Multi-Stream Approach To Robust Speech Recognition

Sangita R. Sharma
B.E., University of Bombay, 1992
M.S., Oregon Graduate Institute of Science and Technology, 1997

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science and Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Electrical and Computer Engineering

October 1999

The dissertation "Multi-Stream Approach To Robust Speech Recognition" by Sangita Sharma has been examined and approved by the following Examination Committee:

Hynek Hermansky
Professor
Thesis Research Adviser

Jont Allen
Technology Leader
AT&T Labs.

Bishnu Atal
Technology Director
AT&T Labs.

Michael Macon
Assistant Professor

Misha Pavel
Professor

Dedication

To Ravi.

Acknowledgments

First and foremost, I wish to express my profound gratitude to my advisor, Prof. Hynek Hermansky for giving me the opportunity to pursue doctorate studies under his guidance. The work in this dissertation would not have been possible without his support, valuable ideas and encouragement. In fact he is the co-author of this work.

I would like to thank Prof. Misha Pavel for his involvement and technical discussions during the earlier part of this work. I would also like to thank all my committee members, Dr. Jont Allen, Dr. Bishnu Atal, Dr. Michael Macon and Dr. Misha Pavel for reviewing this work and offering insightful comments. I would like to acknowledge Dr. Nelson Morgan and Dr. Nikki Mirghafori at the International Computer Science Institute (ICSI), Berkeley, and Dr. Herve Bourlard, Stephane Dupont and Christope Ris from the Faculteé Polytechnic de Mons, Belgium for our multi-band collaboration. ICSI also provided the special purpose hardware and software for running the recognition experiments reported in this dissertation. I would also like to thank Prof B. Yegnanarayana for his valuable inputs during the summers he spent at OGI.

I would like to thank the past and present members of my group at OGI — Carlos, Sarel, Takayuki, Nobo, Naren, Sachin, and Pratibha for their enthusiastic contribution to the stimulating discussions on this work. Pavel Chytil and Amy Todd deserve a special mention for tolerating and accomodating my numerous demands on computational resources. Special thanks to my friends Priya, Neena, Radhika, Javed, Shafqat, Bala, Anurag, Sai, Suchanda, Wei and Xin for making my stay at OGI memorable.

My infinite gratitude to my husband, Ravi, for being my inspiration, and giving me all the love and support necessary to reach my goal. I would also like to thank my family for their love, patience and encouragement.

Finally, I would like to thank the Department of Defense, the National Science Foundation, Intel and Texas Instruments for providing the financial support for the work presented in this dissertation.

Contents

Dedication iii

Acknowledgments iv

Abstract xv

1 Introduction 1
- 1.1 Applications of Automatic Speech Recognition Systems 2
- 1.2 Challenges Facing ASR 3
- 1.3 Generalized Multi-stream Speech Recognition Model 4
- 1.4 Conventional (Single-Stream) ASR Model 8
  - 1.4.1 Drawbacks of the conventional ASR approach 8
- 1.5 Our Approach 9
  - 1.5.1 Single-input multi-stream model 11
  - 1.5.2 Issues in the multi-stream model 12
- 1.6 Dissertation Contribution 13
- 1.7 Organization of the Dissertation 15

2 Overview of the Hybrid HMM/MLP ASR system 16
- 2.1 Stages in an ASR System 16
  - 2.1.1 Feature extraction 16
  - 2.1.2 Probability estimation 17
  - 2.1.3 Decoding 20
- 2.2 Databases 21
- 2.3 Summary 23

3 Multi-Band Model 24
- 3.1 Conventional System 25
  - 3.1.1 Conventional system in narrow-band noise 25
- 3.2 The Human Way — Evidence on How Humans Process and Recognize Speech 26
  - 3.2.1 Fletcher's study on Articulation Index 26

3.2.2 Additional studies of human performance on filtered speech 28
3.3 Other Research on Recognition of Partially Corrupted Speech 30
3.4 Our Approach — the Sub-band Model for ASR 31
3.4.1 Previous work using the multi-band model 33
3.5 Experimental Setup 34
3.5.1 Baseline system 36
3.6 Feasibility Study 37
3.7 Issues in the Design of the Sub-band Model 38
3.7.1 Definition of frequency sub-bands 38
3.7.2 Features 39
3.7.3 Merging techniques 41
3.7.4 Merging level 46
3.8 Extension of the Multi-band Model to a Large-Vocabulary Continuous Speech Task 51
3.9 Summary 52

4 Multi-band in Noise 54
4.1 Sub-band Model on Band-limited Speech 55
4.1.1 Comparison of merging classifiers trained at different merging levels 57
4.2 Techniques for the Selection of Merging Classifiers 58
4.2.1 Signal-based technique — SNR estimation 58
4.2.2 Classifier-based techniques 59
4.3 Controlled Experiment with Frequency-localized Sinusoidal Additive Noise . 61
4.4 Experiments with Real Additive Noise 65
4.4.1 Selection strategies on real noise conditions 69
4.5 Summary 71

5 Analysis of Temporal Information — TempoRAl Patterns (TRAPS) . 73
5.1 Introduction 74
5.1.1 Spectral processing 74
5.1.2 Towards temporal processing 75
5.2 Our Approach to Analyzing the Nature of Temporal Information 77
5.2.1 Temporal evolution of phonemes — Mean TRAPs 80
5.2.2 Variation in temporal patterns across frequency 80
5.2.3 Variance of the temporal patterns 80
5.3 Clustering of Temporal Patterns 81
5.4 Summary 84

6 TRAP Classifiers in ASR **85**
6.1 Our Approach 86
6.2 Experimental Setup 87
6.3 Phonetic Classification using Mean TRAPs 88
6.3.1 Similarity scores obtained from the TRAP classifiers 90
6.3.2 Combination of the Mean TRAP classifiers 94
6.3.3 Combination of the Baseline and TRAP-based systems 96
6.4 Neural TRAPs 97
6.4.1 Robust Neural TRAPs 98
6.4.2 Combination of Neural TRAPs 99
6.4.3 Combination of the Baseline and Neural TRAP-based system 100
6.5 Broad TRAPs 101
6.5.1 Classification using Broad TRAPs 102
6.6 Performance of the TRAP-based System in Noise 103
6.6.1 TRAPs in convolutive noise 103
6.6.2 TRAPs in additive sinusoidal noise 104
6.6.3 TRAPs in realistic additive noise 104
6.7 Summary 105

7 Conclusions and Future Work **107**
7.1 Conclusions 108
7.2 Suggestions for Future Work 109
7.3 Final Remarks 111

Bibliography **112**

A TIMIT Phoneme Set **120**

B Mean TRAPs **121**

Biographical Note **134**

List of Tables

3.1 Cut-off frequencies of the 1-Bark spacing critical band filters used in PLP feature representation for 8 kHz sampling frequency. 35

3.2 Cut-off frequencies of the 0.5-Bark spacing critical band filters used in PLP feature representation for 8kHz sampling frequency. 36

3.3 Comparison of the word-level performance of the two sub-band model with that of the baseline system. The critical band definitions are obtained from Table 3.1. The MLP probability estimator has 200 hidden units and 61 output units. 37

3.4 Word-level performance of the sub-bands in the 2, 4 and 7 sub-band models using PLP critical band energies as features. The cut-off frequencies are as defined by the 15, 1-Bark spaced critical bands, given in Table 3.1. The MLP probability estimator has 200 hidden units and 61 output units. . . . 39

3.5 Word-level performance of the sub-bands in the 4 and 7 sub-band models using cepstral based features. The cut-off frequencies are as defined by the 30, 0.5 Bark spaced critical bands (Table 3.2). The MLP probability estimator in each sub-band has 200 hidden units and 61 output units. . . . 41

3.6 Comparison of the word errors (%) of the Baseline system and the sub-band systems obtained using different merging techniques. 43

3.7 Word accuracies in the individual sub-bands of the 4 sub-band system using cepstral based features. 45

3.8 Word error (%) obtained by merging at the frame-level and the word-level. 49

3.9 Comparison of the frame/phone errors (%) of the baseline system, the sub-bands of the 4 and 7 sub-band systems, the sub-band systems obtained by merging at the frame-level, and Fletcher and Stewart's multichannel model of phone perception given by Eq.3.8. 49

3.10 Word error % on the SWITCHBOARD database. 51

4.1 Different combinations of sub-bands in the 7 sub-band model. 55

6.1 Frame-level performance of the Mean TRAP classifier in each critical band. 89

6.2 Frame errors of the Mean TRAP classifiers as a function of the size of set k, where k corresponds to the set of classes having the top k similarity scores for a given frame. 89
6.3 Performance of the Baseline system and the Mean TRAP-based combined system on the OGI Numbers corpus. 95
6.4 Analysis of the frame-level decisions of the Baseline and the Mean TRAP-based combined system on the cross-validation set. 96
6.5 Performance of the Baseline system, the Mean TRAP-based combined system and a combination of these two systems. 97
6.6 Frame-level performance of different TRAPs on OGI Numbers corpus. Robust Neural TRAPs refers to the case when the input time trajectories are mean subtracted, variance normalized and windowed. 98
6.7 Performance of the combined TRAPs on the OGI Numbers corpus. 100
6.8 Performance of the Baseline system, and combination of the baseline system with the TRAP-based systems. 101
6.9 Performance with the Broad TRAPs. 103
6.10 Comparison of the degradation in performance from clean test condition to condition corrupted by convolutive distortion. 104
6.11 Frame and Word errors (%). 104
6.12 Frame and word errors (%) for the additive noises from the NOISEX-92 database. 105

A.1 61 TIMIT phonemes used in the multi-band experiments. 120

List of Figures

1.1 Generalized multi-stream speech recognition model. 4
1.2 Conventional ASR model. 8
1.3 Single input multi-stream concept. 11
1.4 Multi-stream ASR model. 11

2.1 Block diagram of the hybrid HMM/MLP ASR system. 17

3.1 Conventional ASR. 25
3.2 Sub-band Model. 31
3.3 Block diagram for merging of a 2 sub-band four class system. 42
3.4 Block diagram showing merging of the sub-band outputs of a 2 sub-band system at the word-level in an isolated word task. This system is composed of four phonetic classes and two words. 47
3.5 Phonetic alignments in the individual sub-bands of a 4 sub-band system as compared to that of the baseline system for an utterance of the word *seven*. 47
3.6 Block diagram of merging the sub-bands of a 2 sub-band system at the frame/state-level. This system is composed of four phonetic classes. 48
3.7 Block diagram of merging of the sub-bands of a 2 sub-band system at the phone-level. This system is composed of four phonetic classes and 2 words. 50

4.1 Word error (%) with merging networks having 0 to 6 sub-bands excluded. Merging is done at the frame level. 56
4.2 Word error (%) with merging networks having 0 - 6 sub-bands excluded and trained for merging at the frame-level and word-level respectively. The left-hand side graph is the same as Fig. 4.1 and has been repeated for easy comparison. 57
4.3 Distributions of the outputs (log-likelihoods) of the merging networks for the right decisions (top panel) and wrong decisions (bottom panel). 60
4.4 Distribution of the confidence estimates of the merging classifiers based on the d-prime measure and difference of top 2 outputs (log-likelihoods) for the right and wrong decisions. 62

4.5 Sub-band errors (%) on speech corrupted with additive sinusoidal noise at 1500 Hz. at different SNRs. 63
4.6 Performance of the sub-band system using different selection techniques, on clean speech and speech corrupted with additive sinusoidal noise at 1500Hz and SNRs 30dB, 20dB, 10dB and 0dB. For each SNR level the bars from left to right are in the same order as referred to in the legend. 64
4.7 Average log critical band noise spectra for the six noise samples from the NOISEX-92 database and the noise spectrum of a locally recorded Volvo noise (closed-volvo). Critical band frequency ranges are as given in Table 3.2. 66
4.8 Sub-band word errors (%) for each of the noise conditions. The dashed line corresponds to the sub-band errors for the clean speech and the solid line represents the sub-band errors for speech corrupted by the corresponding noise. The bar plot next to each sub-band error plot compares the word error of the conventional system (leftmost bar), the 7 sub-band system which uses frame-level merging of the sub-band outputs (middle bar) and that which uses word-level merging (rightmost bar). 67
4.9 Average word error rates (%) for speech corrupted with babble, factory, pink, volvo noises from NOISEX-92 database and Volvo noise recorded locally. Fig.(I) shows the case where the conventional system performance degraded to around 25% error. Fig.(II) shows the case where the conventional system degraded to around 50% error. 69
4.10 Average word error rates (%) for speech corrupted with white and high-frequency radio channel noise. 70

5.1 Single input multi-stream concept. 74
5.2 Computation of the temporal evolution of phoneme */ah/* for critical band f_i from a labeled database. 78
5.3 Mean TRAPs for 45 phonemes at 500 Hz (i.e., the fifth critical band). The dotted line for each of the TRAPs represents the center frame (i.e., at time $t = 0$ ms). The patterns separated by the thin solid lines represent sounds with similar temporal patterns. 79
5.4 Temporal evolution patterns of four phonemes for all the 15 critical bands. Each temporal evolution pattern is normalized to zero mean for convenient display. 81
5.5 Variance of the 1 sec temporal trajectories of the 45 phoneme classes for the fifth critical band. 82

5.6 Five Broad TRAP clusters of the fifth critical band TRAPs. The thinner lines in each plot represent the individual TRAPs of the phonemes clustered in one category. The thicker line is the Broad TRAP and represents the weighted mean of the constituent phoneme TRAPs. 83

6.1 Temporal Paradigm for ASR. 87
6.2 Template matching approach to phonetic classification in a frequency band, f_i, using the Mean TRAPs. 88
6.3 Average similarity scores for examples of /iy/ with (a), (b), (c) and (d). . . 90
6.4 Average similarity scores for examples of /ah/ with (a), (b), (c) and (d). . . 91
6.5 Average similarity scores obtained for the 2 classes, /f/ and /t/, from the Mean TRAP classifiers. 92
6.6 Mean TRAP combiner block diagram. Each TRAP classifier has the form shown in Fig. 6.2 up to the decision-making block. 95
6.7 Block diagram of combination of baseline (spectral-based) system with TRAP-based (temporal) system. 97
6.8 Neural TRAP. 98
6.9 Effect of mean subtraction, variance normalization and windowing on a 1 sec spectral time trajectory of clean speech and speech corrupted with additive white noise. Spectral trajectory shown is at 500 Hz. 99
6.10 Average similarity scores for each of the phonemes /iy/, /ah/, /f/ and /t/. The regions bounded by dotted lines represent the 15 frequency bands within each of the 5 Broad TRAP categories, as labeled on the x-axis. . . . 102

B.1 Comparison of the average similarity scores obtained for the 2 classes, /d/ and /k/, from the Mean TRAP classifiers. 122
B.2 Comparison of the average similarity scores obtained for the 2 classes, /dcl/ and /tcl/, from the Mean TRAP classifiers. 123
B.3 Comparison of the average similarity scores obtained for the 2 classes, /kcl/ and /s/, from the Mean TRAP classifiers. 124
B.4 Comparison of the average similarity scores obtained for the 2 classes, /z/ and /th/, from the Mean TRAP classifiers. 125
B.5 Comparison of the average similarity scores obtained for the 2 classes, /v/ and /n/, from the Mean TRAP classifiers. 126
B.6 Comparison of the average similarity scores obtained for the 2 classes, /l/ and /r/, from the Mean TRAP classifiers. 127
B.7 Comparison of the average similarity scores obtained for the 2 classes, /w/ and /ih/, from the Mean TRAP classifiers. 128

B.8 Comparison of the average similarity scores obtained for the 2 classes, /eh/ and /ey/, from the Mean TRAP classifiers. 129
B.9 Comparison of the average similarity scores obtained for the 2 classes, /ay/ and /ao/, from the Mean TRAP classifiers. 130
B.10 Comparison of the average similarity scores obtained for the 2 classes, /ow/ and /uw/, from the Mean TRAP classifiers. 131
B.11 Comparison of the average similarity scores obtained for the 2 classes, /er/ and /ax/, from the Mean TRAP classifiers. 132
B.12 Comparison of the average similarity scores obtained for the class /sil/ from the Mean TRAP classifiers. 133

Abstract

Multi-Stream Approach To Robust Speech Recognition

Sangita R. Sharma

Supervising Professor: Hynek Hermansky

The recognition accuracy of current automatic speech recognition (ASR) systems deteriorates in the presence of signal distortions caused by the background noise and the transmission channel. Improvement in the recognition accuracy in such environments is usually obtained by re-training the systems or adaptation with data from the new testing environment. In this dissertation, we present a new *multi-stream* approach to improve the robustness of ASR systems, without the need for re-training.

In our approach, simultaneous estimates of the phoneme probability are made in parallel in multiple sub-streams. Each sub-stream operates on different features obtained from the same input. These partial phone probability estimates are then combined for the final recognition. This approach potentially allows for de-emphasizing unreliable sub-streams during recognition. We present two multi-stream models based on the features used in each sub-stream.

In the first model, the *multi-band model*, different frequency sub-bands of the speech spectrum are processed independently in different sub-streams. This is in contrast to the conventional spectral processing approach in which the entire spectrum is processed as a single stream. The sub-band phone probability estimates are then non-linearly combined (using a neural network) for final decoding. This model is motivated by the perceptual

studies of phone recognition in humans (Fletcher, 1953) and helps in localizing the effect of narrow-band noise. The multi-band model which is trained only on the clean data yields around 50% reduction in word error rate on isolated digits corrupted by realistic, frequency-selective, additive noise.

The second model is an extension of the multi-band model. In addition to independent frequency processing, it uses medium-term (around 200-500ms) temporal information in each frequency sub-band for phone probability estimation. The use of such syllable-length time spans is supported by psychoacoustic studies (Hermansky, 1998). The features (1 sec long time trajectories of critical band energies) that we use in each sub-band characterize the patterns of temporal evolution of the phonemes in that sub-band. This system yields errors which are complementary to that of the conventional spectral-based system. This system (trained on clean data) when combined with a spectral-based system yields around 25% reduction in word error rate on a continuous numbers task when tested under both clean and noisy conditions.

To summarize, the results obtained in this dissertation indicate that estimating phoneme probabilities from medium-term temporal-based features, independently from multiple frequency sub-bands, followed by non-linear combination of these probability estimates, improves the noise robustness of ASR systems.

Chapter 1

Introduction

Speech is the most natural means of communication in human-to-human interactions. The goal of automatic speech recognition (ASR) systems is to make speech equally viable in human-to-machine interactions. In recent years, ASR systems have found their way in several commercial applications such as in telephony, desktop dictation systems, etc. Even so, ASR technology has still a long way to go before reaching a stage of being the natural means of human interaction with machine (Lippmann, 1997).

The biggest drawback of ASR systems is their limited ability to deal with the diversity and variability of real-world environments. For example, recognition errors of ASR systems increase rapidly in the presence of noise. Humans, on the other hand, can recognize speech fairly well under normally occurring environmental conditions. This suggests that the human speech recognition mechanism is quite different from that used in ASR systems and that ASR can gain by an understanding of what makes the human recognition system robust.

The objective of this dissertation is to improve the noise robustness of ASR systems. To achieve this objective we propose to incorporate two concepts into the ASR system - 1) partial recognition in frequency sub-bands and 2) temporal processing over medium time (about syllable-length) spans. Psychoacoustic evidence suggests that the human auditory system is capable of such processing. We hypothesize that these phenomena account to some extent for the robustness of the human speech perception system to adverse environments and should also improve the noise robustness of ASR systems. To incorporate these two concepts into the ASR system, we propose and explore the *multi-stream* approach to speech recognition.

In this chapter, Section 1.1 provides a brief review of the commercial applications of ASR. Section 1.2 discusses some of the drawbacks of ASR as compared to human speech recognition and suggests an approach towards addressing these drawbacks. Section 1.3 presents our generalized multi-stream framework and reviews its application in the field of pattern recognition. In Section 1.4, we briefly discuss the conventional state-of-the art ASR system. Section 1.5 motivates and presents a multi-stream concept which forms the basis of the work in this dissertation. In Section 1.6, we specify the scope of the thesis and outline the main contributions. We conclude this chapter in Section 1.7 with the organization of the remaining chapters.

1.1 Applications of Automatic Speech Recognition Systems

Speech recognition technology is being used in many commercial applications. Some of these applications, as summarized by Rabiner (Rabiner, 1997), include

- Command-and-control applications where the user gives a single word voice command from a small vocabulary of word commands and the system recognizes and executes the desired command.
- Specialized telephony-based applications such as voice dialing which are especially convenient in hands-free mobile environments, directory assistance, and credit card authorization.
- Desktop applications such as dictation systems, voice browser for the internet and voice navigation of desktop computers.
- Business applications such as data entry, airline reservation systems, etc.

Although these applications have achieved moderate success, their widespread acceptance is limited due to their reduced performance in real world acoustic environments.

1.2 Challenges Facing ASR

Recognition in adverse environments

Most speech recognition systems rely on high quality speech input, usually obtained from close-talking microphones under quiet laboratory-like conditions, for good performance. However, most real world environments are characterized by the presence of interference caused by noise, for e.g., fans or computer-generated noise in an office environment, machinery noise on a factory floor, engine noise in mobile telephony environments (cellular phones or car phones) and background speakers. These are examples of noise that get added to the speech signal. Besides these additive noise conditions, there also exists speech signal distortion due to different acquisition channels such as different microphones and telephone lines. In general, recognition errors of ASR systems increase rapidly in the presence of such noise since there is a mismatch in the training and the testing conditions under which speech is recorded. The recognition of words in an utterance typically reduces in accuracy by 20% to 50% due to environmental noise (Gong, 1995).

Successful commercial applications often require additional tuning, data collection and analysis after field deployment, to adjust the system parameters for differences between laboratory test conditions and actual usage conditions (Thomson, 1997). Such re-training of the system for every new condition is both time-consuming and expensive and should be avoided, if possible.

Comparison of speech recognition performance of humans and machines

The human speech recognition system demonstrates remarkable robustness to the common speech signal degradations mentioned above. Lippmann (Lippmann, 1997) has reported comparisons of the performance of humans and machines on speech recognition tasks ranging in vocabulary size from 10 words to more than 65,000 words. The comparisons show that even under quiet conditions, machine error rates are an order of magnitude higher than those of humans. This difference in performance becomes larger in the presence of noise.

Bridging the gap between human and machine speech recognition

The gap in the performance of humans and machines suggests that the human speech recognition mechanism is quite different from that used in ASR. An approach towards addressing this issue is to understand the robust mechanism of the human auditory system and try to incorporate it into the ASR model. This approach focuses on understanding the psychoacoustic and neurophysiological aspects of the human speech communication process that makes the system resistant to adverse environmental changes.

Several auditory speech processing techniques that aim to explicitly mimic certain peripheral and central auditory processing phenomena observed in humans, have been proposed over the years. These include the ensemble interval histogram (EIH) (Ghitza, 1987), the Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980), the perceptual linear predictive (PLP) technique (Hermansky, 1990), phenomenon of lateral inhibition in the system (Shamma, 1985), among others (see (Hermansky, 1998) for review). There also exist techniques which explicitly emulate the temporal properties of human hearing (Seneff, 1985; Cohen, 1989; Hermansky and Morgan, 1994).

These techniques produce new acoustic feature representations of the speech signal and hence represent alternate speech signal processing techniques for feature extraction. This dissertation proposes to take the perceptual understanding into the acoustic modeling stage of ASR, using a multi-stream framework.

