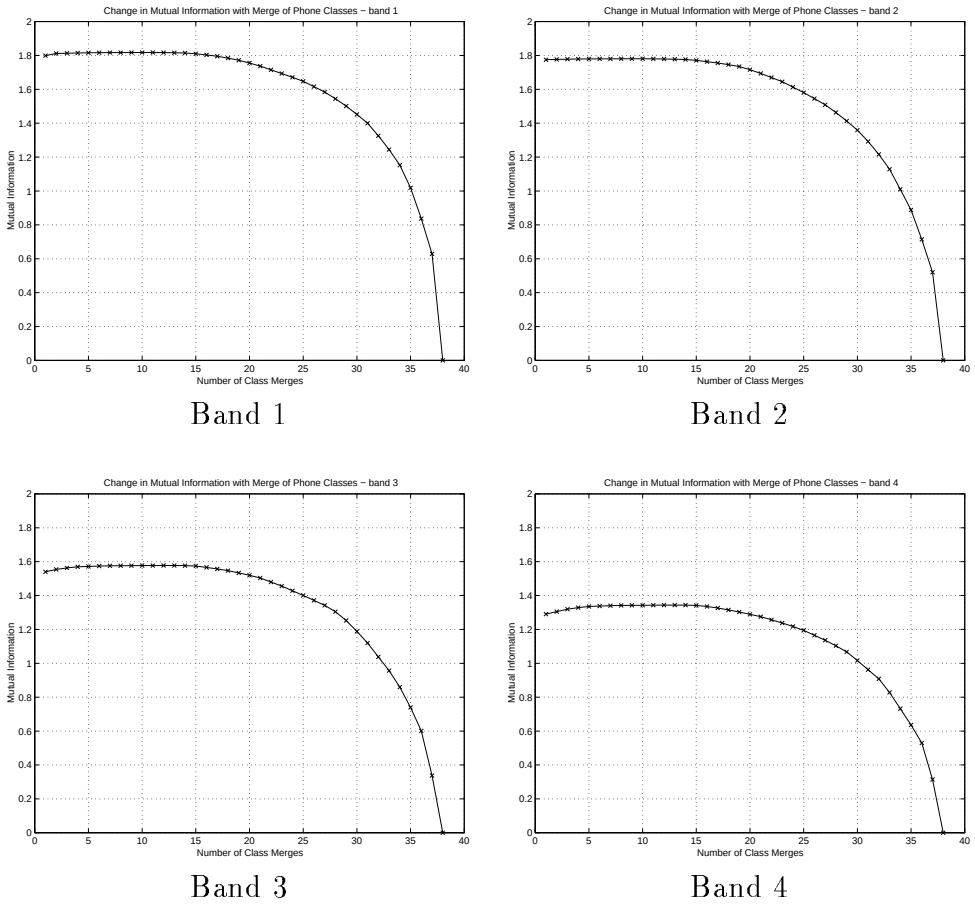

Figure 12: Change in mutual information for bands 1 through 4 as phone classes are merged.

7.1 Mutual Information as the Criterion

The first idea that we explored was to greedily merge the most confused phone classes, as defined by mutual information (MI). Our simple algorithm calculates the change in mutual information for every possible pair merge and then chooses the merge with the highest gain (or the least loss) in MI. In other words:

For every pair of classes C_i and C_j
    Calculate the change in MI (ΔMI) if C_i^* and C_j^* were merged
Choose C_i^* and C_j^* with the highest ΔMI
Merge class C_i^* and C_j^*

As we see in Figure 12 MI stays constant for roughly 13 to 18 class merges, and then decreases. Our hope was that different set of classes would be merged in different

bands. Upon examining the merged classes (shown in Table 9), we see that the same set of classes are merged in every band. More precisely, if we stop the merging just before the MI falls below the initial MI level, the largest merged classes in each band are made up of:

- Band1: aa, ae, ax, axr, b, dh, g, hv, m, sh, and y
- Band2: aa, ae, ax, axr, b, d, dcl, dh, hh, hv, sh, and y
- Band3: aa, ae, axr, b, d, dcl, dh, g, hh, hv, m, sh, and y
- Band3: aa, ae, ao, ax, axr, b, d, dcl, dh, g, hh, hv, l, m, sh, and y

Not only are the same classes merged in each band, but more disturbingly, it appears that the classes with the smallest priors have been chosen (see Table 10). Furthermore, the collapsing of phone classes with significant priors appears to lower the MI significantly. It is not surprising that classes with small priors are the ones which are most easily confused, as there are too few learning examples for adequate training. The following two questions arise: would the collapsing of infrequently occurring and highly confused phones actually decrease the word error? Secondly, is mutual information the right metric for choosing classes to merge?

To reduce on level of complexity, we decided to answer this question in regards to a full-band experiment. Surely, if collapsing such classes is to be useful in multi-band, we would see similar benefits for a multi-band system.

We performed class merges using the CM generated from the probability distributions of the full-band system. This system was identical to the one discussed in Section 4.1, except that the MLP has 400 hidden units instead of 1000.