1.3 Generalized Multi-stream Speech Recognition Model

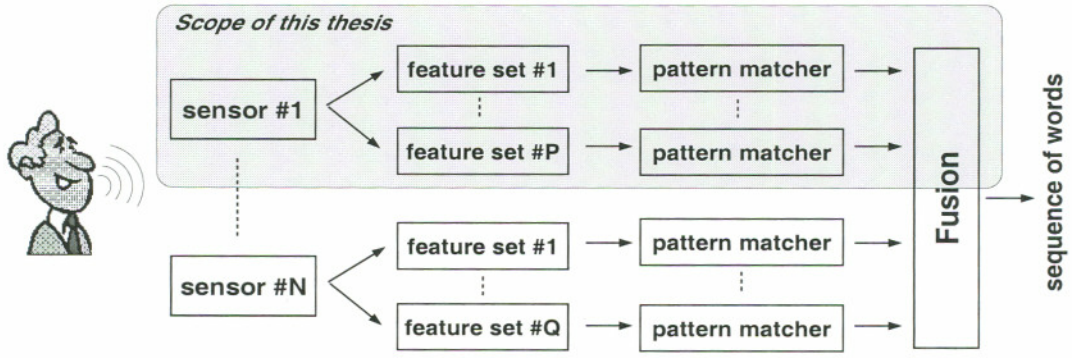


Figure 1.1: Generalized multi-stream speech recognition model.

Figure 1.1 represents a generalized block diagram of a multi-input, multi-stream speech recognition model. Speech input can be captured by a number of sensors such as telephone, microphone arrays, and even a video camera which may track the lip movements of the speakers. Each input signal, in turn, can be processed in a number of different ways, for e.g., by different feature extraction techniques, each of which focuses on a different aspect of the signal subspace. Hence, the extracted features will possibly contain some amount of complementary information. Pattern matching (i.e., classification) can then be performed on each feature stream, either obtained from the same input and/or from different inputs. The outputs from all these parallel streams can then be fused or merged to decode the spoken utterance. Thus, this generalized multi-stream model is based on combining the outputs of two or more classifiers that produce information independently. The output of one classifier does not affect the output of the other classifiers. Such a system is an example of a weakly coupled data fusion system (Clark and Yuille, 1990).

This multi-stream model for speech recognition shown in Fig. 1.1, can be considered as an engineering model of human speech perception. Firstly, it is well-known that human speech perception is at least bi-modal (Massaro and Cohen, 1983; Dodd and Campbell, 1987). In addition to using acoustic information, humans also use visual information, if available, in the form of positional information about the visible speech articulators (for e.g., lip movements, tongue and teeth positions). Secondly, even within a single modality, say for example the acoustic signal, there exist processes of partial recognition (Fletcher, 1953; Allen, 1994) (to be discussed further in Chapter 3), which support the notion of a multi-stream paradigm within a single-input framework.

The advantages of the multi-stream approach are:

- If some operating environment corrupts only a subset of the streams, it allows for the possibility to de-emphasize or drop those sub-streams in the merging process.
- It allows for independent and even different processing in each of the sub-streams.
- It allows the flexibility to incorporate additional streams without re-training the existing ones.

In this dissertation, we focus on a single speech input multi-stream framework (see shaded region of Fig. 1.1). This allows us to address the drawbacks related to the processing of single input speech signal before considering the possible extension to multiple inputs. The defined scope has many applications, especially in telephony, where only a single-input is available.

Multi-stream approach in pattern recognition

The multi-stream framework, also called the multi-classifier system in literature, is not unknown to the field of pattern recognition. For example, it has been used for unconstrained character/handwriting recognition (Xu et al., 1992; Ho, 1992; Lyon and Yaeger, 1996). The use of the multiple classifier systems in pattern recognition has been motivated by the existence of not only alternative feature representations but also alternative classification methodologies such as Markov models, neural networks, nearest neighbor classifiers, etc. Each methodology in turn leads to a solution that could complement the other, in terms of which patterns are correctly classified. Hence, proper integration could make it possible to yield better overall performance than any one technique used in isolation. Many a times the co-existence of multiple techniques also stems from incomplete knowledge about the true underlying model of the data. Different techniques or methodologies facilitate wider coverage and better modeling of the effective feature space in the absence of a solid assumption about a unique underlying model and hence help to complement each other (Shimshoni and Intrator, 1996).

Multiple neural network systems, also referred to in literature as *ensembles* or *committees of learning machines*, are another example of the use of the multi-stream framework for pattern classification and regression problems. In these systems, the input feature representation and classification methodology usually remain the same. The multiple neural networks, however, typically differ in initial conditions and/or architectures. The ensemble members are sometimes trained on the same data set (Hansen and Salamon, 1990) or on different subsets of the data (Perrone and Cooper, 1993; Wolpert, 1992). Ensemble neural networks have been used for several classification tasks such as seismic signal

classification (Shimshoni and Intrator, 1996) and classification of underwater acoustic signals (Ghosh et al., 1992). Mixture models such as adaptive mixture of experts (Jacob et al., 1991) and hierarchical mixtures of experts (Jordan and Jacobs, 1994) also use multiple neural networks. However, they are different from simple ensemble methods in that they are based on the "divide and conquer" approach. In this approach, a mixture of experts compete to gain responsibility in modeling the output in a given input region. The underlying model is based on the assumption of mutual exclusivity — i.e., a single expert is responsible for a particular data point. Boosting (Schapire, 1990; Drucker et al., 1993) is another example of a committee machine in which the experts are trained on data sets with entirely different distributions. The multi-stream framework has also been used in the field of image sensor fusion, to integrate information from various sensors for object detection and identification (Luo and Kay, 1988; Sharma et al., 1999).

Multi-stream approach in speech recognition

The multi-stream approach has been used for speech recognition in the form of *modular networks* (Haykin, 1998). Modular networks are also based on the principle of "divide and conquer", in which each module (subnetwork) operates on distinct inputs without communicating with each other. Modularity has been used in the design of large phonemic networks for recognition of all consonants (Waibel et al., 1989). In this approach, several time-delay neural networks are developed for different subsets of confusable consonants and the outputs of these subnetworks are combined to determine the consonant class. The goal was to provide scalability from smaller modular networks to a larger network without increasing computational complexity. Ensemble methods, such as boosting and mixture of experts, have also been used for phoneme classification (Waterhouse and Cook, 1997; Cook et al., 1997) and found to yield improvement in classification performance. The multi-stream approach has also been used in lipreading (speechreading) systems. The two streams representing audio and visual inputs respectively, are combined to improve speech recognition accuracy, especially in degraded acoustic conditions (Petajan, 1987; Stork et al., 1992; Silsbee and Bovik, 1996).

In this dissertation, we propose new sub-stream definitions in the multi-stream framework, to improve ASR robustness.

1.4 Conventional (Single-Stream) ASR Model

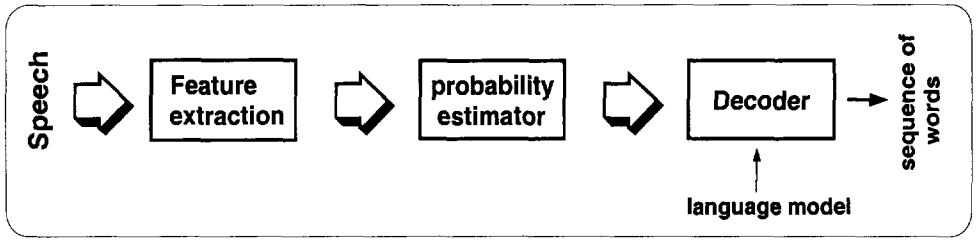


Figure 1.2: Conventional ASR model.

Figure 1.2 represents a simplified block diagram of the conventional ASR system. The feature extraction block divides the incoming signal into short-term frames (typically 10 to 20 ms) which are equally spaced in time, and extracts a feature vector from each frame. Each feature vector typically characterizes the spectral properties of the speech frame. The probability estimator block (for example, a hidden Markov model or a neural network) then estimates the log-likelihoods or posterior probabilities of the frames with respect to predefined speech sub-units (such as phones) being modeled. This probability estimation stage constitutes the acoustic modeling stage in ASR. The estimated probabilities are further processed by the language processing module (decoder), which uses language modeling information (i.e., a model of syntax and semantics) to generate the most likely sequence of words present in the incoming speech.

We call the model in Figure 1.2 a single-stream model because information from the speech signal flows along a single path through the system.

1.4.1 Drawbacks of the conventional ASR approach

There are at least two major limitations in the acoustic modeling stage of the conventional approach:

1. Since the single stream model employs only a single feature vector per speech frame, all the elements of the feature vector are treated as one entity by the probability

estimator. As a result, localized degradation of even a few elements results in a degraded feature vector, which often leads to misclassification. For example, even if only a few frequency regions get corrupted by narrow-band noise, the recognition performance degrades rapidly.

2. As mentioned above, features are extracted on a short-time basis (typically every 10ms) to account for the non-stationarity of the speech signal. The common assumption of the subsequent HMM modeling is that the feature vectors are independent across time. The HMM model attempts to capture the temporal aspects of the speech signal by assuming that speech can be modeled by a succession of states defined by locally stationary segments, with transitions between the states. Any correlations between feature vectors is overlooked (Bourlard and Morgan, 1994). However, the speech production phenomenon of coarticulation suggests that the feature vectors can be correlated over several phonemes, due to the influence of the surrounding phonemes on the production of the current phone. The effects of coarticulation often extend across syllable boundaries (O'Shaughnessy, 1987).

1.5 Our Approach

Independent frequency processing

Fletcher and Stewart (Fletcher, 1953) proposed the multi-channel model of phone perception, which shows that the probability of errors in sub-bands multiply to yield the overall probability of error[1]. Allen interpreted this model in terms of speech recognition to imply that humans could be processing phones in several independent frequency channels (Allen, 1994). The resulting partial phone estimates are then fused for final decoding of phones. The multiplication of the sub-band errors also implies that errors (probably due to noise) in one frequency channel will not affect the errors in the other frequency channels.

An interpretation of this is that as soon as some sub-band combination yields relatively reliable information, information from the other possibly corrupted sub-bands need not

[1]To be discussed in Section 3.2.1.

be used for recognition. The concept of independent processing of frequency sub-bands thus indicates that degradations due to narrowband noise could be localized and hence could improve robustness.

Across-time processing

The effect of a phoneme is known to last over several neighboring phones due to coarticulation. One way to account for this temporal spread of phonetic information is to use temporal information from time-spans extending across phone boundaries to decode the current phone.

Also, it is difficult to distinguish between stationary noise (such as those introduced by linear transmission channel) and speech by looking at just 20ms of the signal. Longer time spans are required to suppress such noise effects. In ASR, the success of dynamic cepstral features (Furui, 1981), cepstral mean subtraction and RASTA (Hermansky and Morgan, 1994) techniques, which process sequences of feature vectors covering 50ms to several hundred milliseconds, indicates that temporal processing improves noise robustness of ASR systems. Besides, use of temporal information from around 200ms is supported by the auditory phenomenon of temporal masking which states that the perception of one sound is diminished by the presence of a preceding sound (O'Shaughnessy, 1987; Moore, 1989). The sensitivity of human hearing to modulation frequencies around 4-6 Hz (Drullman et al., 1994; Arai et al., 1996), also suggests temporal processing of the order of 150-250ms. This ability of the auditory system to process temporal information could account to some extent for its robustness.

Hence, we hypothesize that the use of temporal information from around syllable-length (around 200ms) durations should not only yield considerable phonetic information, but also help in reducing noise effects.

Our Proposal

Combining the above two concepts of independent frequency processing and across-time processing, we propose a shift in the conventional ASR paradigm (Figure 1.3),

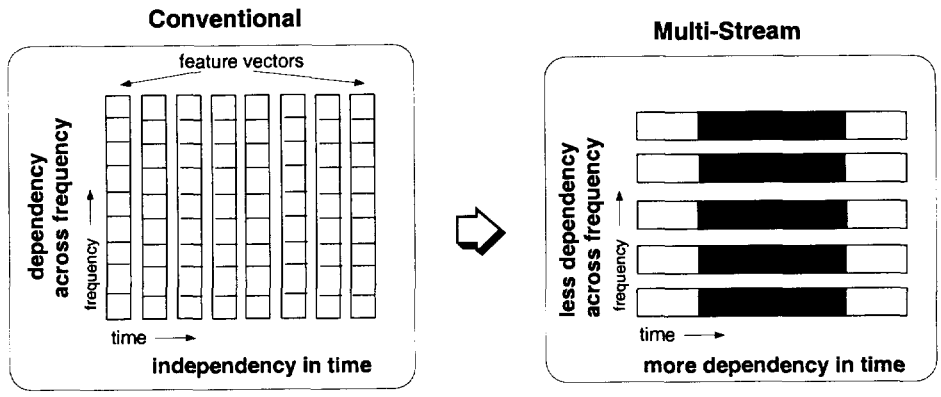


Figure 1.3: Single input multi-stream concept.

- from the assumption of dependence across the different frequency regions to the assumption of relative independence between them, and
- from the assumption of independence of the feature vectors across time to the assumption of syllable-length time dependencies between them.

We further hypothesize that a speech recognition model based on the above paradigm will be robust to environmental degradations because 1) errors in frequency can be localized and 2) temporal processing will help suppress noisy signal components.

In this dissertation, we present an approach to incorporating the above two assumptions in the acoustic modeling stage of speech recognition, using the following single-input multi-stream model.

1.5.1 Single-input multi-stream model

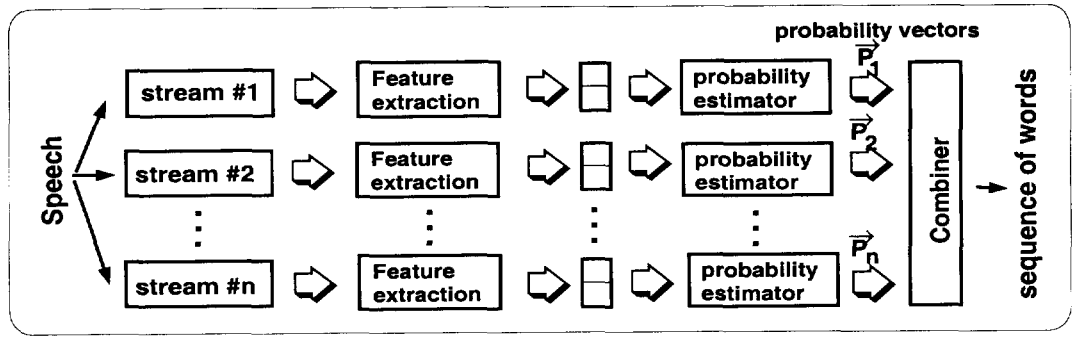


Figure 1.4: Multi-stream ASR model.

Figure 1.4 is the block diagram of the single-input multi-stream model for ASR. The idea is to derive different information sub-streams from the signal, each of which provides a certain view of the underlying linguistic process (speech information). Independent feature extraction and probability estimation is then carried out on each of the sub-streams. The output vectors of probability estimates (phoneme sub-unit probabilities) are then combined. This block diagram does not explicitly show the decoder because the combination can be done before the decoding stage or the decoder itself could perform the merging[2].

1.5.2 Issues in the multi-stream model

There are several issues involved in the implementation of the single-input multi-stream model shown in Fig. 1.4 :

1. **Definition of sub-streams**: The definition of the sub-streams is a critical issue in the design of the multi-stream model for ASR. The streams should be such that each of them provides, to some extent, independent and complementary information about the linguistic process. Another criteria for stream definition is that the sub-streams should cover the input space in such a way that there exist acoustic degradation conditions under which at least a few of them continue to model their input subspace with the same reliability as in clean (or undegraded) acoustic conditions.

2. **Features to be used in each sub-stream**: The issue of what features should be used in the sub-streams is similar to the equivalent issue in conventional systems to obtain the best features for speaker-independent speech recognition. Thus, the features in each sub-stream should robustly describe their input subspace with respect to speech variabilities due to speakers and noise.

3. **Technique to merge the streams**: The effectiveness of the multi-stream model depends on the technique used to merge the outputs obtained from the multiple

[2]To be discussed in Section 3.7.4.

classifiers (probability estimators) being used in the sub-streams. An important issue is whether the merging technique should be linear or non-linear and how should the contribution (i.e., weights) of the individual sub-streams be determined.

4. **Speech-unit level at which the streams should be merged**: As mentioned in Section 1.5.1, the streams can be merged at the outputs of the probability estimators, — i.e, the frame-level in case of neural network probability estimators or the state-level in case of hidden Markov model based probability estimators. The streams can also be merged at a higher sub-unit level such as the phoneme, syllable, word or utterance-level, in which case the merging has to be implemented as a part of the decoder.

1.6 Dissertation Contribution

This thesis focuses on the definition of sub-streams for use in the single-input multi-stream approach to robust speech recognition. The emphasis is on the definition of streams based on the concepts discussed in Section 1.5. The main contributions of the work described in this dissertation are the following:

- **The multi-stream model** : We define the framework of the multi-stream model for a single input ASR system. The motivation and the concept behind this framework are described in Sections 1.3 — 1.5. The resulting framework for a single-acoustic-input system is a significant departure from the single-stream conventional ASR paradigm, as discussed in Section 1.5.

- **The multi-band model** : We propose a sub-stream definition in which each sub-stream is defined by frequency sub-bands characterizing different regions of the speech spectrum. The features obtained in each frequency sub-band are independently processed by a probability estimator to generate phonetic probabilities. These sub-stream probabilities are then non-linearly merged for subsequent classification. The resulting model is called the multi-band model. We show that in matched conditions, merging of independent phoneme estimates from different frequency sub-bands

yields performance comparable to the conventional system, which processes the entire spectrum as one entity. We also discuss the issues in the design of this model.

- **Robustness of the multi-band model**: We demonstrate, on an isolated-digit task, the inherent robustness of the multi-band ASR model to frequency-localized degradations. One of the advantages of the multi-stream framework discussed in Section 1.3 is that it allows for de-emphasizing unreliable sub-streams before merging them. An important issue, therefore is, the determination of the sub-stream reliability. We present several techniques for determining the sub-band reliability. We show that an additional improvement in performance is obtained by using these techniques.

- **Data-driven analysis of temporal structure of speech** : We present a new approach to analyzing the nature of phonetic information present in the temporal structure of speech. The analysis, which is carried out on several hours of labeled speech, yields patterns (called TRAPS for TempoRAl Patterns) representing the temporal evolution of phonemes. The syllable-length ($\approx$ 200 ms) temporal spread of the patterns indicates that the effect of phonemes lasts for considerable duration in the temporal domain. This indicates that temporal processing of syllable-length time spans can yield additional phonetic information.

- **Incorporation of temporal information within each sub-stream** : We propose a model called the TempoRAl Pattern (TRAP) classifier model, which incorporates temporal information (of the order of 200 to 500ms) in each frequency sub-stream. The TRAP model is an extension of the multi-band model. The features used in each frequency sub-stream are the temporal trajectories of logarithmic critical band energies. We present a complete ASR system based on this concept of independent processing of frequency dependent temporal trajectories. We show, using a continuous numbers task, that this system is competitive with a conventional ASR system and yields considerable amount of complementary information. We demonstrate the robustness of the combined conventional and TRAP system in noise.

1.7 Organization of the Dissertation

Chapter 2 covers a short overview of the hybrid (hidden Markov model/multi-layer perceptron) speech recognition system developed at the International Computer Science Institute (ICSI), Berkeley, California. This system represents the conventional (single-stream) ASR system. We use this ASR system for all the experiments reported in this dissertation. Chapter 3 presents the multi-band model which independently processes different sub-bands of the speech spectrum. We discuss the issues concerned with the design of the multi-band model. We demonstrate the feasibility of the model in matched conditions. Chapter 4 demonstrates the effectiveness of the multi-band model in frequency-selective noise. We describe various techniques for the selection of reliable sub-bands in noise. In Chapter 5, we present a new approach to analyzing the temporal structure of speech. We examine the resulting patterns of temporal evolution of different phonemes. Chapter 6 presents an approach to incorporating temporal information in the ASR multi-band system. The proposed TempoRAl Pattern (TRAP) Classifier model incorporates medium-term temporal dependency in the multi-band model. We present a complete ASR system based on TRAPs and demonstrate its benefits in noisy conditions. Chapter 7 describes the conclusions from the work in this dissertation and discusses possible research directions for the future.

Chapter 2

Overview of the Hybrid HMM/MLP ASR system

The combination of hidden Markov models (HMMs) with neural networks is referred to as the hybrid approach to ASR. The speech recognition system used for all the experiments in this dissertation is such a hybrid hidden Markov model/multi-layer perceptron (HMM/MLP) ASR system developed at the International Computer Science Institute (ICSI), Berkeley, California. In this chapter, we give a brief description of this ASR system. A complete description of this system along with a discussion of the underlying theory can be found in (Bourlard and Morgan, 1994) and (Morgan and Bourlard, 1995). The description provided in this chapter is intended to give a quick overview of the system. We also describe the databases that we use for the experiments in this dissertation.

2.1 Stages in an ASR System

A typical ASR system consists of three main stages — 1) the feature extraction stage 2) the probability estimation stage, and 3) the decoding stage. Fig 2.1 represents the block diagram of a conventional ASR system. Specifically it represents a hybrid ASR system.

2.1.1 Feature extraction

The first step in any pattern recognition system is to extract meaningful features from the incoming signal. In speech recognition, the incoming acoustic waveform is segmented into overlapping frames at a typical frame rate of around 10ms to account for the non-stationary

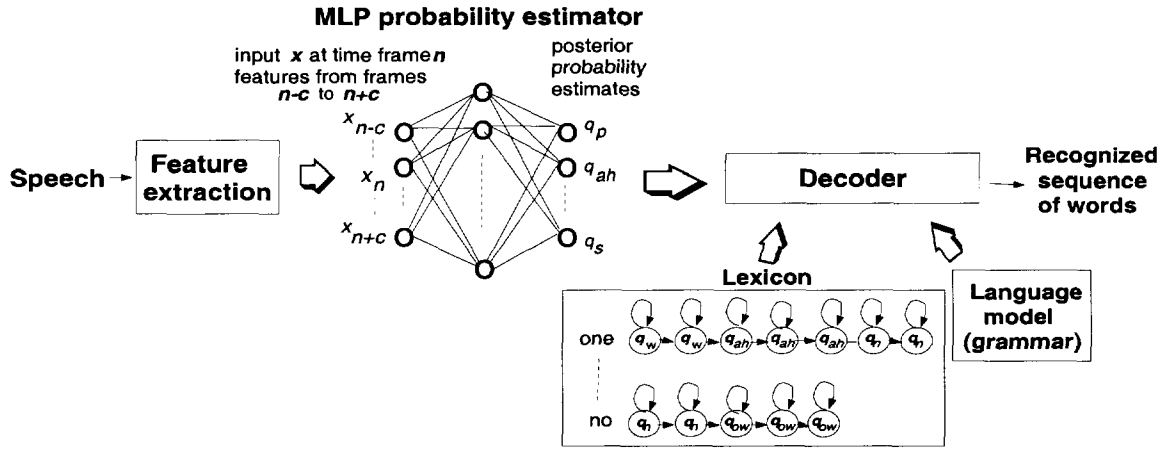


Figure 2.1: Block diagram of the hybrid HMM/MLP ASR system.

nature of the speech signal. Features are extracted from each frame using short-time discrete Fourier transform based signal-processing techniques. Typical features extracted include filter-bank energies, linear predictive coefficients (LPC), cepstral coefficients, Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) or the perceptual linear predictive (PLP) coefficients (Hermansky, 1990). The latter two features are based on auditory characteristics of the human hearing. These features represent the short-time spectral envelope of the speech frame. For the experiments described in this dissertation we use PLP based features.