The change in MI was similar to the sub-band cases, that is, the MI continued to stay relatively unchanged and after approximately 15–18 merges started to decrease rapidly (see Figure 13). All class merges are reported in Table 12. To run our feasibility experiment, we chose the first 15 phone class merges which merged the least-occurring most-confused phones into one large class. These merged phone classes were: aa, ae, ao, ax, axr, b, d, dcl, dh, g, hh, hv, l, m, sh, and y. In other words, we defined a broad class which was a super-class of all these low frequency classes. We changed the lexicon accordingly, that is, we substituted the label for the super-class for every occurrence of the above sixteen classes.

As we see in Table 11, the word error of the merged-phone system (measured on NUMBERS95 development set) is slightly poorer, though the difference is not significant. Examining some differences in error patterns, we see, for example, that the word "forty" has been mis-recognized (as"thirty"), as /dcl/, /d/, and /ao/ have been merged into the super-class phone.

It appears that although merging the most highly confusable classes may make the discrimination task easier (as seen in the frame error), it may still increase the recognition word error rate since confusion among the word models in the lexicon also increase.

	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 9: An ordered list of phone class merges in each band.

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 10: NUMBERS95 phonetic prior probabilities for the development set.

Test Condition	Word Error (sub/del/ins)
Unmerged FB	8.5 (4.8/1.6/2.2)
Merged FB	8.9 (5.2/1.7/2.0)

Table 11: Word error for numbers95 core-set development set; full band feats. are RASTA-PLP8, deltas, and delta energy, MLP has 400 hidden units.

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

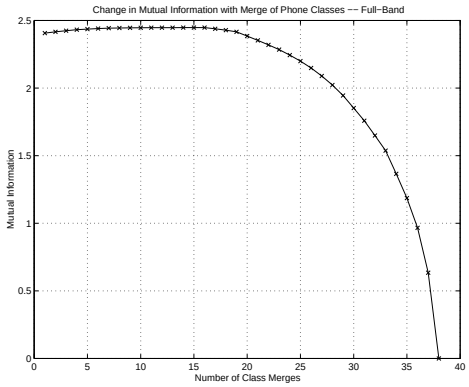

Figure 13: Change in mutual information for bands 1 through 4 as phone classes are merged.

7.1.1 Conclusion

If the full-band system did not improve with this modification, it is unlikely that the multi-band system will improve. One problem may have been that the decision to merge the classes is a local one, whereas maybe we should make a global decision based on the decision of the other bands. Either collapsing classes is not a good idea or using MI as the criteria is not good since it does not capture any idea about the priors. In the next section, we experiment with other criteria for collapsing classes.

7.2 Using Global Discriminators for Choosing Merged Classes

Using global discrimination 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. Using an MLP as the global merger may yield the best optimization. However, the insurmountable computation and time requirements make this an infeasible task, as training an MLP (on the full, unmerged, phone probabilities) on our current sub-band system takes approximately 2 hours, and there are roughly 350K configurations[7]. Due to practical considerations (e.g., hope of graduating in this millennium), we decided to use simpler merger functions. One such simple function is the nearest neighbor algorithm. Means and standard deviations of the features are calculated from the training data. Class membership for each test feature vector is decided based on the distance to the class means. The distance is calculated using one of the following distance metrics:

[7]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, as 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.

- Manhattan $\sum_i |p(x_i) - q(x_i)|$
- Euclidean $\sum_i \sqrt{(p(x_i)^2 - q(x_i)^2)}$
- Relative entropy $\sum_i p(x_i) log \frac{p(x_i)}{q(x_i)}$

Note that the input features for this discrimination task are the probability distribution output of the sub-bands. As demonstrated in Figure 14, we calculate means of the feature vectors from the sub-band output probabilities.

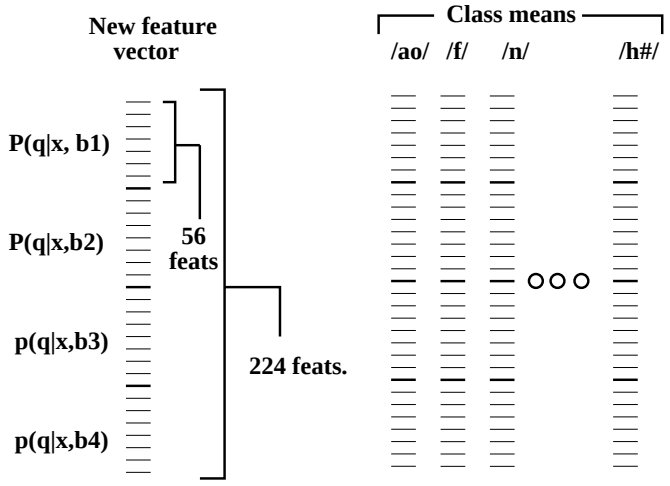

Figure 14: Visual demonstration of the nearest neighbor calculations.

To have fast turn-around time for this experiment, we limited the number of frames used. We use half of the cross-validation data (roughly 30K frames) for training (finding the means) and the remaining half for deciding which classes to merge.

First, we performed word recognition experiments on the full-band setup to observe how these simple discriminators perform in comparison to the MLP merger. The word error on NUMBERS95 development set using relative entropy was 11.6% compared to 8.4% using a fully connected MLP with 400 hidden units. The increased word error for the linear system may also be partly due to the fact that it has been trained on roughly 1/20th of the amount of the data used for MLP training. Also, the number of parameters in the nearest neighbor systems is 500 times less than for the 400 hidden unit MLP.

We choose classes to merge as follows: In every iteration, every class pair in each sub-band is considered for merging, the means are recalculated assuming the merge of the pair, the test set is classified using this new configuration and the frame error for the test set is calculated. The class pair leading to the highest decrease in frame error (or lowest increase) is chosen, and the best classes are merged. The process above is repeated in a greedy fashion until all classes are merged. In Figure 15, we see the change in frame error upon successive merges. Table 13 shows the final frame scores.

Frame Error	Manhattan	Euclidean	Rel. Entropy
Baseline	38.0	30.7	28.2
Merged	28.8	29.4	27.0
Difference	9.2	1.3	1.2

Table 13: The decrease in frame error as phone classes are merged according to Manhattan, Euclidean, or relative entropy error criteria using a nearest neighbor algorithm.

The percentage reduction in the number of classes are 28.8%, 43.1%, and 35.0% for Manhattan, Euclidean, and relative entropy, respectively.

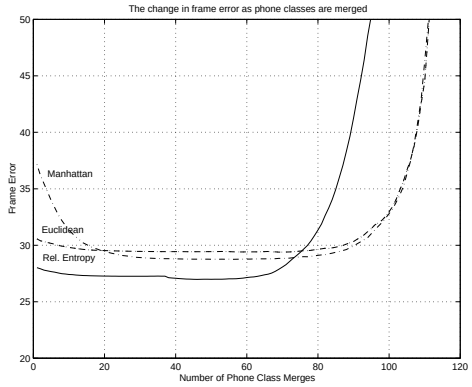

Figure 15: The change in frame error in NUMBERS95 cross validation set as phone classes are merged according to Manhattan, Euclidean, and relative entropy error criteria in a nearest neighbor search.

On each graph in Figure 15, we chose the merged configurations at the point at which the frame error was at a minimum. Hence, the number of merged classes were 56, 69, and 46, reducing the total number phone classes by 29%, 43%, and 35% for Manhattan, Euclidean, and relative entropy criteria, respectively. It is interesting to note that upon collapsing classes, frame error decreases significantly ($p < 0.0005$). Since the priors are taken into account [8], we see that infrequently occurring highly confusable classes are *not* merged first (as opposed to the MI merging criterion previously). As we examine the merged phone patterns, we see 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 we see in the confusion matrix for

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

Condition	Baseline	Manhattan	Euclidean	Rel. Entropy
Percent class reduction	0	28.8	43.1	35.0
WERR (w/ MLP merger)	8.3	9.9	10.0	9.8

Table 14: The word error rate on NUMBERS95 development set for multi-band systems with a frame-level merger MLP trained on the merged classes. The collapsed classes have been chosen using a nearest neighbor algorithm using Manhattan, Euclidean, and relative entropy error criteria.

Word Error	Manhattan	Euclidean	Rel. Entropy
Baseline	12.5	51.8	11.6
Merged (w/ nearest neigh. merger)	13.5	-	11.5

Table 15: The recognition word error rate comparison for baseline and collapsed multi-band systems. The frame-level merging function is the nearest neighbor algorithm using the Manhattan, Euclidean, and relative entropy distance criteria.

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 in the Appendix.

Next, we wanted to compare the performance of a multi-band system trained on the new merged classes with the baseline multi-band system. We trained MLPs mergers on the new merged classes. Recognition word error rates of the merged-class systems were worse compared to that of the baseline multi-band system. Table 14 shows the recognition word error rates.

The reason for the increase in error may be that we are using one touchstone (e.g., relative entropy and nearest neighbor) for choosing merged classes, and another (i.e., sigmoid functions in the MLP) for the phonetic probability estimation. To rid the experimental setup of this potentially problematic inconsistency, it would be best to use the same function for both tasks. As we noted earlier, using the MLPs for choosing merged classes is unrealistic because of time requirement. However, we could use the simple nearest neighbor discriminators for phonetic probability estimation if we keep in mind that the overall results will be worse. A comparison of the baseline multi-band results (no classes merged) with merged multi-band results are reported in Table 15.

It appears that although the frame error for the merged system is lower, the word recognition error rate is either similar or slightly poorer.

Condition	unmerged	merged
Fast Speech	20.2	21.2
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 16: Word error rate for NUMBERS95 development set using relative entropy based nearest neighbor probability estimator in degraded speech conditions.

7.2.1 Degraded Speech Results

It is perhaps not surprising that using coarser class definitions does not improve performance for the clean test data. However, another hypothesis is that in degraded speech conditions, distinguishing between different phonemes becomes more difficult. Hence, if collapsing confused phone classes decreases the confusion in discriminating phone classes, this advantage should help improve degraded noisy conditions. We tested fast speech, reverberant speech, speech with Volvo and pink noises. The results are reported in Table 16.

We generated a digitally-reverberated version of the data as described in Section 6. We trained the system on clean speech and tested on reverberant speech. To determine fast speech, we set to one standard deviation above the mean rate of the training set. The speaking rates were determined from a count of manually transcribed phones over non-silence regions. Based on this criterion, 206 out of 1206 sentences in the development set were considered fast.

The performance of the merged phone system slightly worse than the non-merged phone system for Volvo noise, pink noise, and fast speech, though the differences are not statistically significant. For reverberant speech, however, the merged phone system's performance as significantly better.