2.1.2 Probability estimation

The next stage in a typical ASR system uses statistical techniques to characterize the properties of the extracted sequence of feature vectors. This stage is referred to as the acoustic modeling stage which estimates the class-conditional probabilities (called emission probabilities) of the extracted features.

The hybrid ASR system uses a neural network such as a multi-layer perceptron (MLP) to estimate the conditional probabilities, instead of the multivariate Gaussian mixtures-based probability estimation typically used in HMM systems. HMM-based algorithms are based on a well-developed mathematical framework for learning and representing the time-sequential nature and the variability of speech. However, several assumptions are

commonly made to make the optimization of HMMs computationally feasible. HMM training algorithms generally maximize the likelihoods (maximum likelihood estimation) instead of the posterior probabilities (maximum-a-posteriori estimation) — i.e., each class model is trained using data belonging to that class only and hence the training does not have access to examples from another class, necessary for discriminative training. As a result, the trained models can suffer from poor discrimination. HMM models require a priori choice of probability density functions such as the use of mixture of Gaussian densities. To reduce the number of free parameters, the components of the input feature vector are generally assumed to be uncorrelated — i.e., the covariance matrix is assumed to be diagonal.

The use of a MLP as a probability estimator attempts to alleviate some of the possible limitations of the HMM models, mentioned above. MLPs are trained using discriminant-based training algorithms that minimize the error rate for misclassification while maximizing the accurate modeling of the correct class. When trained as a classifier (i.e., with binary outputs), the outputs of the MLP estimate the corresponding class posterior probabilities (Richard and Lippmann, 1991) without requiring strong assumptions about the underlying probability density functions of the input features (as is required in the case of HMMs). MLP training does not require any assumptions of the independence of feature components. MLPs also allow for the use of contextual inputs, to explicitly account for any correlations between successive speech frames, which is not typically used in most HMM models.

Training

The MLP used in the ICSI hybrid system is a fully-connected feed-forward network with a single hidden layer. The number of inputs, hidden units and output units in the MLP will be mentioned along with the experiments in the following chapters. The MLP is trained using the on-line error-back-propagation algorithm with relative entropy error criterion (Haykin, 1998). It uses sigmoid non-linear activation function on the hidden-layer and softmax function on the outputs. The training uses the performance of the network on an independent data set, referred to as the cross-validation set, to determine

the stopping criterion for the training of the net as well as to determine the learning rate after each epoch of training. The use of the cross-validation set attempts to address the problem of overtraining of the network to maintain its generalization capabilities. The training starts at a learning rate (e.g., 0.008) that is held constant until the performance on the cross-validation set does not improve. For every subsequent epoch the learning rate is then divided by a factor of 2. The training is stopped when the performance of the cross-validation set shows no further improvement. Such tuning of the learning rate based on the cross-validation was found to result in faster convergence of the network. Another technique used in the ICSI system is random presentation of the input feature vectors. This technique helps speed up the on-line training.

Input

To take into account any correlations between successive speech feature vectors, the system allows for the use of contextual input, — i.e., in addition to the features belonging to the current frame the input can contain features corresponding to 'c' frames preceding it and 'c' subsequent frames. We have used $c = 4$ in most of the experiments — i.e., the input consists of features corresponding to 9 frames at any given time. Experiments at ICSI have found the use of 8 contextual frames (in addition to the current center frame) to yield good recognition performance. The input features are also normalized to have zero mean and unity variance. The normalizing mean and variance values are computed only over the training data and applied to the test data during recognition. It is also possible to compute the normalizing mean and variance using the test data by an on-line adaptation technique (Tibrewala and Hermansky, 1997). Such an adaptation is found to reduce the mismatch between the means and variances of the training and testing data. However, such an adaptation was not used for the experiments in this dissertation.

Output

The probability estimator converts the input feature vectors into estimates of posterior probability for each of the output classes. The ICSI system uses context-independent phones as the output classes. The phones used in our experiments are a subset of the 61

phones used in the TIMIT database and listed in Appendix A.

2.1.3 Decoding

The next stage in an ASR system is to find the most probable sequence of words given the sequence of emission probabilities (generated by the probability estimator for the incoming speech feature vectors) with additional use of knowledge sources like a language model and lexicon (dictionary). This stage is referred to as speech decoding. The ICSI hybrid system uses the *y0* (called "why not") decoder (Robinson et al., 1993) which uses Viterbi decoding (a variant of dynamic programming) to find the best possible utterance.

As mentioned earlier the MLP actually estimates posterior probabilities (i.e., probability of a phone given the input feature vector) while in the HMM formalism the decoder uses emission probabilities which are actually the likelihoods (i.e., the probability of the input vector given the phone). In a hybrid system, these emission probabilities are estimated by applying the Bayes' rule to the MLP outputs as given by the equation

$$\frac{p(x_n|q_k)}{p(x_n)} = \frac{p(q_k|x_n)}{p(q_k)} \tag{2.1}$$

where $p(x_n|q_k)$ represents the emission probability for the feature vector x_n given the phoneme q_k, $p(q_k|x_n)$ represents the MLP output corresponding to q_k which represents the estimated posterior probability and $p(q_k)$ represents the estimates of class priors. The class priors are estimated by the relative frequencies of the class as determined by the class labels from the training data. The ICSI system uses the scaled likelihood, $\frac{p(x_n|q_k)}{p(x_n)}$, on the left-hand side of Eq. 2.1 as the emission probability for HMM during decoding.

Lexicon

In ASR, each word is described as a succession of phones. The sequence of phone models which constitute a particular word in the recognizer vocabulary — i.e., the pronunciation(s) of all the words, is specified by a lexicon, also called the dictionary. The lexicon is one of the constraining knowledge sources used by the decoder. The lexicon specifies the sequence of HMMs associated with each word. In a hybrid system, each phoneme q_k is associated with a single MLP output $P(q_k|x_n)$ — i.e., each phoneme is modeled with

a single conditional density associated with the corresponding MLP output. To introduce minimum phoneme duration constraints, each phone q_k is actually modeled by $D/2$ states, where D is a prior estimate of the phoneme duration. Each of these states share the same posterior probability estimate $P(q_k|x_n)$ and have equal transition probabilities (0.5). In the ICSI system, the lexicon, in addition to specifying the word pronunciations, also incorporates this state repetition for incorporating duration constraints, as shown in the lexicon block in Fig 2.1.

Language model

For conversational ASR systems, which require the recognition of word sequences, the decoder generally uses additional constraints in the form of a language model (also called grammar). The language model specifies the probability of occurrence of a word w_i following the word sequence $w_{i-1}, ..., w_{i-N+1}$. This is called an N-gram language model. The N-gram probabilities can be computed from the training data by counting the number of occurrences of each word sequence.

Additional decoder parameters

The decoder uses two additional parameters — word transition penalty and the language model scale factor, which are used to balance the proportion of word insertion errors to word deletion errors. The word transition penalty is a penalty imposed when moving from the end of one word to the beginning of another word during decoding. The language model scale factor weights the influence of the language model score over the acoustic score. The values of these parameters are empirically determined during testing.

2.2 Databases

The experiments reported in this dissertation are based on three databases — the Bellcore isolated digit database, the OGI-Numbers database and the OGI-Stories database.

Bellcore Digit database

The multi-band experiments described in Chapters 3 and 4 are based on this database. This database is a telephone-quality isolated digit database. It consists of a 13 word vocabulary consisting of the ten digits (zero, oh, one, two, three, four five, six, seven, eight, nine) and two control words (yes, no). The data is sampled at 8 kHz. The database comprises 200 speakers. Each speaker has recorded one utterance per word. The training set consists of 150 speakers of which 132 speakers (1716 utterances) are used for training and the remaining 18 speakers (234 utterances) are used for cross-validation. The remaining 50 speakers (650 utterances) comprise the test set. Four such cuts of the training and test set are defined to efficiently use the available data. The reported recognition performance is the average of the performance on these four cuts of the database (unless stated otherwise) and thus effectively comprises of 2600 (650×4) utterances. The training set has approximately 65,000 frames for each of the four defined cuts and the cross-validation set comprises of about 9000 frames. This database is labeled in terms of the 61 phonemes from the TIMIT database (refer to Appendix A).

The lexicon used in the recognition experiments was developed at ICSI. It consists of single pronunciations for each of the 13 words based on the most likely TIMIT pronunciations for these words. Since it is an isolated digit task, it does not use any grammar, i.e., each word is assumed to be equally likely to occur.

OGI-Numbers database

The recognition experiments described in Chapter 6 are based on this database. The OGI Numbers corpus (Cole et al., 1995) consists of a set of continuous, naturally spoken utterances collected from many different speakers over the telephone and sampled at 8kHz. The utterances represent the numbers portion of utterances where people recite their addresses, telephone numbers or zip codes. The utterances vary in length from one to ten numbers per utterance (most utterances have five numbers). The task has a 32 word vocabulary consisting of the words — zero, oh, one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen,

twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety, hundred, uh and um.

Three independent subsets of this database (defined at ICSI) of approximately 1.7 hours, 0.6 hours and 0.2 hours respectively have been used for recognition experiments. The 1.7 hours subset is the training set, the 0.2 hours subset forms the cross-validation set (approximately 64,000 frames) on which the frame-level errors are reported, and the 0.6 hours subset (4670 words) comprises the test set on which the word-level errors are reported. The Numbers database consists of 29 phonetic classes which are a subset of the 61 phonemes from the TIMIT database.

The lexicon for OGI-Numbers database comprises of multiple pronunciations for each of the 32 words, developed at ICSI. The pronunciations are based on the hand-labeled phonetic transcriptions of the training set. This task uses a bigram language model.

OGI-Stories database

We use this database in Chapters 5 and 6 for analysis of the phonetic information in the temporal domain. It comprises of the English portion of the OGI multi-lingual database (Cole et al., 1994). The database consists of telephone quality conversational speech. A subset of approximately three hours is phonetically hand labeled and comprises approximately 50 seconds of extemporaneous speech from each of the 210 different speakers. We have considered 45 most frequently occurring phonetic classes (subset of the 61 phonemes in the TIMIT phone set) from this database for our analysis.

2.3 Summary

In this chapter, we have described the broad framework of the ASR system used for the experiments reported in this dissertation. This ASR system also represents a conventional ASR system. The work in this dissertation involves modification of the feature extraction and the phone probability estimation stages of the ASR system. The modifications are aimed at improving the overall recognition performance under mismatched test conditions. The rest of the dissertation describes and discusses these modifications, their motivations and the resulting effect on ASR performance.

Chapter 3

Multi-Band Model

In this chapter, we present a new parallel model of automatic speech recognition, the multi-band (also called the sub-band) model. The model is motivated by Fletcher's work on the Articulation Index. It combines independent phonetic probability estimates obtained from different sub-bands of the speech spectrum for phonetic classification.

This chapter is organized in two parts. The first part, which consists of Sections 3.1 through 3.4, discusses the motivation leading to the development of the sub-band model. In Section 3.1, we revisit the conventional system from the point of view of its performance in narrow-band noise. Section 3.2 describes the psychoacoustical results obtained by Fletcher which form the motivation of the proposed sub-band model. It also reviews other psychoacoustic studies on human intelligibility in various band-limited speech conditions. Section 3.3 reviews the on-going research efforts to address the degradation of ASR performance in narrow-band noise. Section 3.4 describes our multi-band approach.

The second part of the chapter, comprising Sections 3.5 through 3.8, is based on our empirical studies. Section 3.5 describes our experimental setup. Section 3.6 demonstrates the feasibility of the sub-band model. Section 3.7 aims to address the four main issues in the design of the sub-band model — 1) the definition of the sub-bands, 2) the features to be used in each sub-band, 3) the merging technique for combining the sub-band outputs and 4) the linguistic unit level at which the sub-bands should be merged. Section 3.8 demonstrates the applicability of the sub-band model to a large-vocabulary continuous speech task. The chapter concludes with a summary in Section 3.9.

3.1 Conventional System

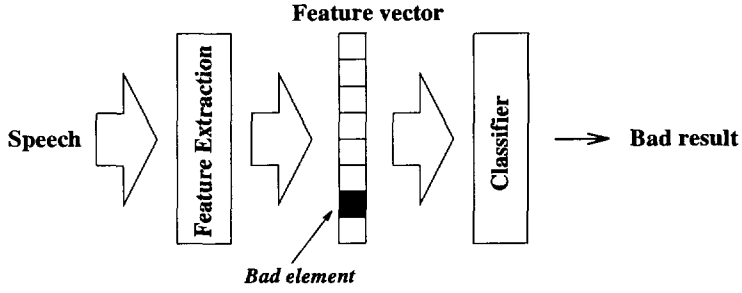


Figure 3.1: Conventional ASR.

The first step in most of the current ASR systems is to convert the incoming speech signal into series of short-term (10-20 ms) frames. A feature vector is extracted from each frame. Each element of the vector describes part of the information carried by the signal. For example, each element of the short-term spectral vector represents the energy of the speech signal in a given frequency band. A typical ASR system uses the entire feature vector as one entity, for subsequent classification into one of the linguistic classes such as phonemes, syllables, words or utterances.

3.1.1 Conventional system in narrow-band noise

Consider the case when some of the elements of the short-term vector are corrupted or carry unreliable information, while the remaining ones are still uncorrupted (Fig. 3.1). This can occur, for example, when the speech signal gets corrupted by narrowband noise. In the current conventional ASR system, since the entire feature vector is used as one entity, even a single corrupted spectral element can severely degrade the performance of the recognizer. This degradation in the performance of the recognizer can be annoying to a system user who has relatively little difficulty in understanding such partially corrupted speech.

3.2 The Human Way — Evidence on How Humans Process and Recognize Speech

3.2.1 Fletcher's study on Articulation Index

Fletcher's work on Articulation Index[1] (Fletcher, 1953) (reviewed in (Allen, 1994; Allen, 1996b; Allen, 1996a)) suggests that human auditory perception works in a different manner than conventional ASR systems. Fletcher and his colleagues at Bell Laboratories extensively studied human speech recognition, in order to quantify the quality of speech sounds in the telephone network, with a view to improve speech intelligibility over the telephone. They used a balanced set of nonsense CVC syllables for these perceptual studies.

Fletcher defined the term *articulation* as the probability of correct recognition for nonsense speech sounds. He empirically determined (by counting the number of times the listeners correctly identified the speech sounds) that the relation between nonsense CVC syllable articulation (S), and the consonant (c) and vowel (v) articulations is given by

$$S(\alpha) = c^2(\alpha)v(\alpha) \approx s^3(\alpha), \tag{3.1}$$

for all values of the speech gain, α, which defines the speech signal-to-noise ratio. This relation meant that within the bounds of this experimental setup, the phones constituting the syllable are decoded independently by the auditory system. Fletcher also showed that, within a small error, the syllable articulation equals the cube of the average phone articulation, $s = (2c + v)/3$.

To further understand how humans decode phones, Fletcher studied the effects of channel frequency response and channel noise on phone perception, by low-pass and high-pass filtering the speech. His studies showed that the partial phone articulations s_L and s_H in the low-pass and high-pass bands respectively, do not sum to the wide-band articulation s — i.e.,

$$s(\alpha) \neq s_L(f_c, \alpha) + s_H(f_c, \alpha), \tag{3.2}$$

[1] Articulation Index is a speech recognition measure which characterizes speech intelligibility under conditions of filtering and noise.

where f_c is the filter cut-off frequency. However, he hypothesized the existence of a non-linear transformation on s, $A(s)$, which would satisfy the additive relation,

$$A(s(\alpha)) = A(s_L(f_c, \alpha)) + A(s_H(f_c, \alpha)), \tag{3.3}$$

for all values of f_c and α. This non-linear transformation was defined as the articulation index. He empirically determined $A(s)$ by finding for all values of the gain α the cutoff frequency, $f_c = f_c^*$, such that $s_L(f_c^*, \alpha) = s_H(f_c^*, \alpha)$. At this frequency, f_c^*, since the low-pass and high-pass speech articulations are equal, they are related to the wide-band articulation by

$$A(s_L(f_c^*, \alpha)) = A(s_H(f_c^*, \alpha)) = 0.5A(s(\alpha)). \tag{3.4}$$

Thus, empirically determining $A(s)$ for all values of α, Fletcher found that for nonsense CVC syllables, $A(s)$ is given by the functional form

$$A(s) = \frac{log_{10}(1-s)}{log_{10}(1-0.985)}, \tag{3.5}$$

where 0.985 is the maximum observable articulation under ideal conditions ($\alpha = 1$). Using this transformation, $A(s)$, in Eq. 3.4 gives

$$(1-s) = (1-s_L)(1-s_H), \tag{3.6}$$

$$\text{and} \qquad e = e_L e_H, \tag{3.7}$$

where e represents the articulation error given by: $e = 1 - s$. Based on the additive nature of the articulation index given by Eq. 3.3, Fletcher and Stewart further generalized Eq. 3.7 to the multi-channel model of phone perception given by

$$e = e_1 e_2 ... e_K, \tag{3.8}$$

where the speech is filtered by K band-pass filters.

This relation between the wide band phone articulation and the partial phone articulations suggests that the errors in a given frequency band are independent of the errors in the other frequency bands. Allen interpreted Eq. 3.8 to imply that the phones are processed in independent frequency sub-bands and the resulting independent phone estimates are then combined to yield the final recognition.

Consider a two frequency band case. If only the high-pass filtered speech is corrupted by noise while the low-pass filtered speech is clean, the recognition error should be dictated largely by the error in the low-pass filtered frequency band according to Eq. 3.7. It is unaffected by the errors in the high-pass filtered band. For example, say, the recognition error on the low-pass filtered speech is 10% and the recognition error on the high-pass-filtered speech is 90% due to noise. Then according to Eq. 3.7, the error on the wideband speech will be $e = 0.1 \text{ x } 0.9 = 0.09$, which is close to the error on the clean low-pass filtered sub-band.

One interpretation of Eq. 3.8 is that the information from the corrupted sub-bands does not have to be used for speech recognition. This notion also receives support from the human speech perceptual studies described in the next section.

3.2.2 Additional studies of human performance on filtered speech

Other researchers have also performed perceptual experiments to test the effect of filtered (low-pass, high-pass, bandpass and band-reject filtered) speech on intelligibility. The aim of these experiments was to gain a better understanding of human speech perception and, in some cases, to find the contribution of different portions of the speech spectrum to intelligibility.

Lippmann (Lippmann, 1996) showed that humans can recognize speech produced by severe band-reject filtering. Specifically, he showed that when nonsense CVC syllables are filtered with a low-pass filter at 800 Hz, the intelligibility of consonants is only around 44%. However when this low-pass filtered speech is presented along with speech high-pass filtered at 4kHz (i.e., mid frequencies from 800 Hz to 4 kHz are filtered out), the intelligibility increases to around 90%. This demonstrates that humans are able to achieve good recognition performance even on this unnatural disjoint band task.

Riener (Riener et al., 1992) and Warren (Warren et al., 1995) studied the intelligibility of words in sentences heard through narrow spectral slits. They found that relatively little spectral information is required to identify component words in sentences. The intelligibility of sentences filtered with single 1/3-octave bands with steep filter slopes (96 dB/octave) averaged more than 95%, for an extended range of center frequencies (maximum around

1500 Hz). Even with severe filtering (1/20-octave bands) intelligibility remained relatively high, with maximum of 77% around 1500 Hz. Further, the intelligibility was roughly 23% for individual 1/3-octave bands centered at either 370Hz or 6kHz, but increased to 77% (much higher than simple additivity of individual intelligibility) when the two bands were presented simultaneously. The results showed that bandpass filtered sentences maintained high intelligibility, even though features characterizing the component phonemes were severely distorted or absent. They concluded that the listeners possess processing mechanisms and strategies employing limited spectral regions that can enhance comprehension under difficult listening conditions.

Kryter (Kryter, 1960) investigated the effect of several narrow pass bands placed simultaneously at different points along the frequency scale, on speech intelligibility. The test set consisted of 50 phonetically balanced words on a carrier phrase and band pass filters with 500Hz bandwidth and slopes 70dB/octave. He found that the intelligibility of single pass bands was around 30%. If two pass bands are presented simultaneously intelligibility increases to 50 - 75 %, depending on which center frequencies are combined. Intelligibility further increases to 70 - 85% when three pass bands are presented simultaneously.

Other studies of human recognition of bandlimited speech include the work by Miller and Nicely (Miller and Nicely, 1995). The task was the identification of one of sixteen consonants in syllables of the form C-/a/, where the stimuli had been bandlimited to various degrees. Their results (for example, 50% accuracy in passband 200-600 Hz and 73% accuracy in 200-2500 Hz band-limited speech) again point to the relatively high recognition rates achievable by humans for narrowband speech.

The results from the studies discussed above show that humans can recognize speech with relatively high accuracy even with limited (due to filtering) spectral cues. This suggests that the speech spectrum contains significant amount of redundancy. The studies also point to the ability of human listeners to easily integrate acoustic cues from different (disjoint) frequency regions for speech perception. The results in the above studies which use nonsense CVC syllables as stimuli are particularly important, because the recognition of the individual phones in these experiments does not use the prior information that can be used when meaningful words or syllables are used (this prior information is usually

referred to as context). Most of the information available to the listeners is acoustic. This supports the hypothesis that bandlimited speech contains a large number of cues necessary to identify speech elements without the aid of context.

The conventional system, described in Section 3.1, however, cannot deal with such bandlimited speech. Since the entire spectrum is used as a single template in recognition, errors across frequency are not treated as independent by the recognizer. Hence, degradations in a limited frequency region affect the entire spectral template.

3.3 Other Research on Recognition of Partially Corrupted Speech

On-going work specifically towards recognition of partially corrupted speech is being carried out by Cooke and Green et al. (Cooke et al., 1994; Green et al., 1995) and Lippmann and Carlson (Lippmann and Carlson, 1997). Their approach is based on missing data (occluded speech) recognition which deals with recognition when some of the normally available speech evidence is missing.

The work by Cooke et al. is motivated by auditory scene analysis (ASA), which refers to the ability of human listeners to separate out and pay selective attention to individual sound sources (Bregman, 1990). The general approach is to computationally separate the evidence of speech signal from arbitrary intrusions. The speech evidence thus segregated is generally incomplete since some spectro-temporal regions will be dominated by other sound sources. They have proposed two techniques for the recognition of such speech with occluded regions (Cooke et al., 1994). The first technique is based on unsupervised training of Kohonen networks (Kohonen, 1984) using a modified training procedure for data vectors with missing components. The second technique is based on a generalization of the continuous density hidden Markov model (HMM) recognition. This technique modifies the HMM probability computation during recognition by making use of a marginal distribution of only those components present in the feature vector (Green et al., 1995; Cooke et al., 1996). Experiments with simulated occlusions by uniform random deletion of parameter vector (e.g., Mel-filter bank energies or auditory spectral representations)

components show little degradation in recognition performance for up to 50% to 80% deletion of components. Experiments with local SNR-based deletions also show improved performance over conventional recognition, though to a lesser extent than that obtained for random deletions.

Lippmann and Carlson have also used dynamic modification of HMM probability computation during recognition, for dealing with missing features. Their preliminary experiments on filtered (low-pass, high-pass and notch filtered) digit recognition task with prior information (true knowledge) about the missing features, have shown promising improvement in recognition accuracy.