7.3 Conclusions

In the previous sections, we described our experiments with merged phone classes in multi-band. Although merging phones seems to reduce the frame error, it neither improves nor degrades word recognition performance significantly. Our main intuition in collapsing phone classes was that some bands do poorly in discriminating between some phones, and thereby increasing the overall randomness and error in the phonetic probability estimation. We had hoped that by making the sub-band-local discrimination task easier, we could improve the overall frame error, and improve recognition performance. We observed that as classes were merged, although frame accuracy showed a decreasing trend, the word recognition accuracy did not follow

suit, probably because the pronunciations of some words became similar.

It may be that in the baseline system the shortcomings of one band is countered by other excelling bands, so that the poor discriminatory abilities of one band do not compromise the overall performance.

8 A Summary of Conclusions

In this technical report we summarized our efforts on advancing the state of knowledge on the multi-band approach. We performed analysis of multi-band feature transmission and transition asynchronies, studied different design parameters for a baseline multi-band system, and explored multi-band paradigm based on merged sub-band classes. Our conclusions are as follows:

- Contrary to the objection of some multi-band ASR critics, we did *not* observe multi-band ASR to be inferior to a full-band approach because of phonetic information being lost due to the division of the frequency space into sub-bands; in fact, even when using a simple multi-band merging method, phonetic features are transmitted better (47.15% for our database) than the comparable full-band system (45.48%).

- Although the phone transition sub-bands are not consistently delayed or advanced in a frequency-dependent manner, they are dependent on both frequency and contrast conditions (speaking rate and reverberation). In particular, roughly one-third of the sub-band transitions in the control condition do not occur within 50 msecs of each other. Furthermore, the high frequency band timings have a spread that is strongly dependent on speaking rate. Thus, we would expect that there is a potential for improvements in acoustic modeling if longer time-scale information stream merging (i.e., phone or syllable) is used.

- Although merging phone sets seems to reduce the frame error, it neither improves nor degrades word recognition performance significantly. We observed that as the number of phone classes were drastically reduced, even though frame accuracy showed a decreasing trend, the word recognition accuracy did not follow suit. It may be that in the baseline system the shortcomings of one band is countered by other excelling bands, so that the poor discriminatory abilities of one band do not compromise the overall performance. There was some improvement observed for reverberant speech, however.

- There were many parameters to adjust and spaces to search for designing our baseline multi-band system. In our parameter search, we observed that:

  - Narrow band RASTA-PLP features are better than spectral features.
  - A non-linear scheme is better than a linear one for frame by frame merging of the sub-band probabilities.

- When merging the sub-band on a frame by frame basis, merging the probability distributions of the sub-bands is superior to only retaining the identity of the phone with the highest probability.
- In allocating parameters in the merger MLP, it is best to increase the size of the hidden layer and not the input window size.
- Embedded alignment of each sub-band MLP independently and then the merger MLP did not help in our experiment.
- Combining the probability streams of the full-band and the multi-band systems gave us our best results. Our guess is that the errors of the two systems are somewhat orthogonal, though we need to perform more analysis.
- The multi-band system did not perform better than the full-band system on fast and reverberant speech. They were not tested on cases that have been clearly shown to be advantageous for multi-band recognition, most notably narrow-band noise.

9 Future Work

9.1 More Work on the Baseline System

Although we spent much time searching the space of basic parameters, clearly there is much more work that may be done to improve the basic multi-band system. A few are:

- As discussed above, performing embedded alignment only on the merger MLP. There is some anecdotal evidence that varying the input window size may be helpful in this case.
- Varying the number of bands, for example developing a two-band system.
- Using different set of features.
- Experimenting with different merging criteria.

9.2 Merging Bands on Longer-Time Scale

In our analysis in Section 4.3, we observed that the phone transitions in each band occur asynchronously. Allowing the streams in each band to merge asynchronously on a longer time scale seems promising. The HMM decomposition algorithm [69] and the two-level dynamic programming [53] are viable algorithms for implementing asynchronous merging. We have started experimenting with these algorithms, though it is too early for us to make any conclusive statements about the results.

9.3 Transition-Based Phone Models

Transition-Based (TB) Phone Modeling, a simplified version of the Stochastic Perceptual Auditory-event Model (SPAM) [47], is one of the most immediate extensions of the multi-band paradigm. 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 transitions in each sub-band may occur asynchronously (see Figure 4), and by breaking the frequency bands into sub-bands, we can determine regions of significant change more accurately.

Acknowledgments

My gratitude to Nelson Morgan for direction, advice, support, not to mention his careful reading and valuable comments on this TR. I would like to thank Brian Kingsbury for the CM script; Eric Fosler-Lussier, Su-Lin Wu, and Dan Gildea for helpful discussions on lexicon creation and decoding; Jeff Bilmes on the discussion of his results as extra motivations for multi-band; and other ICSI Realization group members for various exchanges. I wish to acknowledge our colleagues Hervé Bourlard, Stéphane Dupont, Steve Greenberg, Hynek Hermansky, and Sangita Tibrewala for multi-band collaboration and numerous exchanges of ideas over the past two years. Thanks to Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, now at the University of California, Davis, for collecting the room impulse responses and making them available to us. This work was supported by Mentored Research Fellowship by the University of California, European Community Basic Research grant (Project Sprach), and the International Computer Science Institute.