These approaches to recognition of partially corrupted speech do not explicitly model partial sub-band phonetic recognitions, as hypothesized by Fletcher's perceptual studies. Our approach, the sub-band model, explicitly models this hypothesis.

3.4 Our Approach — the Sub-band Model for ASR

In many environmental conditions (as discussed in Section 3.1.1), ASR recognition performance could be improved if the recognizer had the human ability to de-emphasize the unreliable frequency sub-bands, provided the remaining clean sub-bands still supply sufficiently reliable information. Our approach to ASR is based on such a paradigm.

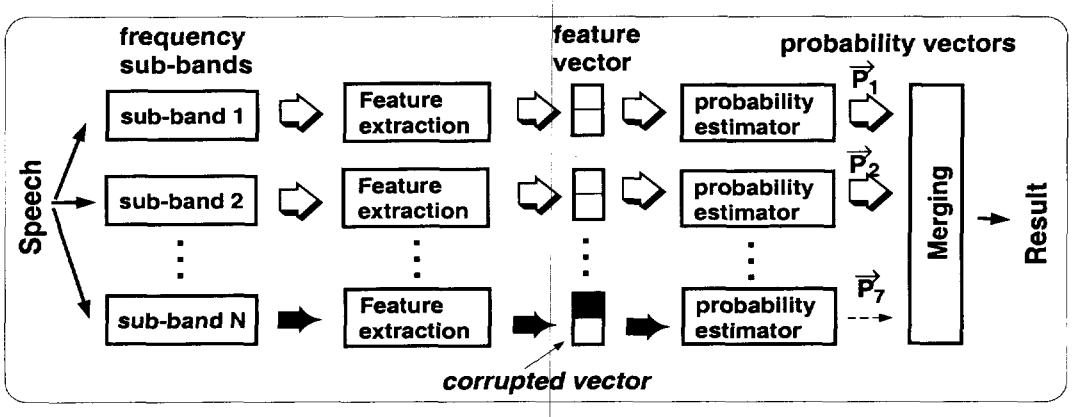


Figure 3.2: Sub-band Model.

Fig. 3.2 is the block diagram of the sub-band model. In this model, we subdivide the speech spectrum into 'N' sub-bands. Spectral features are extracted from each of these

bandlimited regions. This is followed by independent class-conditional probability (or likelihood) estimation in each sub-band, based on the corresponding sub-band features. The probability estimation can be carried out by using standard probability/likelihood estimators like HMM or artificial neural networks such as multi-layer perceptrons (MLP). This stage yields N vectors of class-conditional probability estimates for each input speech frame. The probability estimates from each of the N sub-bands are then merged by a suitable merging algorithm. The output of the merging stage is a vector representing the merged probability estimates. These can be further used in decoding. The output of the merging stage may also directly represent the final decoded result, if the merging is implemented as a part of decoding (discussed in detail in Section 3.7.4).

Fig. 3.2 shows that when there is noise which corrupts only part of the speech spectrum (e.g., last sub-band in the figure), and the identity of the corrupted sub-band(s) can be determined, the model allows for the de-emphasis of the outputs of those sub-band(s) during merging. The hypothesis is that the use of the more reliable sub-bands should yield less degradation in recognition performance as compared to that of the conventional system.

The sub-band model thus represents a practical engineering approach to Fletcher's model for human speech recognition. The use of features extracted independently from the frequency sub-bands also receives support from the work on spectral shape analysis in the central auditory system (Wang and Shamma, 1995), which suggests that the auditory system employs local descriptors of features in the different frequency regions.

Let,

- N represent the number of sub-bands.
- x_b represents the feature vector in sub-band b corresponding to a single frame or the sequence of feature vectors corresponding to a sub-unit such as a phone, syllable, word. Since there are N sub-bands, $x_1 ... x_b ... x_N$ represents the feature set corresponding to the respective sub-bands.
- M is the number of classes (phones, syllables or words) in the recognition task.

- $c_1, ..., c_k, ...c_M$ represents the M class symbols respectively.

The output of the b^{th} sub-band probability estimator is then a vector with elements

$$P(c_k|x_b) \quad \text{for posterior probability estimation}$$
$$\text{or,} \quad P(x_b|c_k) \quad \text{for likelihood estimation}$$

The desired output of the merging stage is then a vector with elements

$$P(c_k|x_1, ..., x_b, ..., x_N) \quad \text{in the posterior probability framework,}$$
$$\text{or,} \quad P(x_1, ..., x_b, ...x_N|c_k) \quad \text{in the likelihood framework,}$$

which represents the joint class-conditional probability of the sub-band input feature vectors.

If the merging is carried out at the decision-making stage of the recognizer, for example, after sub-band word decoding in case of an isolated word recognition task, then the decision is based on -

$$c_k = \max_j P(c_j/x_1, ..., x_b, ...x_N) \quad \text{for } j = 1, ..., M \text{ in the posterior probability framework}$$
$$\text{or, } c_k = \max_j P(x_1, ..., x_b, ...x_N/c_j) \quad \text{for } j = 1, ..., M \text{ in the likelihood framework.}$$

We use the likelihood approach that is typically used in HMMs and also the hybrid (HMM/MLP) systems. When referring to MLP outputs for sub-band probability estimation in the hybrid approach, or in the case of non-linear merging (to be discussed in Section 3.7.3), the use of likelihoods implies that the posterior probabilities estimated by the MLP have been transformed to scaled likelihoods using Eq. 2.1.

3.4.1 Previous work using the multi-band model

A model similar to the sub-band model described above has been previously used in (Duchnowski, 1992). The approach, referred to as the *post-labeling integration* approach, was motivated by the human perceptual experiments on band-limited nonsense syllables by Miller and Nicely (Miller and Nicely, 1995) (discussed briefly in Section 3.2.2). In this approach, tentative decisions or phonetic labels of the identity of a given speech frame

were assigned in parallel by sub-recognizers, each operating on a band-limited portion of the speech waveform. Four non-overlapping frequency sub-bands roughly spanning the formant frequency regions were used. The labels generated by the four sub-band HMM recognizers for each frame were subsequently integrated using a maximum-a-posteriori approach. To address the problem of frequent single frames of different labels following one another in the stream of frame labels generated by the label-integrator, a second HMM was trained to produce the final phonetic segmentation. Phonetic recognition on the TIMIT database using the post-label integration model obtained performance close to that of a comparable conventional recognizer.

The main difference between this approach and our proposed multi-band approach is in the integration of the sub-band recognizer outputs. Post-labeling integration allows for merging based on only the top choice (the discrete class label) from each of the sub-bands. The sub-band recognizers inherently yield poor performance due to their band-limited nature (discussed later in Section 3.7.1). Hence, limiting the integration to have access only to the top choice label could limit the performance of the integrator. In our approach, we propose to use all the probabilities estimated by the sub-band recognizers for merging. In addition, our approach does not require the use of a second stage for phonetic segmentation. Besides, the work on the post-labeling integrator did not deal with recognition in the presence of noise. In our work (Chapter 4), we specifically discuss the applicability of the multi-band model in noisy conditions.

3.5 Experimental Setup

Our experiments on the sub-band model use the telephone-quality isolated-digit Bellcore database described in Section 2.2. Each feature vector represents a 25ms speech frame (using Hamming window) with 12.5 ms of overlap between consecutive frames. The features used in all the experiments are PLP (Hermansky, 1990) based features. We use as features either the critical band energies followed by the perceptually motivated equal-loudness pre-emphasis and cube-root compression, or the cepstral coefficients obtained from the autoregressive coefficients of an all-pole model of the equalized and cube-root compressed

critical-band energies. The cepstral coefficients are appended with the delta cepstral coefficients, 1 delta energy coefficient and 1 delta-delta energy coefficient. The critical band filters used in the feature computation are either the 15 filters at approximately 1 Bark spacing or 30 filters at approximately 0.5 Bark spacing. The cut-off frequencies of the filters are as specified in the Tables 3.1 and 3.2. The multi-band system uses sub-bands defined in terms of groups of critical-bands and not derived by explicit bandpass filtering of the speech signal.

Table 3.1: Cut-off frequencies of the 1-Bark spacing critical band filters used in PLP feature representation for 8 kHz sampling frequency.

Critical band number	Lower cut-off frequency (Hz)	Upper cut-off frequency (Hz)
1	17	161
2	115	265
3	216	375
4	323	495
5	438	629
6	565	778
7	707	949
8	868	1144
9	1051	1370
10	1262	1632
11	1506	1937
12	1790	2292
13	2122	2709
14	2509	3197
15	2962	3769

The recognizers used in all the experiments are hidden Markov model/multi-layer perceptron (HMM/MLP) based hybrid recognizers (Bourlard and Morgan, 1994), as described in Chapter 2. The MLP generates estimates of phone posterior probabilities for each input frame. The phoneme set used in this experiment consists of 61 context independent phonemes based on the TIMIT phone set, (listed in the Appendix A). The output of the MLP thus has 61 units, one unit per phonetic class. The input to the MLP consists of the features from the current frame to be classified, along with features from 8 additional surrounding frames (4 frames each from the past and the future respectively) to provide contextual input, as mentioned in Section 2.1.2.

Table 3.2: Cut-off frequencies of the 0.5-Bark spacing critical band filters used in PLP feature representation for 8kHz sampling frequency.

Critical band number	Lower cut-off frequency (Hz)	Upper cut-off frequency (Hz)
1	14	158
2	62	208
3	109	258
4	158	310
5	207	365
6	258	421
7	310	480
8	364	542
9	420	607
10	479	677
11	541	750
12	607	828
13	676	911
14	749	1000
15	827	1094
16	910	1196
17	998	1305
18	1093	1422
19	1195	1548
20	1304	1684
21	1421	1830
22	1547	1987
23	1682	2157
24	1828	2340
25	1985	2537
26	2155	2750
27	2337	2980
28	2535	3229
29	2747	3497
30	2977	3788

3.5.1 Baseline system

We have defined two baseline systems based on the type of features used. The first baseline system uses the 15 equalized and compressed PLP critical band energies as features. The MLP thus has 135 inputs (15 x 9 context frames), 200 hidden units and 61 output units. The second baseline system uses 8 cepstra, 8 delta cepstra, 1 delta and 1 delta-delta energy coefficients. The MLP in this system has 162 inputs (18 x 9 context frames), 200 hidden units and 61 output units. Note that for the sub-band experiments we have not optimized the size of the hidden layer.

3.6 Feasibility Study

In the case of the multi-band system, an important question to be answered is — in terms of the recognition performance, is it feasible to independently process different regions of the speech spectrum and then merge these independently processed outputs, without degrading recognition performance as compared to a conventional (full-band) system? In other words, the question is — does independent processing of the frequency sub-bands result in loss of discriminatory information, which may be present in the sub-band correlations available to the conventional system?

In order to conduct the feasibility study we divided the speech spectrum into two sub-bands. One possible frequency division of the spectrum is around 1000 Hz, which approximately corresponds to the frequency beyond which the critical bandwidth as well as the perceived pitch increase logarithmically with frequency. Our experiment with two sub-bands approximately followed this frequency division.

Table 3.3: Comparison of the word-level performance of the two sub-band model with that of the baseline system. The critical band definitions are obtained from Table 3.1. The MLP probability estimator has 200 hidden units and 61 output units.

Recognizer	Critical bands	# of critical bands	Lower cut-off frequency (Hz)	Higher cut-off frequency (Hz)	# of inputs to MLP	Word error (%)
Baseline	1-15	15	17	3769	162	3.85
2 sub-band model						
sub-band 1	1-8	8	17	1144	72	14.0
sub-band 2	9-15	7	1051	3769	63	10.65
Merged system (using non-linear merging)					26:26:13	2.27

Table 3.3 describes the baseline features and the sub-band features and also compares the performance of the baseline system with that of the two sub-band system. Non-linear merging using MLP was used for combining the outputs of the two sub-band word probability estimators (the merging technique and merging level are discussed later in Sections 3.7.3 and 3.7.4 respectively). The table shows that:

- The performance of the individual sub-bands is poor as compared to that of the baseline system, as each sub-band operates on a band-limited spectrum.

- In comparison to the baseline (full-band) ASR system, the improvement in performance obtained using the 2 sub-band system is significant at 0.1% level using the binomial significance test.

The results demonstrate that partial information from the individual sub-bands can be successfully merged and also results in improved performance as compared to the baseline system. The results thus demonstrate the feasibility of the multi-band model. However, for a good design of the multi-band system several issues need to be addressed.

3.7 Issues in the Design of the Sub-band Model

The main issues involved in designing the sub-band model are — 1) the definition of the frequency sub-bands, 2) the features to be used in each sub-band, 3) the merging technique and 4) the temporal unit at which the sub-band outputs should be merged.

3.7.1 Definition of frequency sub-bands

The first issue in the design of the multi-band system is — how to subdivide the available speech spectrum into sub-bands. The more and narrower the sub-bands, the better is the chance to alleviate frequency-localized degradation, with reduced loss in spectral information. On the other hand, exceedingly narrow bands could yield a poor discrimination between the linguistic classes because of reduced spectral information in each sub-band.

An important point to consider is the definition of the frequency bands used by Fletcher in his experiments. He referred to the frequency bands as articulation bands. Allen (Allen, 1994) further noted that the articulation bands were similar, though not equal, to the critical bands. This suggests that the frequency division should be somehow related to the frequency division defined by critical band filters.

Allen (Allen, 1994) further suggested that one articulation band roughly represents two critical bands. Given that we have 15 (Bark-spaced) critical bands (Table 3.1), a frequency division of approximately two critical bands per sub-band yields a seven sub-band system. We used two other frequency divisions — the first yielding two sub-bands (as used in Section 3.6). The second uses four sub-bands which roughly correspond to

the four formant frequency regions, as used in (Duchnowski, 1992). The exact critical bands used and the corresponding frequency ranges for the 2, 4 and 7 sub-band models are further detailed in Table 3.4. The sub-bands are overlapping to some extent due to the overlapping nature of the critical bands. The table also lists the size of the sub-band MLP-based probability estimators.

Table 3.4: Word-level performance of the sub-bands in the 2, 4 and 7 sub-band models using PLP critical band energies as features. The cut-off frequencies are as defined by the 15, 1-Bark spaced critical bands, given in Table 3.1. The MLP probability estimator has 200 hidden units and 61 output units.

Sub-band number	Critical bands	# of critical bands	Lower cut-off frequency (Hz)	Higher cut-off frequency (Hz)	# of inputs to MLP	Word error (%)
2 sub-band model						
1	1-8	8	17	1144	72	14.0
2	9-15	7	1051	3769	63	10.65
4 sub-band model						
1	1-6	6	17	778	54	19.95
2	7-10	4	707	1632	36	20.0
3	11-13	3	1506	2709	27	28.73
4	14,15	2	2509	3769	18	49.8
7 sub-band model						
1	1-3	3	17	375	27	58.93
2	4,5	2	323	629	18	30.93
3	6,7	2	565	949	18	34.05
4	8,9	2	868	1370	18	33.10
5	10,11	2	1262	1937	18	29.7
6	12,13	2	1790	2709	18	40.95
7	14,15	2	2509	3769	18	49.8

Table 3.4 shows that the average word error in each of the sub-bands increases from 12% for the 2 sub-band model, to 30% for the 4 sub-band model, and 40% error for the 7 sub-band model, because the sub-bands progressively become narrower in terms of frequency coverage.

3.7.2 Features

Another issue in the design of the multi-band model is the features to be used in each sub-band. We have used critical band energies as features in the earlier experiments. Cepstral features, which are the coefficients of the Fourier transform of the log magnitude spectrum,

are widely used in most speech recognition systems because they have been found to be more robust and reliable for ASR. Cepstral coefficients can also be computed from the autoregressive coefficients obtained from an all-pole model, by a recursive formula (Rabiner and Juang, 1993). Standard PLP analysis approximates the pre-emphasized (equal loudness pre-emphasis) and cube-root compressed critical band energies by an all-pole model, using the autocorrelation method (Hermansky, 1990). The autoregressive coefficients thus obtained are further converted into cepstral coefficients.

We compared the performance of the sub-bands trained on critical band energies (Table 3.4) to those trained on cepstral and dynamic cepstral coefficients obtained from an all-pole modeling of the corresponding critical band energies. To compute the all-pole model of the critical band spectrum in a particular sub-band, we re-index the critical bands in that sub-band such that the first critical band of that sub-band is at zero frequency and the last critical band is at frequency π. Since the sub-bands are comprised of very few critical bands (as few as 2 critical bands in case of the 7 sub-band model), it restricts the model order that can be used for an all-pole modeling. Hence, we have to upsample the critical bands or use more critical bands in a sub-band, to be able to fit an all-pole model of higher order (for e.g., an order higher than 1 for 2 critical bands). We considered 30 critical bands at a spacing of approximately 0.5 Barks. Table 3.2 gives the corresponding critical band filter cut-off frequencies. In addition to cepstral coefficients, we also use delta cepstra, 1 delta energy and 1 delta-delta energy coefficients as features.

Table 3.5 gives the sub-band word error rates for the 4 sub-band model and the 7 sub-band model. Comparing the error rates of Tables 3.4 and 3.5, we see that the individual sub-band errors decrease from an average of 30% using critical band energy features to 16% using cepstra and dynamic cepstral coefficients as features for the 4 sub-band system. Similarly, for the 7 sub-band system the error decreases from an average of 40% error to 23% error. The improvement in performance is not only due to the use of cepstral features but also due to the use of delta coefficients.

Table 3.5: Word-level performance of the sub-bands in the 4 and 7 sub-band models using cepstral based features. The cut-off frequencies are as defined by the 30, 0.5 Bark spaced critical bands (Table 3.2). The MLP probability estimator in each sub-band has 200 hidden units and 61 output units.

Sub-band number	Critical bands	Low cut-off frequency (Hz)	High cut-off frequency (Hz)	Model order	# of cepstra	# of features per frame	# of Inputs to MLP	Word (error %)
4 sub-band model								
1	1-12	17	828	5	7	16	144	14.5
2	13-20	676	1684	5	7	16	144	8.55
3	21-26	1421	2750	3	5	12	108	12.6
4	27-30	2337	3788	3	5	12	108	27.85
7 sub-band model								
1	1-4	14	310	3	5	12	108	48.13
2	5-10	207	677	5	7	16	144	17.18
3	11-14	541	1000	3	5	12	108	18.75
4	15-18	827	1422	3	5	12	108	15.15
5	19-22	1195	1987	3	5	12	108	10.95
6	23-26	1682	2750	3	5	12	108	20.5
7	27-30	2337	3788	3	5	12	108	27.85

3.7.3 Merging techniques

A critical part of the sub-band model is the merging technique to be used for combining the outputs of the sub-band classifiers.

For a combination of the outputs from multiple classifiers to yield performance better than any single classifier used, an important criterion is that the classifiers should yield complementary information (errors). We analyzed the word level decisions of the sub-band classifiers trained on cepstral based features. For all the possible pairs of sub-bands in the 4 sub-band system (6 combinations of 2 sub-bands), on average the sub-bands yield different decisions for 30% of the words in the cross-validation set. Similarly, all possible pairs of sub-bands in the 7 sub-band systems (21 combinations of 2 sub-bands) yield on average different decisions for 38% of the words. This analysis indicates that the sub-bands yield a significant amount of complementary information.

The merging technique should combine the outputs of the sub-band classifiers so as to get maximum benefit from the complementary information. In other words, it should be able to take advantage of the strengths and weaknesses of the individual classifiers, with

the aim to improve the overall accuracy beyond that of any single classifier.

Several combination techniques have been proposed in the pattern classification literature. These include voting schemes like majority voting (Xu et al., 1992) and label ranking techniques (Ho, 1992; Honathan et al., 1994). These techniques can be applied to classifiers that output only discrete labels as outputs. Many classification algorithms (including the hybrid HMM/MLP algorithm that we use) provide continuous numbers as outputs, which represent the degree to which the input belongs to a particular class. For example, neural network classifiers provide estimates of posterior probabilities and HMM classifiers give likelihood estimates. In such cases, linear as well as non-linear merging techniques can be used.

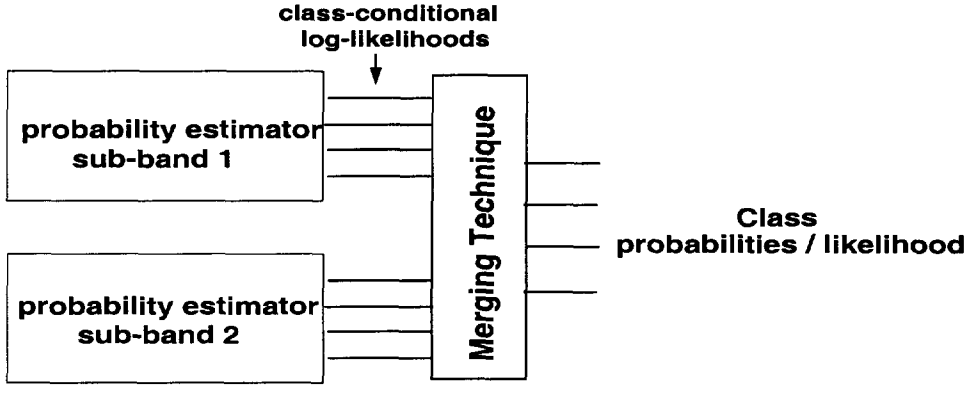


Figure 3.3: Block diagram for merging of a 2 sub-band four class system.

Fig. 3.3 is the general block diagram for merging the outputs of a 2 sub-band, four class system. The four (continuous-valued) outputs each of the 2 sub-bands have to be combined to yield the merged probability estimates for each of the four classes. We have examined weighted linear combination techniques which use simple averaging, as well as weights based on accuracies in each sub-band, and a non-linear combination technique using a neural network.

Linear merging

Linear merging techniques include averaging the classifier outputs or using some weighted linear combination (Xu et al., 1992; Hashem, 1994; Adjoudani and Benoit, 1996; Ghosh

Table 3.6: Comparison of the word errors (%) of the Baseline system and the sub-band systems obtained using different merging techniques.

Baseline	2.0	
	4 sub-band system	7 sub-band system
Individual sub-band	9.0 - 27.5	11 - 48
Linear merging techniques		
equal weights	3.0	4.27
sub-band accuracy weights	2.85	4.23
sub-band class accuracy weights	4.13	5.68
Non-linear merging		
neural network (MLP)	1.31	1.5

et al., 1992). In our case, as mentioned in Section 3.4, the desired output of the merging classifier is some estimate of the joint class-conditional probability $P(x_1, ..., x_b, ..., x_N|c_k)$. If the sub-bands are considered to be independent, then it is given by

$$P(x_1, ..., x_b, ..., x_N|c_k) = \prod_{b=1}^{N} P(x_b|c_k) \quad \text{for } k = 1, ..., M. \tag{3.9}$$

The sub-band classifiers used in our experiments are HMM/MLP based hybrid classifiers, which are based on the likelihood framework (as is the case in HMM classifiers). Specifically, the outputs represent the log-likelihoods of the different words in the vocabulary. Taking log on either sides of Eq 3.9 gives,

$$\log\{P(x_1, ...x_b, ..., x_N|c_k)\} = \sum_{b=1}^{N} \log\{P(x_b|c_k)\}. \tag{3.10}$$

This equation assumes equal weighting of each of the sub-band outputs . However, the sub-bands can be reliable to different extents and a measure of this reliability can be used as a weighting in the linear combination, as given by

$$\log\{P(x_1, ...x_b, ..., x_N|c_k)\} = \sum_{b=1}^{N} w_b \log\{P(x_b/c_k)\}, \tag{3.11}$$

where w_b represents the reliability of the b^{th} sub-band classifier. Eq. 3.11 represents the weighted linear merging of the sub-band classifier outputs.