10 Appendix A – Related Work

10.1 Previous Work on Multi-band ASR

In this section, we will discuss some related work on multi-band approach: the PhD thesis of Paul Duchnowski [18], our earlier work, and the work of our collaborators, Bourlard and Hermansky and their students.

The first work published on the multi-band ASR has been by Paul Duchnowski [18] at MIT as his PhD thesis. His goal was to apply multi-band processing to the task of phonetic, speaker-independent, phone recognition as cue-ing aid for the deaf. He divided the frequency band into four non-overlapping bands [100-700Hz], [700-1500Hz], [1500-3000Hz], [3000-4500Hz] loosely based on the formant regions. His experiments were done on the TIMIT phone recognition task. The focus of his experiments were on comparing different acoustic feature sets (LPC, cepstrum, autocorrelation, and their deltas), choosing the best subset of the features, number of code-books, and ways of combining the results of the band-limited recognizers on a frame by frame basis. The highest phoneme recognition accuracy achieved on the TIMIT test set was 58.5%, which was within the range of the performance achieved by the established phonetic recognizers [34, 29, 74].

Our work on multi-band processing has more differences than similarities with that of Duchnowski. Clearly the basic idea of multi-band ASR is similar: acoustic features are derived from narrow bands and fed to local decision makers, then the local decisions of the narrow-bands are merged to form the final recognition decision. There are at least two fundamental differences in our approach: In Duchnowski's work, local sub-band information is in the form of phone labels (the phone with the maximum probability is declared the winner), whereas in our approach, sub-band stream is a probability distribution over all phones. Secondly, our goal is to merging of the sub-band streams asynchronously, whereas in Duchnowski's work, local decisions were merged on a frame by frame basis, albeit allowing some smoothing. More minor differences are our choice of test-bed (telephone quality continuous speech recognition versus phone recognition task), speech recognition systems (a hybrid HMM/MLP paradigm versus an HMM ASR system), band overlap (i.e., our narrow-bands have a limited amount of overlap by the virtue of overlapping critical band filters), and acoustic features (RASTA-PLP versus LPC).

More recently, work by us [42] and our collaborators Bourlard and Dupont [7, 8] and Hermansky and Tibrewala [25, 67] has focused on multi-band for continuous speech recognition. Comparable or better performance for normal speech, and superior performance for band-limited noisy speech were demonstrated. We briefly summarize these results below.

The goal of our first set of experiments (in 1995) [42] was to establish a proof of concept by developing a two-band multi-band system for word recognition. For features we chose a vector of 15 power spectrum values obtained after PLP critical band filter analysis, cube-root compression, and equal loudness equalization [23]. To

keep the turn-around time of our experiments short, we chose the Bellcore Digits database for testing. The Bellcore digits database is 25 minutes of speech comprising 13 words: {*zero, one, two, ..., nine, oh, yes, no*}, spoken by 209 speakers and is recorded over the telephone (sampled at 8 KHz). To choose the optimal cutoff frequency for the multi-band system[9], we trained 29 systems, each of which was trained on either a low-pass or high-pass condition. The optimal frequency cutoff point of 1400 Hz was the intersection of the low-pass and high-pass word error curves (see Figure 16). The intersection point may be considered as the frequency threshold 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 by Miller and Nicely [41] in their psycho-acoustic experiments. We 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 had 135 input, 200 hidden, and 61 output units (40K parameters). The low-pass system had 63 and the high pass system had 72 input, 200 hidden, and 61 output units (keeping the total number of parameters at 40K). The merging was simply done on the word level by training an MLP on the normalized log likelihoods obtained from the Viterbi decoding distances. The word error rate of the full-band was 4.6% and the two-band system's was 4.3%. The improvement was not statistically significant.

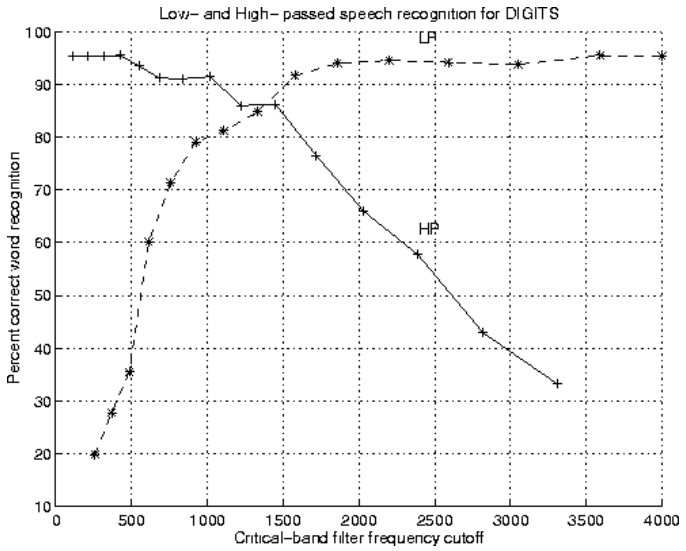

Figure 16: Word recognition percent correct for low- and high-pass speech of DIGITS corpus.