Averaging - equal weights

The simplest form of merging is averaging the corresponding log-likelihoods. This implies equal reliability of all sub-band classifiers — i.e.,

$$w_b = \frac{1}{N}. \tag{3.12}$$

Weights based on sub-band accuracies

However, Table 3.5 shows that the accuracies $(1 - error)$ of the individual sub-bands vary between 72% - 91% for the 4 sub-band system and 52% - 89% for the 7 sub-band system. These accuracies can be used as a reliability measure to weight the different sub-bands during combination. Let, r_b represent the accuracy of sub-band b for a N sub-band system. Then,

$$w_b = \frac{r_b}{\sum_{b=1}^{N} r_b}. \tag{3.13}$$

Weights based on sub-band class accuracies

The reliability of the individual sub-bands not only varies depending on the frequency coverage of the sub-band, but it may also vary depending on the class (for e.g., phoneme, syllable or word) being classified.

Table 3.7 gives the word accuracies for the each of the four sub-band classifiers in the 4 sub-band system. The table shows that the word accuracies vary significantly across the sub-bands. Similar variation is observed across the sub-bands in the 7 sub-band system. These sub-band word accuracies can also be used as weighting coefficients for merging. Let r_{bk} represent the accuracy for the class k in sub-band b. Then, the weights can be computed as

$$w_{bk} = \frac{r_{bk}}{\sum_{b=1}^{N} r_{bk}}. \tag{3.14}$$

Results of linear merging

Table 3.6 compares the performance of the baseline system (trained on cepstral based features) with that of the multi-band systems which use the three different types of weightings

Table 3.7: Word accuracies in the individual sub-bands of the 4 sub-band system using cepstral based features.

Classes	Accuracies (%)			
	sub-band 1	sub-band 2	sub-band 3	sub-band 4
oh	77.78	93.06	68.06	56.94
zero	91.67	95.83	93.06	86.11
one	87.50	100.00	80.56	79.17
two	76.39	93.06	87.50	83.33
three	75.00	100.00	81.94	76.39
four	93.06	100.00	86.11	68.06
five	98.61	94.44	73.61	76.39
six	98.61	94.44	93.06	87.50
seven	98.61	94.44	93.06	58.33
eight	88.89	73.61	94.44	80.56
nine	56.94	69.44	63.89	47.22
yes	91.67	94.44	80.56	70.83
no	50.00	68.06	51.39	26.39

discussed above for linear merging. The results indicate that the linear combination technique yields significantly reduced word error rate as compared to that of the individual sub-bands. However, the performance of the merged system is worse than that of the baseline system.

Non-linear merging

The linear combination technique is based on the assumption of independence of the sub-band inputs, x_b. But this assumption may not be true. One of the possible reasons is the fact that though the sub-bands are defined on different regions of the frequency spectrum they are overlapping to some extent, due to the overlapping nature of the critical bands. It has been shown that if the sub-classifiers in a multiple classifier system are not independent, the optimal fusion rule has to be non-linear (Pavel and Hermansky, 1997). For merging in this case, a non-linear classifier can be used. In this case, the sub-band classifiers can be thought of as feature extractors instead of classifiers. With this interpretation, the outputs of the sub-band classifiers can be regarded as features to be used for further classification (Huang and Suen, 1994).

We used a multi-layer perceptron (MLP) for non-linear merging. The MLP used is a

three-layered network which is trained using the error back-propagation algorithm. The network should ideally be trained on independent training data, if available, for good generalization, but in this experiment the data used in the training are the same as that used in the training of the sub-band classifiers. The input to the neural network is a concatenated vector of the outputs of the sub-band classifiers. Specifically, the inputs represent the class conditional log-likelihoods obtained from each sub-band classifier. Thus, the merging network has (13 x 4) 52 inputs in case of the 4 sub-band system, and (13 x 7) 91 inputs in case of the 7 sub-band system. The number of hidden units used is 26 and the network has 13 outputs corresponding to the merged estimates of the probabilities of the thirteen word classes.

Table 3.6 shows that the non-linear merging outperforms the linear merging techniques and yields better performance than the conventional system. The improvement in performance is significant at the 0.05 level for the 4 sub-band system, and 0.1 level for the 7 sub-band system, using the binomial significance test. The results also indicate that irrespective of how many sub-bands are used (i.e., 4 sub-bands or 7 sub-bands), the performance of the MLP combiner is similar (i.e., 1.3% word error for the 4 sub-band system and 1.5% for the 7 sub-band system). This reiterates our observation (in Section 3.6) that partial information from the sub-bands can be merged. It also indicates that independent processing of the sub-bands up to the point of merging does not result in loss of performance as compared to the conventional system that uses the entire spectrum as one entity. This observation receives further support from (Mirghafori and Morgan, 1998), where it is shown that the phonetic features are recognized better by the sub-band system than the conventional system on a continuous numbers task.

3.7.4 Merging level

Another issue in the design of the multi-band system is the level at which the sub-bands should be merged - i.e., state/frame-level, phone-level, syllable-level or word-level. This is directly related to the question of relaxing the temporal synchrony among different regions of the frequency spectrum up to the point of merging.

In Section 3.7.3, the sub-bands were merged after the likelihoods of the words were

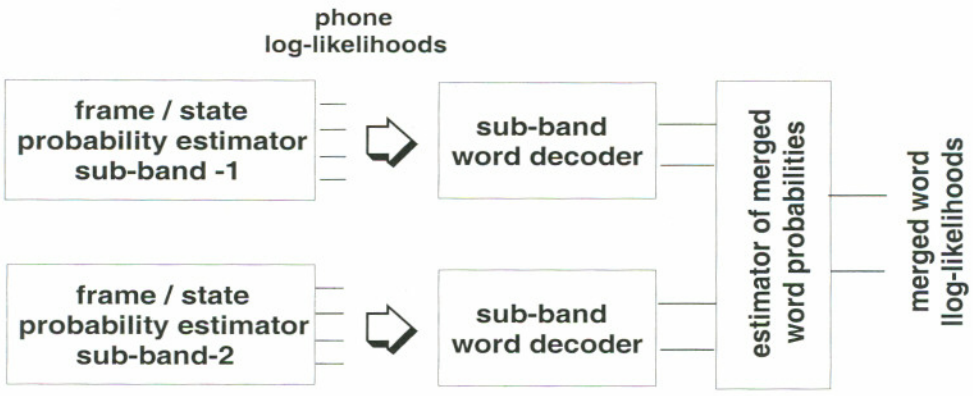


Figure 3.4: Block diagram showing merging of the sub-band outputs of a 2 sub-band system at the word-level in an isolated word task. This system is composed of four phonetic classes and two words.

estimated independently by the decoder in each sub-band, i.e., the sub-bands were merged at the word-level. Fig. 3.4 shows the block diagram of merging the sub-bands at the word-level. Merging at any level higher than the frame-level implies that the optimal paths taken by the decoder could be different in different sub-bands. This could result in different state, as well as different phonetic (and syllable) alignments in the different sub-bands for the same utterance.

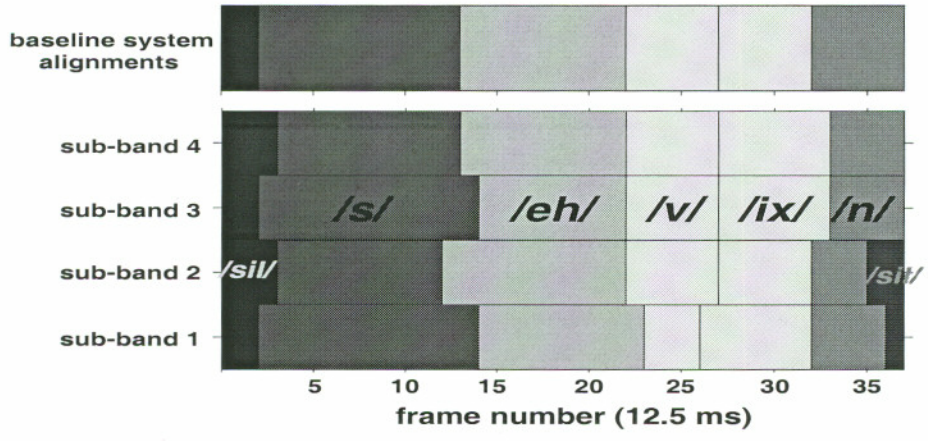


Figure 3.5: Phonetic alignments in the individual sub-bands of a 4 sub-band system as compared to that of the baseline system for an utterance of the word *seven*.

Fig. 3.5 shows the phonetic alignment for an utterance of the digit 'seven', obtained by forced alignment in the different sub-bands of the 4 sub-band model. The sub-bands have

different phonetic alignments, which will result in asynchrony in the different sub-bands if the sub-band outputs are merged at the syllable or word-level. This observation of asynchrony in the different sub-bands is supported by similar observations made in (Mirghafori and Morgan, 1998). We speculated that one of the advantages of the sub-band model could be this relaxation of the temporal synchrony.

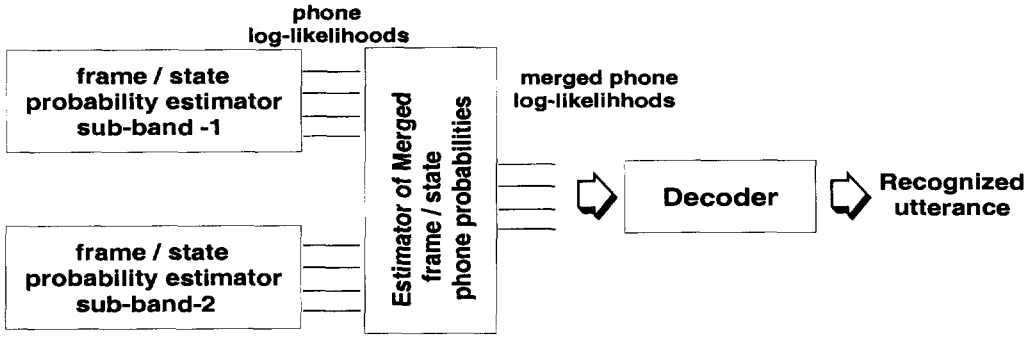


Figure 3.6: Block diagram of merging the sub-bands of a 2 sub-band system at the frame/state-level. This system is composed of four phonetic classes.

To test this hypothesis, we merged the sub-band outputs at the state/frame-level which represents perfect synchrony among the sub-bands. Fig. 3.6 shows the block diagram of merging the sub-band outputs at the frame level. The phone log-likelihoods estimated for a frame in each sub-band are merged. The merged estimates of the class-conditional log-likelihoods are then used in subsequent decoding. Since the non-linear merging technique using a MLP outperformed the linear merging techniques in Section 3.7.3, we used a MLP for computing the merged phone estimates for each frame. The merging network has a single hidden-layer and has (61 x 7) 427 inputs for the 7 sub-band system (61 x 4 = 244 inputs for the 4 sub-band system), 200 units in the hidden layer and 61 outputs corresponding to the 61 phonetic classes.

Table 3.8 compares the performance of the sub-band systems which use a MLP for merging at the word-level (same as in Table 3.6) and at the frame-level. The table shows that the merging levels do not yield a significant (at the 0.1 level of a binomial significance test) change in performance of the sub-band systems on this database.

Since merging at the frame-level combines the partial sub-band phone estimates and

Table 3.8: Word error (%) obtained by merging at the frame-level and the word-level.

Baseline	2.0	
Recognizer	MLP Merging Level	
	Word	Frame
4 sub-band model	1.31	1.52
7 sub-band model	1.5	1.3

Table 3.9: Comparison of the frame/phone errors (%) of the baseline system, the sub-bands of the 4 and 7 sub-band systems, the sub-band systems obtained by merging at the frame-level, and Fletcher and Stewart's multichannel model of phone perception given by Eq.3.8.

Baseline	17.1	
	4 sub-band model	7 sub-band model
sub-band 1	34.41	58.33
sub-band 2	33.62	41.35
sub-band 3	39.26	43.09
sub-band 4	55.98	42.15
sub-band 5	-	38.31
sub-band 6	-	48.77
sub-band 7	-	56.09
Merged system	17.3	17.6
multichannel model	2.54	0.46

yields the combined phone estimate, we can compare the phone errors made by the merged system with that predicted by Fletcher and Stewart's multichannel model of phone perception (Eq. 3.8). Table 3.9 shows this comparison. We see that the merged system yields considerably higher phone error as compared to that obtained by taking a product of the sub-band phone errors (according to Fletcher and Stewart's model). However, even the baseline system which computes the phone error on the entire available spectrum (wide band phone error) yields performance comparable only to that of the merged system and not to that of the multichannel phone model.

Though not conclusive, the results in Table 3.8 indicate that relaxing the time-synchrony

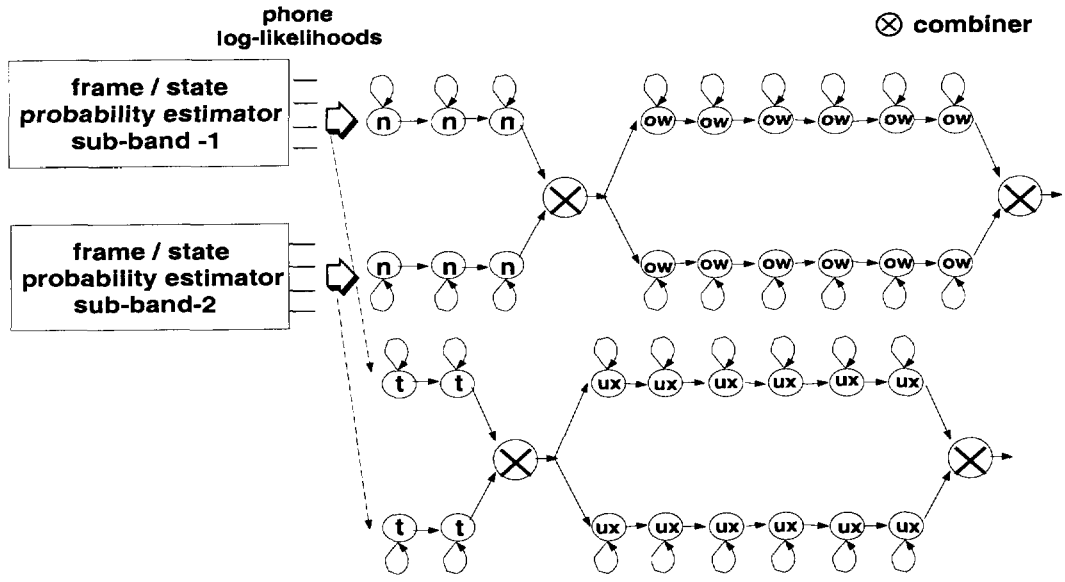


Figure 3.7: Block diagram of merging of the sub-bands of a 2 sub-band system at the phone-level. This system is composed of four phonetic classes and 2 words.

between sub-bands does not improve system performance. Besides extending the applicability of the multi-band model to a large-vocabulary continuous speech task, with limited asynchrony among sub-bands, will require modification of the decoder. The decoder will have to handle multiple sub-band models of the same sub-unit, with synchrony forced at an intermediate sub-unit like phone, syllable or word, as shown schematically in Fig. 3.7. This figure illustrates the merging of two sub-bands at the phone level thus allowing for asynchrony at the state-level. One approach to asynchronous merging can be based on the HMM decomposition technique proposed by Varga and Moore (Varga and Moore, 1990). Another limitation of asynchrony in large vocabulary systems is that it is not straightforward to use non-linear merging. The advantage of synchronous merging of the sub-bands (at the frame/state level) is that it requires modification of only the frame-level probability estimates and not the decoder. In this case, the extension of the multi-band system to large-vocabulary continuous speech tasks becomes straightforward.

3.8 Extension of the Multi-band Model to a Large-Vocabulary Continuous Speech Task

To test the extension of the multi-band model to a large-vocabulary continuous speech task, we used the conversational speech SWITCHBOARD database. The training set consisted of 4 hours (9019 utterances) of male speech. The test data consisted of 240 male utterances. The multi-band model consisted of 7 sub-bands as defined in Table 3.5. PLP cepstral features were used in the baseline system, as well as in each sub-band, with the same autoregressive model order and number of cepstral coefficients, as detailed in Table 3.5 for the 7 sub-band model. In addition to cepstral features, each feature vector also included delta cepstra, delta-delta cepstra, delta energy and delta-delta energy features. Each sub-band MLP probability estimator used 9 frames of input, had 500 hidden units and 56 output units (called the ICSI56 phoneme set) which were a subset of the 61 phonemes used in the Bellcore database experiments. The non-linear merging technique (using MLP) was used for merging the frame-level log-likelihood estimates from the sub-bands. Because of the availability of additional data, the merging network was trained on an independent set of 2 hours of male speech. The decoding was done using the lattice-decoder in the STRUT[2] software.

Table 3.10: Word error % on the SWITCHBOARD database.

Baseline	60.9
7 sub-band system	
Individual sub-bands	68-73
MLP merging	59.0

Consistent with our experiments on isolated digits, the MLP-based non-linear merging again yielded the best performance with about 2% absolute improvement in the error rates. While we also experimented with various linear merging techniques, they were not as effective as the non-linear merging technique.

[2]STRUT is the Speech Training and Recognition Unified Tool developed at the Faculté Polytechnique de Mons.

The improvement in performance of the multi-band system cannot be attributed only to the multi-band paradigm. When we trained an MLP on independent data to re-classify the outputs of the baseline system, the performance of the baseline system also improved to 59% error, which was similar to what we achieved by MLP merging of the sub-bands. However, we can conclude that the sub-band model can be applied for large-vocabulary continuous speech tasks, where it yields performance similar to the conventional full-band baseline system, under matched training and test conditions.

3.9 Summary

In this chapter, we presented a new parallel sub-band model, as an approach to ASR. We empirically demonstrated the feasibility of the model. The multi-band model yields slightly better performance as compared to the conventional system in matched training and test conditions. This indicates that independent processing of the frequency sub-bands before merging does not result in loss of discriminatory information. We also studied some of the main issues to be addressed in the model design stage. Our studies on the issues, though not conclusive, do elucidate the following points —

- It is feasible to sub-divide the speech spectrum in as many as 7 sub-bands given that we have 15 critical bands available up to 4 kHz, without loss in recognition performance on matched testing conditions.

- As has been found in conventional state-of-the art ASR systems, the cepstra and dynamic cepstral features result in good performance at the sub-band level too.

- Since the sub-bands are overlapping to some extent due to the overlapping nature of the critical bands, the sub-band classifiers are not necessarily independent and hence non-linear merging of the sub-band classifiers yields better performance.

- It is possible to merge the sub-bands at various linguistic (frame, phone, syllable, word) levels, but the simplest merging is at the frame level, which can be readily extended to any speech task. Also, allowing temporal asynchrony did not improve

system performance. The next chapter further supports this point using mismatched (noisy) test conditions.

These observations are supported by studies reported in parallel by our collaborators (Bourlard and Dupont, 1996).

The next chapter highlights the benefits of the sub-band model in noise conditions which corrupt only some frequency sub-bands. We also discuss various techniques to practically adapt the sub-band model for such noisy cases.

Chapter 4

Multi-band in Noise

In the previous chapter, we presented a parallel model, the sub-band model, for automatic speech recognition based on merging the independent class-conditional probability estimates in several frequency sub-bands. We empirically demonstrated the feasibility of the model in matched conditions. We also discussed some of the issues involved in the design of the model.

The main motivation for the development of the sub-band model was to deal with noise conditions that corrupt only part of the speech spectrum. In such noise conditions, some part of the speech spectrum still carries reliable information. The two main issues in the recognition of speech corrupted by noise are - 1) how to determine the reliability of the sub-bands, and 2) how to de-emphasize the less reliable sub-bands. This chapter explores techniques to address these two issues. These techniques are subsequently tested in the presence of realistic noise conditions.

Section 4.1 demonstrates the feasibility of recognition from band-limited speech. The recognition performance on limited spectral regions also demonstrates the spectral redundancy in speech and its potential advantage in band-limited noise conditions. In Section 4.2, we discuss various techniques for the dynamic selection of reliable sub-band classifiers during recognition. These techniques include sub-band SNR estimation, classifier confidence estimation, majority voting and adaptation. Since the main advantage of the sub-band model is expected to be its performance in band-limited noise, we demonstrate the feasibility of the model and the effectiveness of the selection techniques, on controlled band-limited noise — i.e., sinusoidal noise at different signal-to-noise ratios. In Section 4.4, we test the effectiveness of the model in several realistic additive noise

conditions from the NOISEX-92 database. The experiments demonstrate the robustness of the sub-band model (more than 50% reduction in word error rate on average) to noise which corrupts some sub-bands more than the others. We conclude with a summary of this chapter in Section 4.5.

4.1 Sub-band Model on Band-limited Speech

In some noise conditions, when only part of the speech spectrum is corrupted, the remaining speech spectrum may still carry reliable information. One of the main advantages of the sub-band model, as mentioned in Section 3.4, is that it allows for the de-emphasis of the unreliable or corrupted sub-bands during the merging process. To determine the feasibility of the sub-band model in such noise conditions, we tested its performance on band-limited speech.

An approach to simulate band-limited speech is to drop some sub-band(s) from the merging process. In the 7 sub-band system, considering all possible combinations of the 7 sub-bands with up to 6 sub-bands dropped, gives 127 different band-limited speech conditions, as shown in Table 4.1. The table entry with 6 sub-bands excluded refers to a system with just one of the seven sub-bands and hence implies each individual sub-band classifier. Since non-linear merging using a MLP yielded the best performance of the sub-band model in our experiment reported in Section 3.7.3, a MLP is trained for each of these 127 sub-band combinations.

Table 4.1: Different combinations of sub-bands in the 7 sub-band model.

# of sub-bands excluded	# of Merging Networks
0	1
1	7
2	21
3	35
4	35
5	21
6	7
Total	127

Fig. 4.1 gives the range of word errors for the 127 merging classifiers trained on all

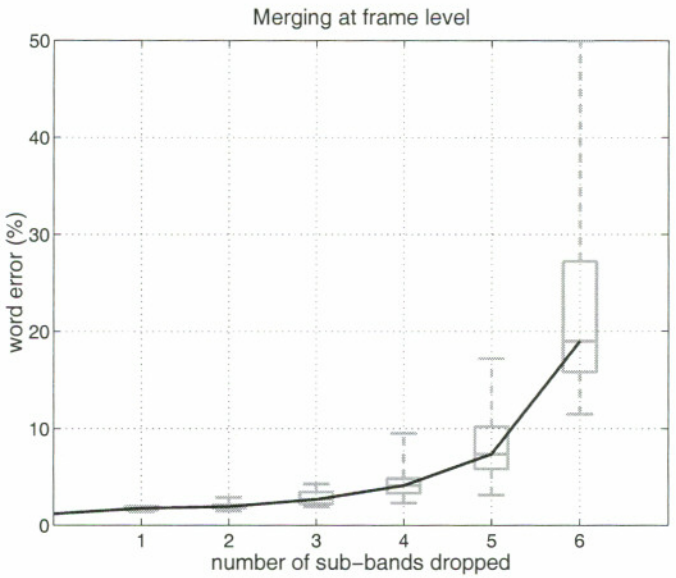


Figure 4.1: Word error (%) with merging networks having 0 to 6 sub-bands excluded. Merging is done at the frame level.

combinations of the sub-bands in the 7 sub-band model. Each sub-band is trained on cepstral based features and the merging is done at the frame-level. As in Section 3.7.4, each of the merging networks is a three-layered network and has 61 inputs from each of the sub-bands to be merged, 200 hidden units and 61 output units. The errors shown in the figure are the average errors on the four cuts of the database. The abscissa represents one of each of the conditions shown in Table 4.1. The range of word errors obtained from the merging classifiers trained on each of these conditions is shown by a box and whisker plot. The box has lines at the lower quartile (25% of the merging classifiers yield word errors below this line), median (50% of the merging classifiers yield errors below this line) and upper quartile (75% of the merging classifiers yield errors below this line) error values. The whiskers cover the range of errors below the lower quartile and above the upper quartile respectively. A line connecting the median values over all conditions represents the trend in the errors as the number of sub-bands dropped from the merging process is incremented.