Hermansky and Tibrewala [25, 65, 67, 66] tested 2-band and 7-band multi-band systems on clean and noisy speech on the DIGITS database. They experimented

[9]This experiment was suggested by Jont Allen.

with various forms of corrupted speech: additive 900Hz sinusoidal with SNRs ranging from 30dB to 0dB, and added noise samples from the NOISEX-92 database (factory, destroyer-engine, pink, white, Volvo, babble, and high frequency radio channel). A full HMM/MLP recognition system was trained on PLP features extracted from each sub-band, and the merging was done on the word level using an MLP trained on the log likelihood distances. They trained 127 MLPs (that is: $\binom{7}{1} + \binom{7}{2} + \cdots + \binom{7}{7}$) for every possible configuration of the 7 bands. An merging MLP configuration was chosen based on various techniques, such as SNR thresholding (i.e., leaving out the sub-bands which yielded SNR estimates [26] below a certain threshold), majority vote, and adaptation. For noise conditions with local frequency degradations (e.g., additive sinusoid, pink, and babble) the performance of the multi-band system was better than the full-band system. In addition, for mild noise cases, choosing a merging configuration was not necessary, as the $\binom{7}{7}$-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 multi-band approach (as also observed by [7]). One does not need to running 127 (or a similar number of MLP merging units) to reap the noise robustness benefits, as for a large task the needed computational power would make this approach infeasible. Another notable result of [67] is that merging on the frame level is as good as on the word level, again, making their approach scalable to tasks with larger vocabulary sizes.

Tibrewala and Hermansky also performed experiments on the Switchboard continuous speech corpus, demonstrating 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 training set, test sentences were 240 male utterances from the development set, and the merging network was trained on an independent set of two hours of male speech. Each of the seven bands roughly covered 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 760K parameters), whereas the baseline full-band system had 234 input, 500 hidden, and 56 output units (145K 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%[10], and for the seven-band multi-band system was 59.0%. The 2% absolute difference between the error rates is not statistically significant, and also, it may be attributed to the additional two hour training data used for training the merging system. However, it is encouraging that the multi-band system is as good as the full-band system for a large vocabulary continuous speech task, though in fairness, we note that this multi-band system has about five times as many parameters as the baseline full-band

[10]The error rate for these HMM/ANN systems is particularly high because only four hours of training data are used to shorten the experimental turn-around time. The error rate of the HTK system with the same amount of training data was similarly around 60%.

system.

In their experiments Bourlard and Dupont [8, 7] 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 [24] 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 HMM-state, phone, or syllable level. A multi-layer perceptron (MLP) was used as merging unit in the frame-level combination experiments, and in syllable- and word-level combination experiments, HMM decomposition [69] was applied 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 multi-band paradigm was either as good or 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 [67]). Dividing the streams into four narrow-bands seemed 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 [55] are also noteworthy, as they show applicability of 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 [65] (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 510K parameters). The full-band system to which they compare their results was made up of four MLPs, each with 342 inputs, 1000 hidden units, and 56 outputs (1,600K 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%.

What distinguishes the work in this TR from that of our collaborators, as cited above, is that we have focused our attention on developing merged classes, transition based narrow-band classes, as well as on the analysis of multi-band.

Tomlinson et. al. [68] devised a two-band system with asynchrony between a high- and a low-pass component of the speech spectrum through a variant of HMM decomposition. Their experiments were performed on an in-house speaker-dependent 500 word ARM (Airborne Reconnaissance Mission) task [57]. They divided the speech to two bands of [0-4kHz][11] and [4kHz-8kHz], using 25 and 4 cosine terms (for low pass (LP) and high pass (HP) bands, respectively) and one energy term each to de-

[11] Note that their lower band region alone [0-4kHz] is equal to our full frequency band, as our experiments are performed on band-passed telephone speech.

scribe the lower and upper bands. They allowed asynchronous merging of the steams 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, generalizing their approach to three bands, as well as removing the phone-boundary asynchrony limitation did not perform as well as the more basic system reported above. Again, it is encouraging that multi-band has been shown to perform as well or better than a full-band system, however, the generalizability of the approach in [68] is uncertain. Furthermore, it is unclear whether the performance gain was purely from asynchronous merging, as the results for two vs. three band systems were not consistent.

A related multi-band work is that of Sankur et. al. [59] with application to the classification of respiratory sound signals into 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 the multi-band system.

10.2 Psycho-acoustic Studies

The research into multi-band approaches is first and foremost motivated by the work of Harvey Fletcher, as summarized and re-reported by Jont Allen [19, 1]. The underlying hypothesis of their work is that human speech perception is based on narrow frequency channels. Fletcher performed human listening experiments using nonsense CVC (consonant-vowel-consonant) sets, and based on the data, he proposed a model for human speech processing with five layers. 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 $W(A) = 1 - (1 - S(A))j$.

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

1. There is no statistical model for the error in the full-band being equal to the product of the errors in the narrow bands, as this requires knowledge of the reliability of a band with respect to the others.

2. The work of Kryter and Lippmann [32, 36] point out the shortcomings of the AI theory in explaining human speech perception (see discussion of [36] below).

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

Richard Lippmann [36] reports on human perception experiments using low and high frequencies. The common belief has been that high-frequency speech energy above 4 kHz contains inadequate information for speech perception. In his paper, 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 [33]. 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, whereas, the untrained human subjects achieve high recognition accuracy on highly unnatural conditions.