The following observations can be made from our experiment -

- There is a gradual decrease in the performance of the merging classifiers with decreasing number of available sub-bands. With as many as three out of seven sub-bands dropped, the upper quartile error is only around twice the error rate obtained when all the seven sub-bands are retained. This gradual decrease in performance indicates that the speech spectrum carries redundant acoustic information.

- It also indicates that even with half the spectrum corrupted, if we can identify and remove the corrupted sub-bands from the merging process, the performance of the system will degrade quite gradually.

- We also observed that for a specific number of sub-bands dropped, leaving out adjacent sub-bands generally yields more error than that obtained when the interleaved sub-bands are left out. For example, for merging classifiers with four sub-bands dropped, leaving out the adjacent sub-bands 2,3,4,5 results in a word error of 6.8%, while leaving out interleaved sub-bands 1,3,5,7 yields a word error of only 2.0%.

4.1.1 Comparison of merging classifiers trained at different merging levels

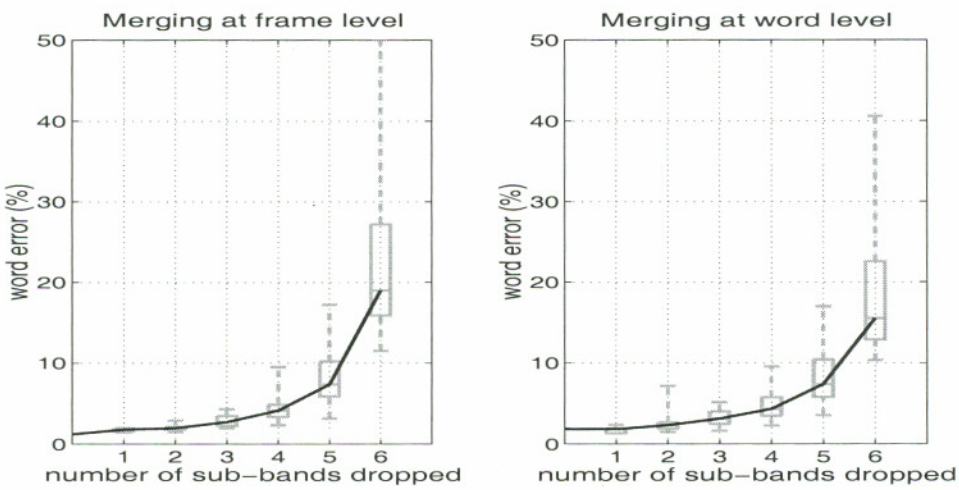


Figure 4.2: Word error (%) with merging networks having 0 - 6 sub-bands excluded and trained for merging at the frame-level and word-level respectively. The left-hand side graph is the same as Fig. 4.1 and has been repeated for easy comparison.

Fig. 4.2 compares the range of errors of the merging classifiers trained for merging at

the frame-level, to those that have been trained for word-level merging. The median error rates for the frame-level and word-level merging classifiers are comparable.

4.2 Techniques for the Selection of Merging Classifiers

In the previous section, we tested the performance of the multi-band system under various band-limited speech conditions. However, in order to practically use the system, we need to know which sub-bands are corrupted. In a practical scenario, with unknown noise corrupting the speech signal, we need to have some estimate of the sub-band reliability. We examined some techniques for estimating reliability of the sub-bands. These techniques can be broadly divided in two categories. The first category consists of signal-based estimation techniques, for e.g., signal-to-noise ratio (SNR) estimation. The second category consists of selection strategies based on the merging classifier outputs, for e.g., majority voting, adaptation and classifier confidence estimation.

4.2.1 Signal-based technique — SNR estimation

One approach to determine the reliability of the sub-bands is to use signal processing techniques to estimate the noise level in each sub-band signal and then compute the signal-to-noise (SNR) ratio in each sub-band. Based on some empirically determined threshold, sub-bands with SNR below a threshold are considered less reliable than the other sub-bands. These sub-bands can then be dropped from the merging process by using the merging classifier which is not trained on these sub-bands.

We estimate the noise level in each frequency sub-band, using the estimation technique proposed by Hirsch (Hirsch, 1993). The noise level in each frequency sub-band is estimated by computing a histogram of the magnitude spectral values in that sub-band. The spectral value at which the smoothed histogram displays a peak is the noise amplitude estimate ($\sqrt{N}$). Since, we do not know the power of the clean speech signal, the power of the available noisy signal in the frequency band is used as an estimate of the noisy signal ($S+N$). The sub-band SNR is then estimated as,

$$\mathrm{SNR} = 10\log_{10}\left\{\frac{(S+N)}{N} - 1\right\}. \tag{4.1}$$

4.2.2 Classifier-based techniques

These techniques estimate the merging classifier reliability based on the classifier outputs, instead of the signal as used in the case of SNR estimation technique.

Majority voting

A simple approach uses majority vote among the 127 merging classifiers. Let L represent the number of merging classifiers, Y_i represents the sub-bands which are merged by the i^{th} merging classifier, for e.g., if the i^{th} merging classifier estimates the joint class-conditional likelihood of sub-bands 1,3 and 5, then Y_i represents the inputs $\{x_1, x_3, x_5\}$ in the respective sub-bands. The merging classifier will then estimate the likelihood $P(Y_i|c_k)$ for the k^{th} class. The majority vote is computed as

$$\Delta_{ik} = \begin{cases} 1 & \text{if } P(Y_i|c_k) = \max_j P(Y_i|c_j) \text{ for } j = 1...M \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

$$c_k = \max_j \sum_{i=1}^{L} \Delta_{ij}. \tag{4.3}$$

Direct use of majority voting among all the merging classifiers gives equal weighting to all the 127 merging classifiers. For band-limited noise condition, some of the merging classifiers are likely to be more reliable than the others, depending on which sub-bands are being merged by the particular merging classifier. Selection of the merging classifiers for majority voting based on some reliability criterion, could hence yield better performance than direct majority voting among all the merging classifiers.

Adaptation

An approach to selecting a merging classifier can be based on the availability of a few known utterances from the noisy test environment, if the application permits. In such a case, the classifier which gives the best recognition performance on these utterances can be used for all further tests in that particular test environment. In cases where more than one merging classifier yields the same performance on the adaptation utterances, a majority voting among them can be used for further classification.

Classifier confidence estimation

Selection of the merging classifiers can also be based on an estimation of the confidence of the merging classifier output. We examined the distributions of the log-likelihoods estimated by the 127 merging classifiers. For every utterance, each merging classifier estimates the log-likelihood (or probability) for each of the competing classes (13 words in our experiment), which constitute the output vector. For all utterances in the test set, we separated the output vectors obtained from all the merging classifiers into two sets — 1) output vectors from the merging classifiers which made the right decision and 2) remaining output vectors (i.e., from merging classifiers which made the wrong decision).

We further divided each set into two more sets of likelihoods - 1) the likelihood of the winning class and 2) the likelihoods of the remaining 12 classes. A plot of the distributions of these two sets of likelihoods, then indicates the extent to which the distribution of the winning class likelihood is separated from the distribution of likelihoods of the other classes.

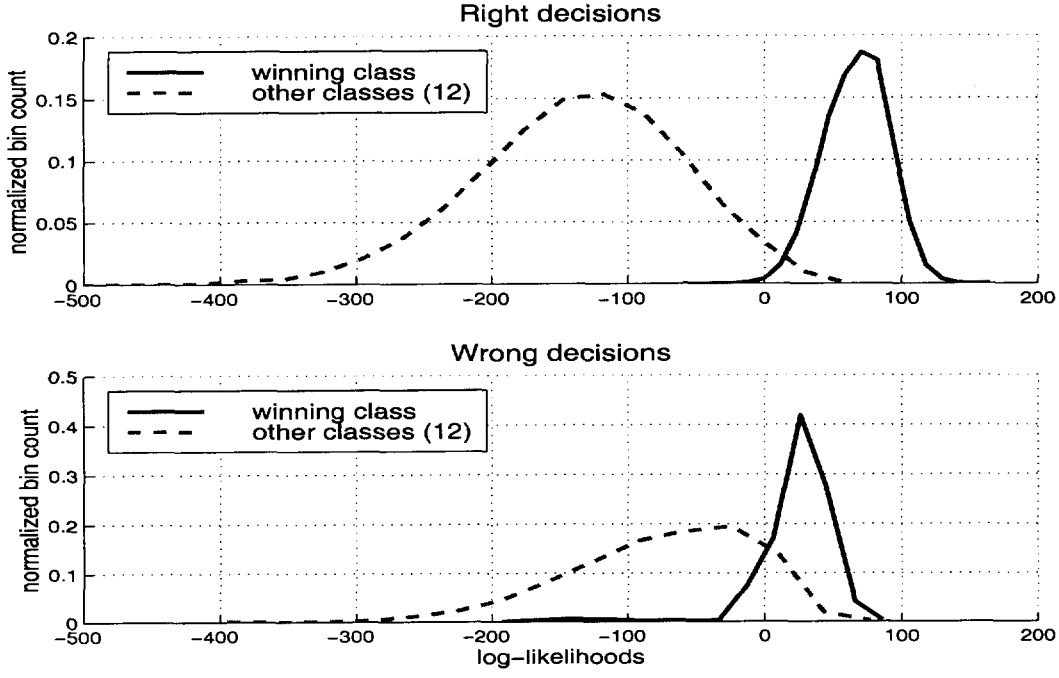


Figure 4.3: Distributions of the outputs (log-likelihoods) of the merging networks for the right decisions (top panel) and wrong decisions (bottom panel).

The plots in Fig 4.3 show that the distributions of the right decisions are better separated than those of the wrong decisions. A measure of this separation can be used as a classifier confidence estimate. We examined two such measures.

The first confidence measure is a sensitivity measure used in signal detection theory (Green and John, 1974) to measure the difference between the means of the signal and noise distributions. It is given by

$$d' = \frac{\mu_s - \mu_n}{\sigma_n}, \tag{4.4}$$

where μ_s is the mean of the signal distribution, and μ_n and σ_n are the mean and standard deviation respectively of the noise distribution, which is assumed to be Gaussian. A higher value of d' (referred to as d-prime) indicates that the distributions have greater separability and hence implies the higher discriminability of the signal from noise. For our purpose, we consider the likelihood of the winning class as the signal and the likelihoods of the other classes as noise. Thus, we replace μ_s by the likelihood of the winning class, and μ_n and σ_n by the estimate of the mean and standard deviation of the other class' likelihoods.

The second confidence measure is based on taking a simple difference between the top two log-likelihoods in the output vector. The assumption is that a greater difference in the log-likelihoods of the winning class and the next best guess implies higher classifier confidence.

Fig. 4.4 compares the distributions of the confidence estimates for the merging classifiers which make the right decisions to those that make a wrong decision, using the d-prime based confidence measure and difference based measure. The distributions show that, as desired, the classifiers making the right decision have higher confidence than those that make the wrong decision. Based on these distributions, the top 10% to 15% of the merging classifiers, having the highest confidence are selected, for majority voting.

4.3 Controlled Experiment with Frequency-localized Sinusoidal Additive Noise

To test the effectiveness of the sub-band model in narrowband noise conditions, we performed a controlled experiment by adding sinusoidal noise at 1500 Hz to the original test

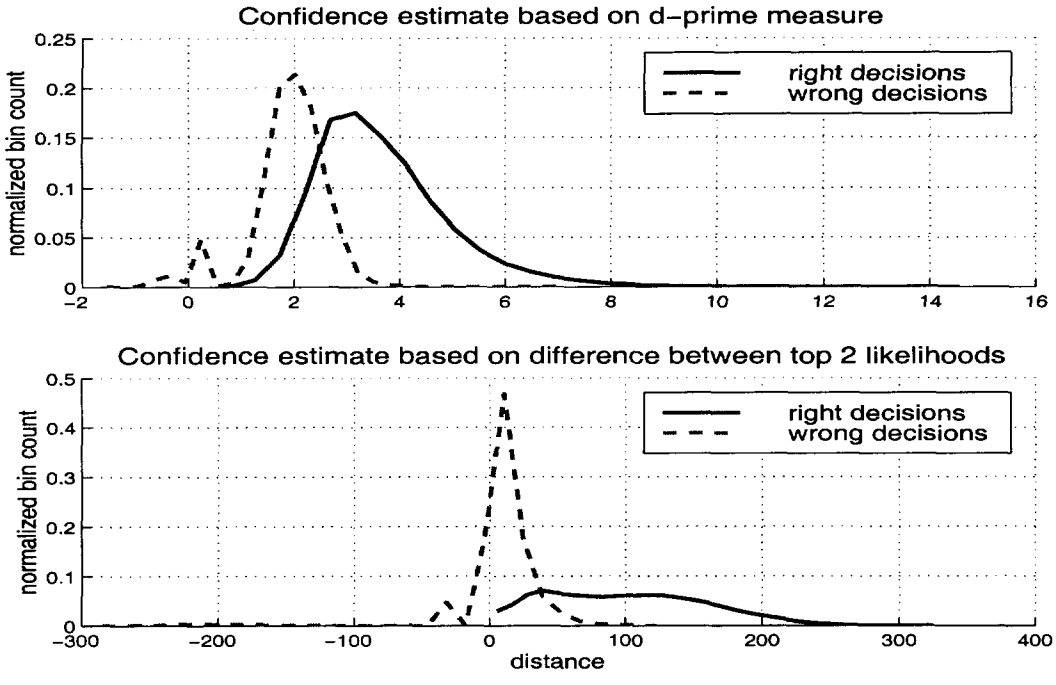


Figure 4.4: Distribution of the confidence estimates of the merging classifiers based on the d-prime measure and difference of top 2 outputs (log-likelihoods) for the right and wrong decisions.

set at different SNRs (30dB, 20dB, 10dB and 0dB).

Fig. 4.5 shows the effect of the noise on sub-band performance at different SNRs. It is seen that the noise affects sub-bands 4, 5 and 6, as these bands integrate the energy at 1500 Hz to different extents. As the SNR decreases, the recognition performance in these sub-bands decreases. The performance of the remaining sub-bands (1,2,3,7) is unaffected by the sinusoidal noise and their performance remains similar to that on the uncorrupted (clean) speech.

The outputs of all the seven sub-band recognizers were then merged at the frame-level using the 127 merging classifiers trained on the clean training data, as described in Section 4.1. We also tested the effectiveness of the techniques for selecting the reliable merging classifiers. For the confidence estimate based selection techniques, the top 15 (12%) merging networks having highest confidence were used in the majority voting. For the SNR estimation based selection technique, sub-bands having SNR estimate below 10dB for any of the constituting critical bands were dropped from the merging process.

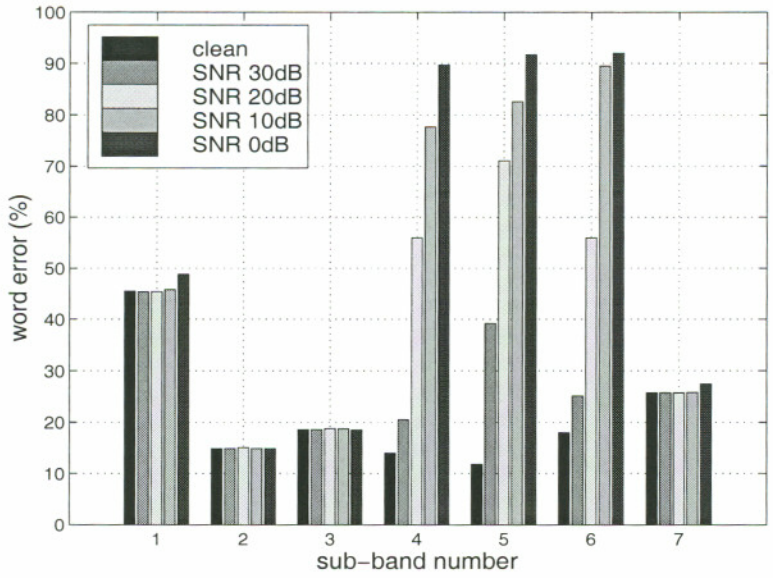


Figure 4.5: Sub-band errors (%) on speech corrupted with additive sinusoidal noise at 1500 Hz. at different SNRs.

For the adaptation based selection technique, 10 utterances from a single speaker were used as adaptation data.

The results of the experiment are shown in Fig. 4.6. The bar labeled *Best* on the plot represents a *cheating* result, obtained by manually picking the merging classifier which gives the best performance on the test data. This represents the performance that could be achieved on each of the test conditions in case we could pre-determine the most reliable sub-bands for the particular test environment.

The following observations can be made from the bar plot —

- The performance of the conventional recognizer (represented by the first bar at each noise level) deteriorates rapidly from 2.5% word error to 62% word error as the SNR decreases from the clean case to 0dB.

- The sub-band system which uses all seven sub-bands (second bar at each noise level) shows less degradation in performance as compared to the conventional system. Similar observations have been reported in (Bourlard and Dupont, 1997) for band-limited noise on a continuous numbers task. At higher SNRs (up to 10dB), it yields significant reduction in word error, almost halving the errors for up to 20dB SNR.

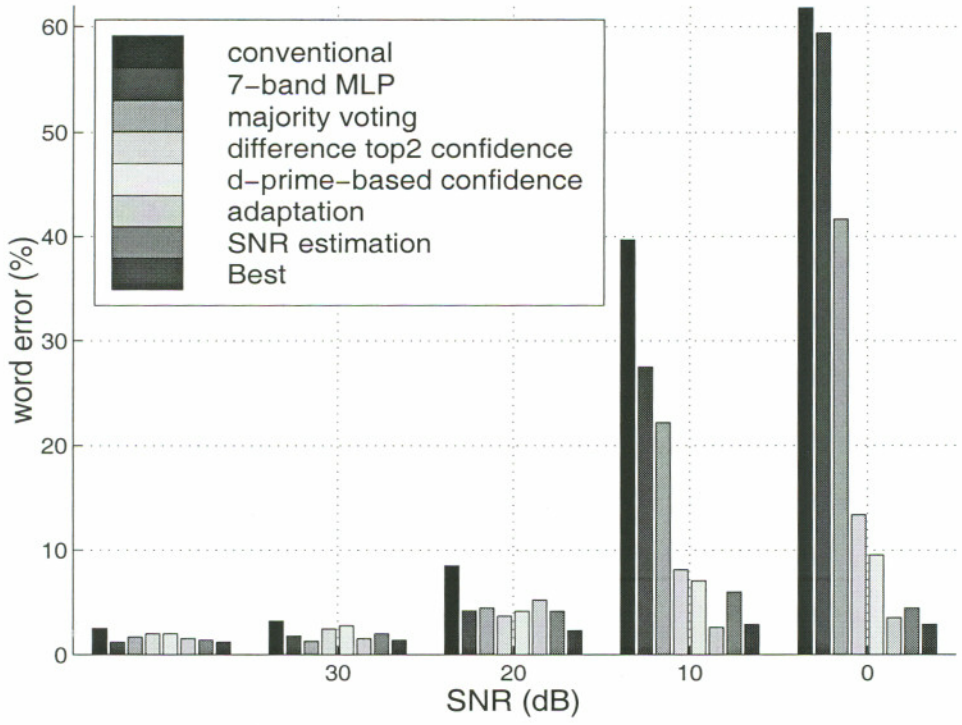


Figure 4.6: Performance of the sub-band system using different selection techniques, on clean speech and speech corrupted with additive sinusoidal noise at 1500Hz and SNRs 30dB, 20dB, 10dB and 0dB. For each SNR level the bars from left to right are in the same order as referred to in the legend.

However, at low SNRs (e.g., 0dB) the sub-band system is not very efficient, yielding only marginal improvement in performance as compared to the conventional system.

- Majority voting using all 127 merging classifiers (third bar at each noise level) results in a further reduction in word error rate, especially at lower SNRs (e.g., 10dB and 0dB).

- The eighth bar at each noise level (labeled *Best*) indicates that there exists at least one sub-band combination (among the 127 combinations) which yields good performance, even in presence of significant degradation by band-limited noise. It shows that the error rate for the noisy speech could improve dramatically (more than an order of magnitude for SNRs 10dB and 0dB). Performance close to this can be achieved using some selection criteria (bars 4 to 7 at each noise level).

- Comparing the two classifier confidence estimation techniques, it is seen that selection of the merging classifiers using the confidence estimate based on the d-prime measure (fifth bar at each noise level) is more robust as compared to the selection which uses the confidence estimate based on the difference of the top two classifier outputs (corresponding fourth bar at each noise level).

- Selection of the merging classifiers based on adaptation (bar 6 at each noise level) yielded performance close to that of the *Best* system. This technique can be used with advantage if the application allows for any adaptation to a given environment.

- Selection of a merging classifier based on SNR-based thresholding (bar 7 at each noise level) also yields recognition performance close to that of the *Best* system.

In conclusion, all selection techniques yield results which are significantly better than the results of the conventional recognizer.

4.4 Experiments with Real Additive Noise

The previous section demonstrated that the sub-band model is effective in the presence of narrowband degradations. The next question is — does noise encountered in real environments exhibit such band-limited characteristics. To answer this question, we examined some of the noise samples (babble, factory2, pink, volvo, white and high frequency radio channel noise) from the NOISEX-92 database (Varga and Steeneken, 1993), as well as a car noise locally recorded in a Volvo with closed windows (referred to as closed-volvo noise). Fig. 4.7 shows the average critical band log power spectra for each of the noise conditions. The figure shows that for the first 5 noise cases - i.e., babble, closed-volvo, factory2, pink and volvo noises, the noise spectral energy appears to be concentrated in some spectral regions. For example, as an extreme case the volvo-noise is mainly concentrated in the low-frequency spectral regions. These noise types thus represent the noise conditions which corrupt some frequency sub-bands more than the others. On the other hand, for the white noise and high-frequency radio channel noise, the noise energy appears to be distributed across all spectral channels.