Another interesting psycho-acoustic study that involves the division of the frequency band is Oded Ghitza's tiling experiments [21] 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 CVC's 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 1000Hz and 2500Hz). He then interchanged a particular tile from one word with the same tile from another word in the same pair. He argues 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, and for exploring transition-based classes [47], which similar to diphones, emphasize the transition regions.

Finally, we wish to briefly discuss the the seminal work of Miller and Nicely [41] on low frequency and high frequency masking. 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 is not

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

only possible, but surprisingly good, with limited narrow band information (see Table 17 for 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 acoustic power, leaving consonants inaudible, causing random confusions. Similarly, we hope that in a multi-band ASR system, the pattern of errors are different for each of the sub-band recognizers. 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* held up, 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 our proposal of designing multi-band recognizers that classify narrow-band merged classes.

Note that the Miller & Nicely results were obtained on a set of sixteen consonants, and perception of continuous narrow-band speech may be different. For continuous speech recognition, our informal listening experiments on TIMIT sentences 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 [63].

Finally, French & Steinberg [20] have also performed human speech recognition experiments with non-sense CVC's with high-pass and low-pass speech. The results, summarized in Figure 17, further re-confirm the ability of humans for narrow-band speech recognition, and suggests the redundancy of information in speech spectrum.

10.3 Work on ASR with Missing Features

The work of [14] and [39] show that speech recognition with incomplete features may be done using missing feature theory. The main idea is to reconstruct the missing

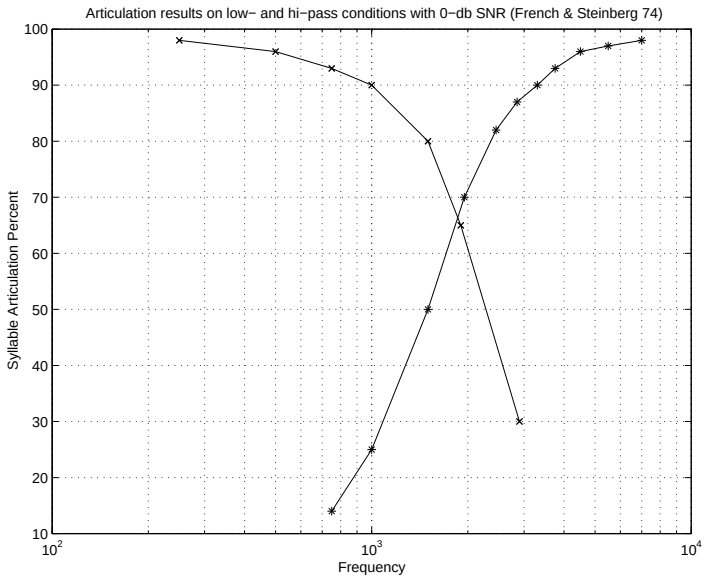

Figure 17: Articulation test results for human nonsense CVC recognition on low- and hi-pass filters with 0 SNR (from [French and Steinberg 1947]).

pieces of information using means and variances of the existing data. [14] has shown 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 according to local levels of SNR, the performance deteriorates gravely. Multi-band appears to be more robust in these situations [8, 66].

11 Appendix B – Confusion Matrices

Figures 18, 19, 20, 21, 22, and 23 show confusion matrices generated for sub-bands 1 through 4, the full-band, and multi-band, respectively. The data has been generated on NUMBERS95 development set. Detailed descriptions of the systems used are in Section 5.

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#
d	-	3	-	1	-	2	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
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	-	-	-	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	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	-	-	7	14	15	39	2	162	-	1	4
ax	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	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 18: 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	-	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	-	-	129
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	15815	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	91	242	324	56	-	70	127	5280	39	186	65	26	-	86	91
ao	-	-	-	6	-	-	-	-	-	-	-	1	-	-	-	-	1	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	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	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 19: 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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
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	-	-	-	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	-	-	-	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	-	1	-	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	-	44	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 20: 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	-	-	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	-	3	-	96	2	1	57	5	60	-	9	123	17	5	-	-	44	-	20	20	-	-	-	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	12269	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	-	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	-	17	9	146	-	5	202	20	39	-	-	545	16	205	52	-	-	-	450	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 21: 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	axr	ax	h#
t	-	31	-	5722	23	4	198	2	252	1	88	121	262	28	-	-	24	-	25	28	-	-	-	9	-	-	-	-	-	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	-	-	-	20	-	7	5	4	-	2	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	23	125	32	1	-	-	907
th	-	16	-	156	4	1	72	1	101	-	49	42	2111	21	-	-	8	-	112	-	-	12	-	4	30	5	1	-	-	-	-	2	9	-	24	-	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	4	-	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	-	-	354	5	383	59	-	-	1	9253	186	26	713	-	4	97	53	8	38	114	11	-	10	245
ih	-	1	-	3	-	3	5	75	126	-	67	35	37	10	1	-	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	-	-	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	-	2	208
ah	-	8	-	13	6	-	3	-	19	-	2	14	2	72	-	2	241	-	220	699	-	-	-	24	214	377	22	-	96	157	6446	41	161	11	19	-	110	89
ao	-	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	5	-	-	-	-	-	-	-	-	-	1	-	6	36	-	-	-	-	4
ow	2	1	-	59	3	-	36	10	118	-	26	115	30	87	-	-	141	186	709	225	-	5	-	58	54	210	70	-	27	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	-	-	35	62	4	228	7378	15	-	3	226
er	-	-	-	14	-	2	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 22: Confusion matrix for full-band.

Confusion matrix for multi-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	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	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	-	-	-	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	62	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	-	-	-	1	-	-	-	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	-	-	-	1	-	-	-	-	10	-	2	2	-	-	-	-	6	-	-	-	-	36	8	-	-	-	-	-	-	-	3	-	9	-	-	-	-	1
hv	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-	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	117	-	2	22	146	1	34	100	95	-	15	38
eh	-	-	-	11	-	-	-	-	74	1	2	17	1	92	-	-	19	1	13	16	-	-	-	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	1	17	1	24	-	24	25	-	3	-	-	150	5	160	53	-	-	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	-	-	-	-	-	-	227	66	26	65	-	-	-	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 23: Confusion matrix for multi-band.

12 Appendix C – Listings of Merged Classes

In this section, we have included the list of merged classes in each band. The details of the collapsing process can be found in Section 7.

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#. Below are the list of the merged classes for each distance criterion in each band:

- Relative Entropy:
  - Band 0:
    - (kcl er)
    - (t ax)
    - (n aa)
    - (hv ah)
    - (k aw)
    - (s hh l)
    - (dcl w)
    - (v ay)
    - (th h#)
  - Band 1
    - (f ay)
    - (n h#)
    - (eh uw)
    - (v ey)
    - (ih ow)
    - (kcl iy)
    - (k th)
    - (ao ax)
    - (t hh)
    - (s aw)
    - (tcl hv)
    - (aa ah)

- Band 2
  - (v ow)
  - (tcl ay)
  - (kcl f)
  - (th n)
  - (ao ax)
  - (dcl uw hv hh s)
  - (l aa k)
  - (t aw)
  - (z iy)
- Band 3
  - (kcl uw)
  - (dcl ay)
  - (iy aa)
  - (l ow ax)
  - (hh er)
  - (tcl hv)
  - (t aw)
  - (k ax)
  - (z eh)
  - (f v)

- Euclidean
  - Band 0
    - (f h# th k dcl)
    - (w ay hv)
    - (ah ax aa ey)
    - (r hh)
    - (kcl ao)
    - (z l)
    - (t aw)
    - (v er)
    - (iy uw)
    - (s ow)
  - Band 1
    - (eh uw)
    - (ay h# ih kcl)
    - (dcl n hv)
    - (t f hh ey)
    - (ah ax)
    - (tcl aa l er k aw)
    - (v ao z)

- Band 2
  - (n h# kcl)
  - (w ay)
  - (z ih)
  - (k v ax aa er)
  - (dcl th l hh ao)
  - (t hv uw)
  - (tcl aw eh)
  - (ah ow)
- Band 3
  - (t ah)
  - (f v uw)
  - (r ow)
  - (iy ax ao ey w z)
  - (k kcl aa l hh ih)
  - (dcl s aw)
  - (tcl hv)
  - (th eh)
  - (er h#)

- Manhattan
  - Band 0
    - (n w aa)
    - (th h# ao)
    - (hv eh hh dcl)
    - (er ax)
    - (kcl ih)
    - (k ey)
    - (t aw)
    - (tcl l)
  - Band 1
    - (n h#)
    - (eh uw er)
    - (iy ao)
    - (k tcl ay)
    - (w ax)
    - (hv ih)
    - (t hh)
    - (dcl aw ah)
    - (l aa)

- Band 2
  - (s iy)
  - (tcl h# ow hv)
  - (kcl f)
  - (z ey k)
  - (l r)
  - (ay ax aa)
  - (t aw)
- Band 3
  - (n h# ay ah t ih iy s r f ey v uw w th kcl eh ao)
  - (dcl ow)
  - (hh aa)
  - (hv er)
  - (k aw)

References

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

[3] 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.

[4] 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*, Seattle, WA, May 1998. To Appear.

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

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

[7] 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*, Philadelphia, PA, USA, October 1996.

[8] 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.

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

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

[11] 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.

[12] Numbers corpus, release 1.0, 1995.

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

[14] 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.

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

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

[17] 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.

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

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

[20] 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.

[21] 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.

[22] Steven Greenberg. Personal Communications, December 1996.

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

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

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

[26] 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.

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

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

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

[30] Fred Jelinek. Personal Communications, Johns Hopkins Summer Workshop on ASR, August 1996.

[31] 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.

[32] 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.

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

[34] 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.

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

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

[37] 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.

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

[39] 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.

[40] Alvin Martin, Jon Fiscus, Bill Fisher, Dave Pallett, and Mark Przybocki. 1997 LVCSR/HUB-5E workshop: System descriptions & performance summary. In *Proceedings of the Conversational Speech Recognition Workshop on DARPA Hub-5E Evaluation*, Baltimore, MD, May 1997.

[41] 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.

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

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

[49] 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*, Seattle, WA, May 1998. To Appear.

[50] 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.

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

[52] 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.

[53] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*, chapter 7.3, pages 395–400. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] Sangita Tibrewala. Personal Communication, October 1996.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

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

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

[72] 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*, 1998. To appear.

[73] 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.

[74] 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.