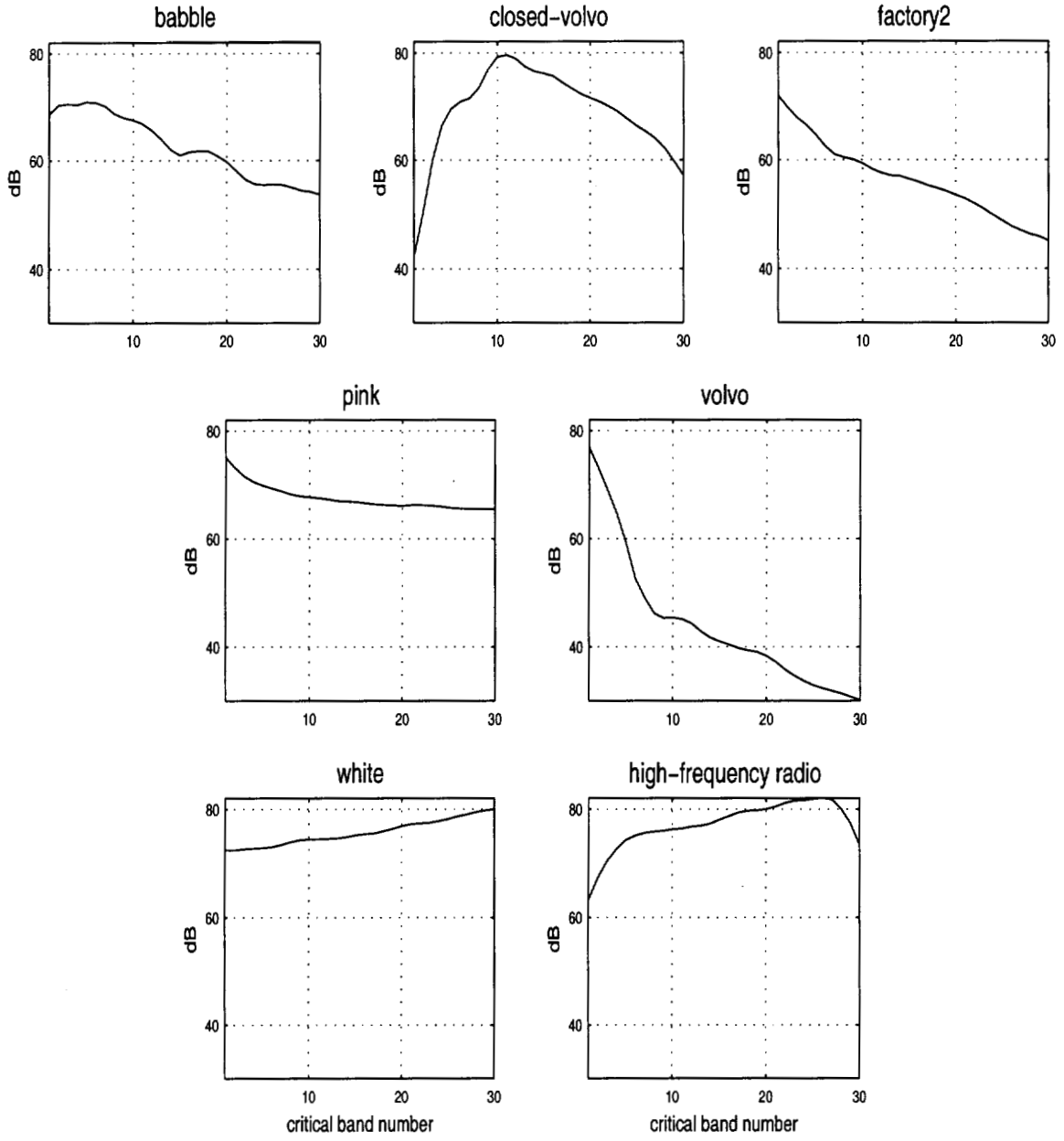


Figure 4.7: Average log critical band noise spectra for the six noise samples from the NOISEX-92 database and the noise spectrum of a locally recorded Volvo noise (closed-volvo). Critical band frequency ranges are as given in Table 3.2.

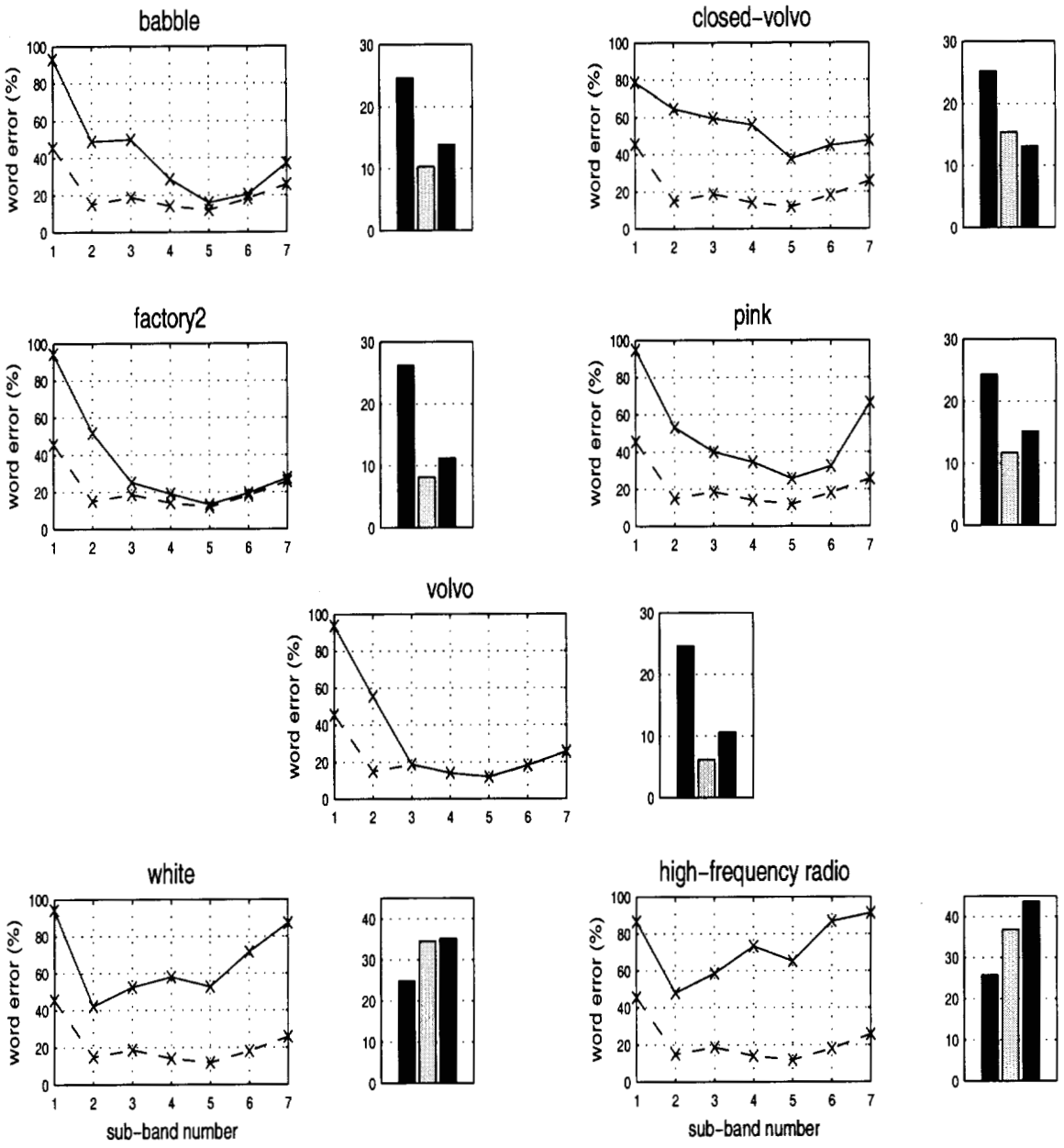


Figure 4.8: Sub-band word errors (%) for each of the noise conditions. The dashed line corresponds to the sub-band errors for the clean speech and the solid line represents the sub-band errors for speech corrupted by the corresponding noise. The bar plot next to each sub-band error plot compares the word error of the conventional system (leftmost bar), the 7 sub-band system which uses frame-level merging of the sub-band outputs (middle bar) and that which uses word-level merging (rightmost bar).

These seven noise samples were then added to test speech data (from the first cut, Section 3.5) after being scaled, so that the performance of the conventional full-band ASR system noticeably degraded from the baseline error of about 2.5% to approximately 25% for each of the noise conditions. The resulting SNR was approximately 15dB for the first five noise cases and around 5 dB for the white and high frequency radio channel noise. Fig. 4.8 compares the sub-band errors for each of the noise corrupted speech cases with that of clean speech case. The figure shows that

- For the first 5 noise cases some of the sub-bands degrade more in performance than the other sub-bands. For example, for the factory noise case, sub-bands 4,5,6 and 7 are degraded to a much lesser extent than the remaining sub-bands. For the other two noise examples — i.e., white and high-frequency radio channel noise, all sub-bands show a significant degradation in performance.

- Also shown in the figure in the form of bar plots, is the comparison of the performance of the conventional recognizer (leftmost bar) to that of the sub-band system which merges the outputs of all seven sub-bands either at the frame-level (middle bar) or the word-level (rightmost bar). The sub-band system which has been trained only on clean speech yields a significant reduction in word error rate (as much as 50%) for the first 5 noise cases. As mentioned earlier, these are also the noise types which corrupt some frequency sub-bands more than the other sub-bands and hence are similar to the band-limited noise cases.

- The multi-band approach is seen to be ineffective for the white-noise and high-frequency noise cases where all the sub-bands show significant degradation in performance.

- The bar plots also indicate that the sub-band system which combines the outputs of the 7 sub-bands at the word-level, in general, gives worse performance than the sub-band system which combines the outputs at the frame level.

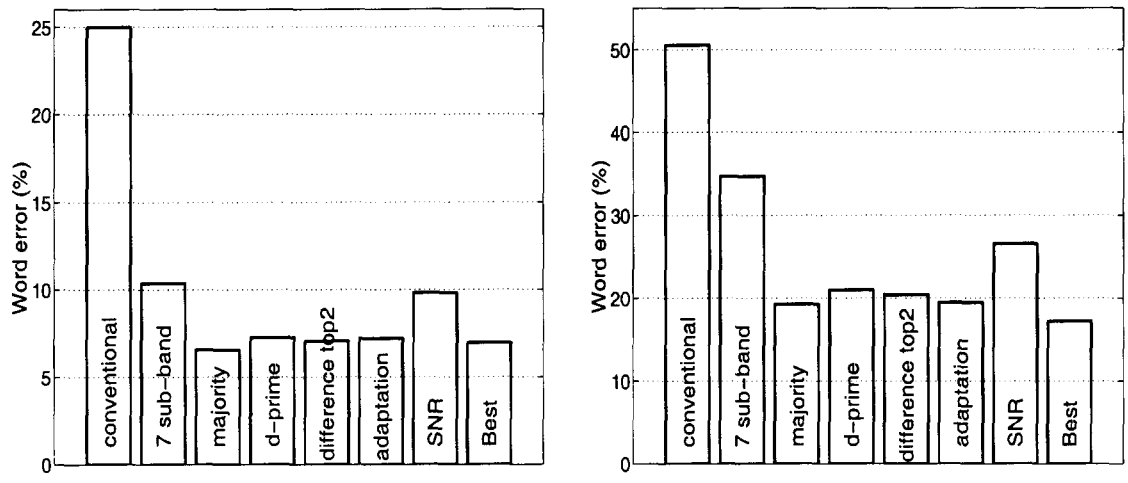


Figure 4.9: Average word error rates (%) for speech corrupted with babble, factory, pink, volvo noises from NOISEX-92 database and Volvo noise recorded locally. Fig.(I) shows the case where the conventional system performance degraded to around 25% error. Fig.(II) shows the case where the conventional system degraded to around 50% error.

4.4.1 Selection strategies on real noise conditions

We evaluated the effectiveness of each of the merging classifier selection strategies for the real noise cases. Fig. 4.9(I) shows the averaged error rates for speech corrupted with the five noise cases - babble, factory2, pink and volvo noise samples from the NOISEX-92 database and another Volvo noise sample collected locally. The sub-band system which combines all 7 sub-bands yields on average about half the error compared to the conventional baseline system. Use of merging classifiers based on the selection strategies give an additional improvement in recognition performance as compared to the sub-band system. The selection strategies which use the two confidence estimates — d-prime based confidence estimate and difference in top 2 classifier output based estimate, give similar performance. The threshold of 5dB was found to yield the best performance for the SNR estimation based selection technique. However, it does not yield as good a performance as the classifier-based selection techniques.

Another point to note is that it is possible for the selection strategies to yield performance better than that obtained by *cheating* (labeled *Best*), since the best combiner is the single merging classifier which yields the best performance on all the test utterances, while the selection strategies dynamically select merging classifiers for each utterance, and hence can yield better performance.

Fig. 4.9(II) shows the averaged error rates for the 5 noise cases when the noise was added at a higher noise level (approximately 5dB SNR), such that the baseline system performance degraded to approximately 50% word error. The sub-band system still yields a significant improvement in performance. However, the improvement due to the selection techniques is more pronounced than seen for the lower noise level (Fig. 4.9(I)).

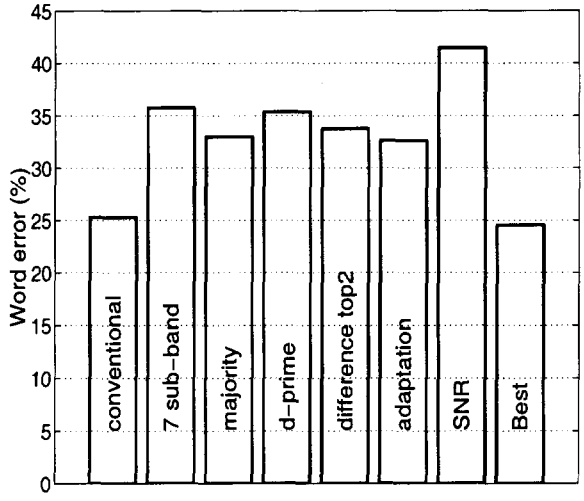


Figure 4.10: Average word error rates (%) for speech corrupted with white and high-frequency radio channel noise.

The selection strategies are ineffective (Fig. 4.10) for white noise and high frequency channel noise, where all the sub-bands were significantly corrupted by noise and hence exhibit significant degradation in performance. For these noise cases the conventional system yields the best performance.

4.5 Summary

In this chapter, we examined the performance of the sub-band model under various band-limited speech conditions. The gradual decrease in recognition performance as the available speech spectrum is reduced, indicates the presence of spectral redundancy in speech. This can be exploited in the multi-band system, by combining different spectral regions using several merging classifiers. These different merging classifiers thus provide a way of obtaining "different views" of the underlying linguistic process. In addition, these classifiers provide an approach to de-emphasizing the unreliable sub-bands by dropping them from the merging process.

We further examined several techniques for the selection of reliable sub-band classifiers or merging classifiers. The effectiveness of the SNR estimate based selection technique depends on the consistency of the SNR estimates as well as the determination of a proper threshold for rejecting the sub-bands. The adaptation approach requires the availability of a few known adaptation utterances in the testing environment. In our experiments, the selection strategy based on classifier confidence estimate performed consistently (performance close to the best merging classifier) for moderate as well as strong noise levels.

We tested the sub-band model in a controlled band-limited sinusoidal noise condition as well as in several realistic noises at various noise levels. For moderate noise levels, the sub-band system itself, without the use of any selection technique, significantly reduces error rates (around 50% reduction in word error as compared to the conventional system). For stronger noises, the selection techniques appear to be more effective. A comparison of the performance of the sub-band system merged at the frame-level with that of the sub-band system merged at the word level showed that frame-level merging yielded better performance in most noise conditions.

The results indicate that the sub-band system is inherently robust to narrow-band additive degradation. Further improvement in performance is obtained by effective de-emphasis of unreliable (noise corrupted) sub-bands.

This chapter and Chapter 3 explored the sub-band model, which treats different regions of the frequency spectrum as independent up to the point of merging. We demonstrated the

feasibility of this approach and showed its efficacy in dealing with narrowband noise. The features used in each sub-band were short-time spectral-based transformations. Hence, the temporal information available to each sub-band classifier at any given time, was in the range of only 10-100ms. Our hypothesis (as discussed in Section 1.5) is that the availability of temporal information of the order of at least syllable-length durations can improve robustness of recognizers. In the next two chapters, we explore the extension of the sub-band model to include medium term (around syllable-length) temporal dependencies in each sub-stream.

Chapter 5

Analysis of Temporal Information — TempoRAl Patterns (TRAPS)

In Chapters 3 and 4, we discussed the sub-band model for ASR. The model comprised of multiple classifiers, where each classifier operated on features derived from different speech frequency sub-bands. This model relaxed the dependency among elements of the feature vector, which in a conventional ASR system are considered dependent.

The frequency localized cepstral features used in each sub-band are short-time spectrum based (10ms to 100ms, due to delta coefficients) features. However, psychoacoustic and physiological evidence (Hermansky, 1998; Greenberg, 1996) suggests that the peripheral human auditory system integrates longer time spans (around 200 ms) of the audio signal. Recent research (discussed in Section 5.1.2) on speech feature extraction has focussed on exploring ways of integrating medium-term temporal information in the short-time features. Such feature extraction techniques yield improved ASR performance, especially in mismatched test conditions. This observation forms the basis of our motivation for the multi-stream concept (as stated in Section 1.5 and shown in Fig. 5.1) of incorporating medium-term temporal dependencies among the short-time feature vectors.

In this chapter, we examine relatively long ($\approx$ 1 sec) temporal trajectories of the speech features to gain an understanding of the nature of phonetic information that is present in time. Based on this study, we propose a set of features which incorporate medium-time temporal dependency in the features used in the multi-band system. The features are the mean TempoRAl Patterns (Mean TRAPs) of spectral energies, and comprise a radically different set of features than those traditionally used.

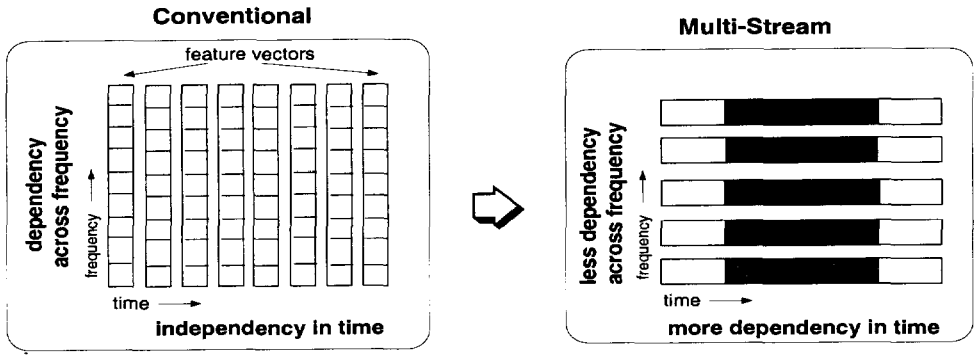


Figure 5.1: Single input multi-stream concept.

In Section 5.1 we review the evidence from the psychophysical and recognition studies, which support the use of medium-time temporal information in ASR systems. Section 5.2 describes our approach towards analyzing the temporal information. This analysis yields the patterns of temporal evolution (Mean TRAPs) of the phonemes. Section 5.3 describes the clustering of the Mean TRAPs to yield the Broad TRAPs, which represent the temporal evolution of the broad phonetic categories. Section 5.4 summarizes this chapter.

5.1 Introduction

5.1.1 Spectral processing

In ASR systems, the speech signal is processed as a series of independent short-time (e.g., 10ms) frames to account for the non-stationary characteristic of the signal, and to facilitate application of the well-developed processing techniques for stationary signals. Spectral features (such as filter bank energies, linear predictive coefficients (LPC), cepstral coefficients, Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and perceptual linear predictive (PLP) cepstral coefficients (Hermansky, 1990)) extracted from short-time frames, form the basis of most feature extraction methods in current ASR. These spectral features describe the spectral envelope of the speech signal in a given frame. The spectral envelope has the drawback that it is quite sensitive to changes in the communication environment, such as characteristics of different channels or background noise. As a result, recognizers which use these features, rapidly degrade in performance

in realistic communication environments.

5.1.2 Towards temporal processing

Psychoacoustic and physiological evidence

Psychoacoustic studies (reviewed in (Hermansky, 1998)) suggest that the peripheral auditory system in humans integrates information from much larger time spans than the temporal duration of the frame used in speech analysis. This time span is of the order of several hundred milliseconds (around 200ms). One such evidence in auditory perception is that of forward masking (discussed in detail in (Moore, 1989)). This is a phenomenon by which, if one signal (masker) is followed by a test signal (probe) closely in time, the perception of the probe is diminished. The effect of forward masking lasts for about 200ms.

It is known that speech is not produced as a string of well-formed sounds. Rather, the production of a speech sound is grossly affected by the preceding and following sounds, due to the mechanical constraints of the articulators. This speech production mechanism is referred to as coarticulation. Coarticulation suggests the temporal dependence of the short-time speech frames even across phone boundaries.

Temporal processing in ASR

Several feature extraction techniques also attempt to incorporate medium-time temporal information within the short-time based ASR framework (refer to (Hermansky, 1995) and (Hermansky, 1998) for a detailed review). These techniques have had some success in improving the noise robustness of ASR.

The dynamic cepstral coefficients (Furui, 1981) are widely used as features in most ASR systems. These coefficients are computed as the first and second order orthogonal polynomial expansions of feature time trajectories and are referred to as delta and acceleration coefficients, respectively. They represent the slope and curvature, respectively, of the feature trajectories and are typically computed over 50ms to 90ms speech segments. These coefficients are robust to slowly varying convolutive distortions introduced by communication channel variations, and are used in both speech recognition and speaker verification tasks.

Another technique that uses temporal processing is the cepstral mean normalization technique, in which the long-term average is subtracted from the logarithmic speech spectrum. This technique thus suppresses convolutive noise introduced due to the communication channel. However it requires computation of long-term averages, which affects a real-time implementation. To minimize the processing delay, the mean from the past data can be updated recursively (Tibrewala and Hermansky, 1997).

The RelAtive SpecTrAl (RASTA) technique (Hermansky and Morgan, 1994) overcomes the long-term average requirement for suppressing the slowly varying convolutive distortions, and in addition, also attenuates the spectral components that vary more rapidly than the typical rate of change of speech (between 1-16 Hz). It employs an infinite-impulse response bandpass filter on the logarithmic spectral time trajectories, with a passband between 1 to 12 Hz (at a frame rate of 10ms, the sampling frequency of the spectral time-trajectories is 100Hz). The impulse response of the optimal RASTA filter has an effective length of around 200ms.

Recent efforts towards exploring the temporal domain

The RASTA filter design was ad hoc as it was optimized on a small series of ASR experiments with noisy telephone digits. To address the optimization issue, RASTA-like filters were designed from a labeled database using a constrained optimization technique (Avendano et al., 1996), and also using the linear discriminant analysis (LDA) technique (van Vuuren and Hermansky, 1997). The criterion for the LDA technique was to optimize the linear discriminability among phonetic classes by suppressing undesirable within-class variability. The three most significant FIR filters (weighting functions) obtained using LDA suggest the use of features derived from at least 250ms long speech segments, for optimal linear discrimination between context independent phonemes.

Other preliminary efforts to explore alternate techniques for deriving medium to long-time dependent features include the RASTA HMM features (Luettin et al., 1997). The design of these features aimed to address the limitations imposed due to the inherent assumptions in the design of the RASTA-like features derived using LDA — i.e., the assumptions of equal within-class covariances, and the use of non-class specific filters.

The approach chosen was to design class-dependent RASTA filters by using LDA.

A study was recently conducted on how the information about the linguistic process is distributed in time. This study used the joint mutual information between the phonetic labels of a labeled database and the logarithmic energies at points in the time-frequency plane, as a criterion (Yang et al., 1999). The results indicate that at a particular frequency, components around 100ms outside the phonetic labeled segment still carry information relevant to the classification of the given phoneme.

The techniques discussed above support the notion that information about a sound (e.g., phoneme) is available not just at its time of occurrence, but rather, it is spread over relatively long (at least syllable-length) duration around it.

5.2 Our Approach to Analyzing the Nature of Temporal Information

Phonemes are the one of the basic sound units used in ASR. To understand the nature of the linguistic information available in the temporal structure of speech, we analyzed the temporal structure of phonemes. In our analysis, we have used logarithmic critical band energies. This domain has previously been used to incorporate temporal information (see Section 5.1.2).

For our analysis, we used the OGI-Stories corpus, described in Section 2.2. We used a subset of approximately three hours of phonetically hand labeled data. We considered the 45 most frequently occurring phonetic classes.

The analysis is carried out on the temporal trajectories of logarithmic critical band energies as shown in Fig. 5.2. We extract all segments from a particular critical band that are labeled as the class (phoneme) under consideration. From each such segment, we form approximately 1 sec long (101 frames at 10ms frame rate) vectors of logarithmic critical band energies, centered around each frame of the segment labeled as the phoneme. For example, say we are considering the phoneme /ah/. If a particular segment labeled /ah/ has five frames, we form five, 1 sec long vectors, each containing one of the five frames of /ah/ in the center. The mean of all such 1 sec vectors, then represents the pattern

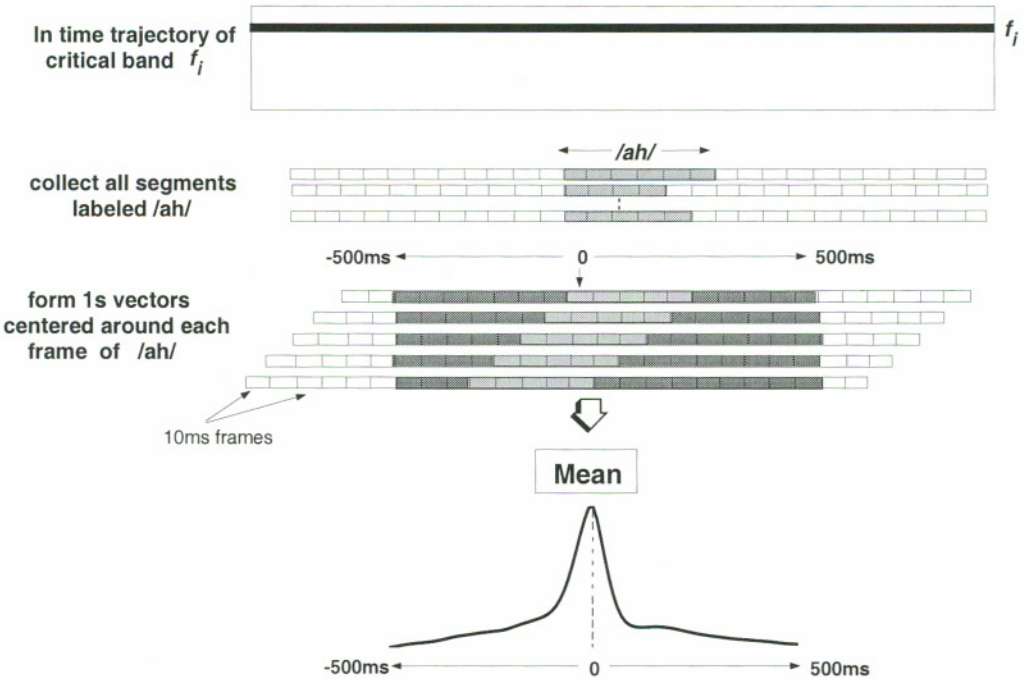


Figure 5.2: Computation of the temporal evolution of phoneme */ah/* for critical band f_i from a labeled database.

corresponding to the average temporal evolution of the phoneme under consideration, at that particular frequency.

In this analysis, only the center frames in all these vectors belong to the same class. Other frames can belong to any other class that occurs in context of the center phoneme, in conversational speech. The mean operation averages over all the surrounding context of the phoneme under consideration. The value of the temporal pattern at the center (i.e., at time $t = 0$ ms, in Fig. 5.2) corresponds to the average spectral value of the considered phoneme at that frequency. The shape of the pattern, away from the center, represents the average coarticulation of that phoneme with all other sounds in whose context it appears.

This process of extraction of the temporal evolution patterns for different phonemes is repeated independently for each of the 15 critical band time trajectories. As a result, 15 temporal patterns are obtained for each phoneme.

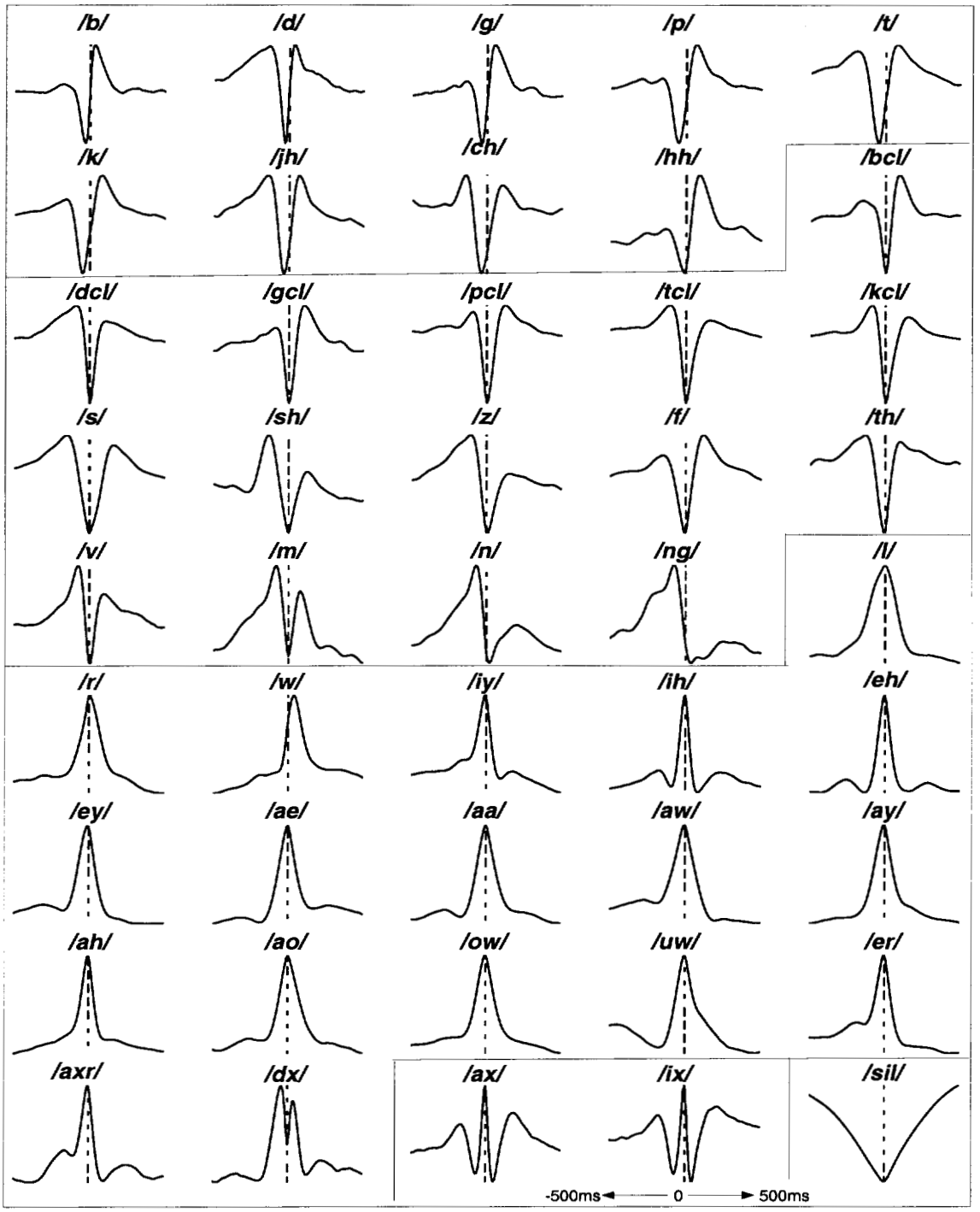


Figure 5.3: Mean TRAPs for 45 phonemes at 500 Hz (i.e., the fifth critical band). The dotted line for each of the TRAPs represents the center frame (i.e., at time $t = 0$ ms). The patterns separated by the thin solid lines represent sounds with similar temporal patterns.

5.2.1 Temporal evolution of phonemes — Mean TRAPs

Fig. 5.3 shows the TempoRAl Patterns (TRAPs) of the 45 phoneme classes. The patterns are shown for the fifth critical band (centered around 500 Hz). The dotted line in each pattern represents the center frame (i.e., time $t = 0$ ms). Two observations from this figure are:

- Some of the phonemes have distinctly different patterns of temporal evolution, for example, the stop-consonant */b/* compared to the vowel */ah/*.
- The temporal evolution of some phonemes is quite similar (represented by patterns separated by the thin solid line in Fig. 5.3), for example, the vowels */ah/* and */ao/*.

5.2.2 Variation in temporal patterns across frequency

Fig. 5.4 shows the temporal evolution patterns of four phonemes across the 15 critical bands. The temporal patterns for a particular phoneme differ in average energy across the critical bands, for all the phonemes. The patterns for some of the phonemes, for example */s/* and */n/*, also differ considerably across the critical bands. The vector consisting of the center frames from all the critical bands represents the average spectral envelope of the phoneme.

5.2.3 Variance of the temporal patterns

The covariances of the temporal trajectories can also be computed for each phoneme class, in a manner similar to the computation of the mean of the temporal trajectories. Fig 5.5 shows the variance vectors (the diagonal of the covariance matrices) for all the 45 phoneme classes. Irrespective of the class, the center frame has the minimum variance and the variance increases as we move away from the center frame in time. In the computation of the pattern of temporal evolution for a particular class, only the center frame of all the vectors belongs to the same class. Hence, the center frame has minimum variance. Frames other than the center frame can belong to any other class in whose context the center class occurs. Hence, the variance increases away from the center, reflecting the contextual variability.

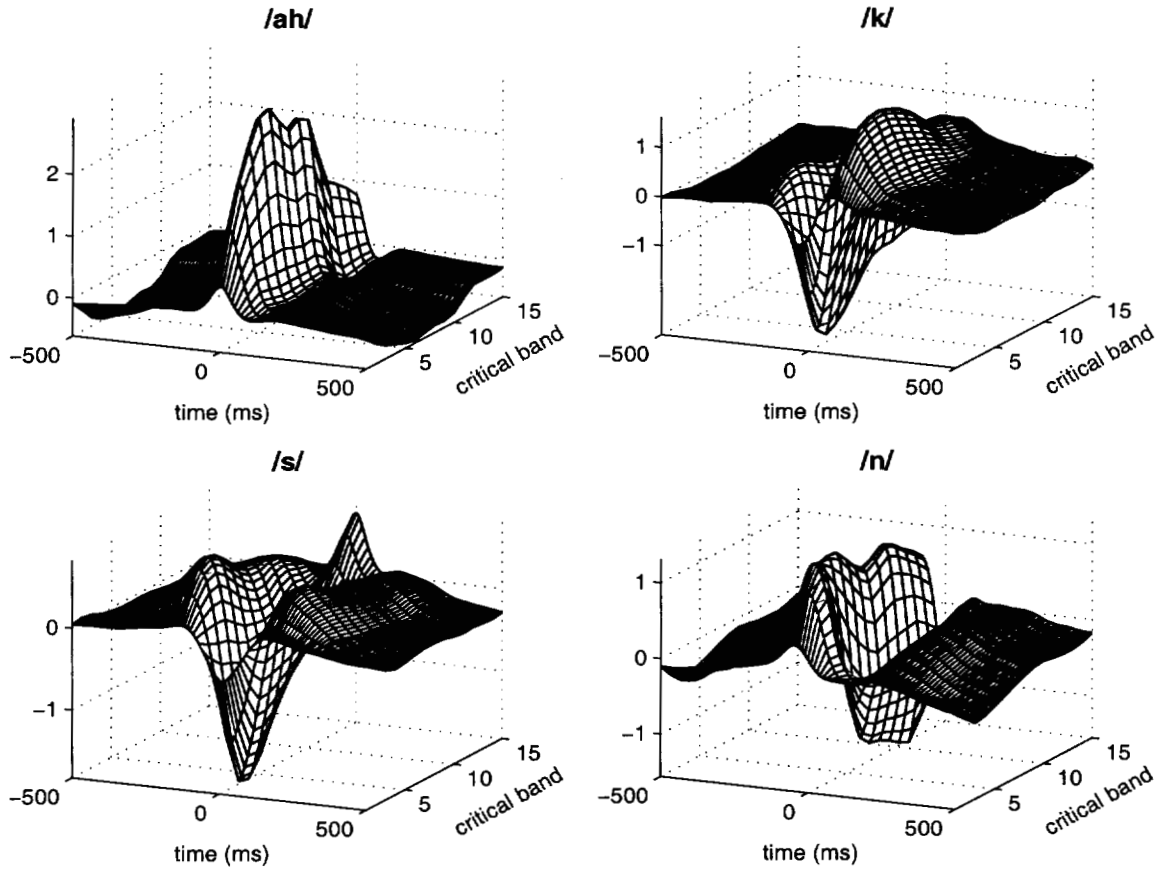


Figure 5.4: Temporal evolution patterns of four phonemes for all the 15 critical bands. Each temporal evolution pattern is normalized to zero mean for convenient display.

5.3 Clustering of Temporal Patterns

As described in Section 5.2.1, some of the classes have very similar patterns of temporal evolutions. Therefore, these patterns can be clustered using a distance measure which takes in to account the similarity. We used the hierarchical clustering algorithm (Arthanari and Dogde, 1981) for clustering the TRAPs, and a correlation based similarity measure, which is given by

$$d(x,y) = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}, \tag{5.1}$$

where $d(x,y)$ represents the distance between the two patterns, x and y, σ_{xy}^2 represents the covariance between x and y, and σ_x and σ_y are the standard deviations of x and y respectively. The similarity measure yields a distance between -1 and 1. A distance of 1

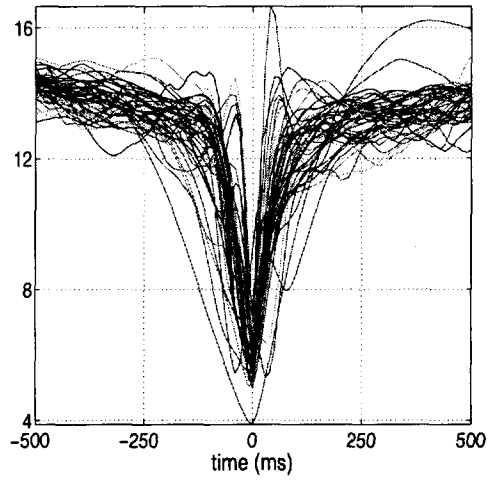


Figure 5.5: Variance of the 1 sec temporal trajectories of the 45 phoneme classes for the fifth critical band.

implies that the TRAPs have the same shape and a distance of -1 implies that the two TRAPs have opposite correlation.

The clustering results in 5 distinct broad-category TRAPs, which we call the Broad TRAPs. Fig 5.6 shows the TRAPs of the phonemes clustered together, along with the 5 Broad TRAPs, for the fifth critical band time trajectory. The Broad TRAP is computed as the weighted mean of the clustered Mean TRAPs. The weighting represents the relative frequency of occurrence of each of the clustered TRAPs in the OGI-Stories database.

The clustering shows that —

- Although no assumptions were made in the clustering algorithm, the TRAPs cluster into the five broad phonetic categories — 1) vowels and diphthongs 2) stop-consonants 3) fricatives 4) schwas (reduced vowels) and 5) silence.

- These Broad TRAPs have distinct and intuitive temporal patterns. For example, the Broad TRAP corresponding to the vowel cluster has a peak in the center since vowels have higher energy than the other sounds. The stop-consonant Broad TRAP has a dip off-center to the left, since a stop-consonant is usually preceded by a closure, which has low energy. In general, the Broad TRAP corresponding to the fricative cluster has low energy at the center where the fricative is located (though this varies for the individual fricatives depending on the frequency band, as mentioned below)

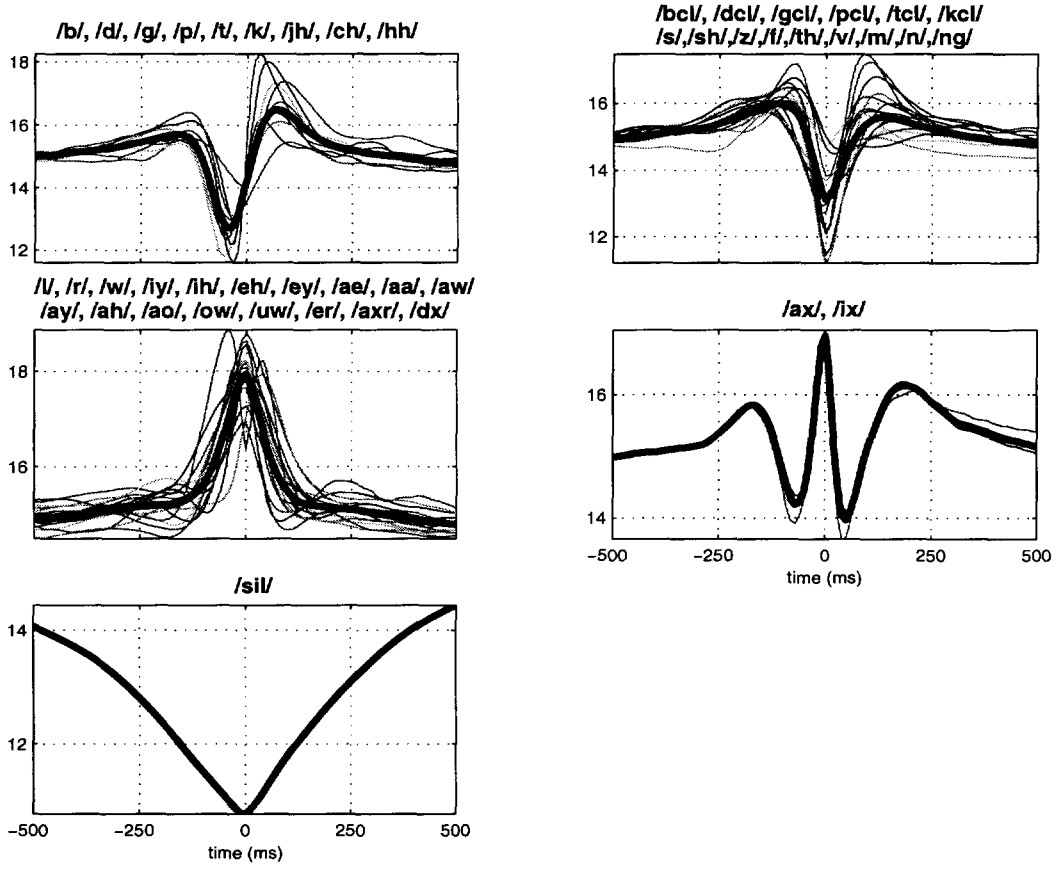


Figure 5.6: Five Broad TRAP clusters of the fifth critical band TRAPs. The thinner lines in each plot represent the individual TRAPs of the phonemes clustered in one category. The thicker line is the Broad TRAP and represents the weighted mean of the constituent phoneme TRAPs.

and the energy increases around it because the surrounding context generally consists of vowels. However, the schwas are characterized by a rather complex temporal evolution pattern.

- Some phonemes change cluster affiliations depending on the frequency bands. For example, the phoneme */s/* clusters as a fricative for frequency bands up to 2 kHz. For higher frequencies, where /s/ has high energy, it displays a vowel-like temporal pattern. The nasal sounds (*/n/*, */m/*, */ng/*) display vowel-like temporal patterns at lower frequencies (up to approximately 500 Hz) and move to the fricative cluster at higher frequencies. The glides (*/l/*, */r/*, */w/*) move from the vowel cluster at lower

frequencies to the stop cluster at higher frequencies.

- Irrespective of the phoneme duration, which varies from approximately 30ms for stop-consonants to 80ms for vowels, the effect of the center phoneme lasts for about syllable-length durations ($\approx$ 200 ms) around it. This result indicates that there is linguistic information beyond the 50-90ms time duration used in the conventional ASR systems.

5.4 Summary

In this chapter we presented a novel technique for analyzing the phonetic information present in the temporal domain. The technique is based on analyzing 1 sec long time trajectories of the critical band energies in the vicinity of the phoneme under consideration. This analysis yields patterns of temporal evolution of the different phonemes, which we call TRAPs. These patterns reveal a syllable-length temporal spread of activity in the region surrounding a phoneme, irrespective of the phoneme duration. This observation supports the hypothesis that the effect of a phoneme lasts for a long duration in time in the region surrounding the phoneme. The TRAPs show distinct characteristics for the broad phonetic categories of sounds — i.e., vowels, stop-consonants, fricatives, schwas and silence. Broad TRAPs, which characterize the broad phonetic categories, are derived based on this observation.

The next step is to integrate the results of this temporal analysis into an ASR system. The next chapter explores this integration, which leads to the generalization of the multi-band system into the TRAP-based classification system.

Chapter 6

TRAP Classifiers in ASR

In the last chapter, we presented a novel approach for analyzing the nature of phonetic information available in the temporal structure of speech. The analysis yielded TempoRAl Patterns (TRAPs) derived from the data. These patterns characterize the temporal evolution of different phonemes in the logarithmic spectral domain. The TRAPs have a syllable-length temporal spread in the region surrounding a phoneme. This supports the notion that information of a phoneme is spread in rather long time durations around it.

The analysis of the temporal structure was motivated by our goal to extend the multi-band model to use temporal information in the individual sub-bands. This chapter discusses the use of TRAPs in the multi-band framework. We examine the use of the temporal patterns (TRAPs) in place of the conventional spectral patterns for ASR in the multi-stream framework. The proposed system is found to yield a significant amount of complementary information to that of the conventional spectral features-based ASR system. A combination of these two ASR systems results in improved robustness to additive and convolutive environmental degradations.

Section 6.1 describes our TRAP-based classification approach. Section 6.2 describes the experimental setup used for the experiments in this chapter. Section 6.3 describes the use of Mean TRAPs, developed in the previous chapter, for ASR . A simple template-matching based approach to phonetic classification using the Mean TRAPs is presented. We demonstrate the presence of considerable amount of phonetic information in each band-limited temporal trajectory of energies. We further analyze the nature of phonetic information available from all the sub-band Mean TRAP classifiers and propose a non-linear integration of these Mean TRAP classifier outputs, in the multi-band framework.

This integration yields high ASR performance. An analysis of the errors of the TRAP-based combined system and the conventional spectral-based system shows availability of complementary information. A combination of these two systems is shown to yield better recognition performance than the conventional system.

In Section 6.4, we present an improvement in the performance of the TRAP-based system, by using a neural classifier (called Neural TRAPs), in place of the template-matching based Mean TRAP classifier. In Section 6.5, we explore the use of broad phonetic categorization using Broad TRAPs, as a first step towards achieving fine phonetic categorization. The results obtained using this simplified system are quite promising. In Section 6.6, we demonstrate the robustness of the TRAP-based ASR system to convolutive as well as additive noise. We conclude this chapter with a summary in Section 6.7.

6.1 Our Approach

The multi-band experiments, discussed in Section 4.1, suggested the presence of spectral redundancy in the critical band filtered speech. This was demonstrated by the relatively gradual degradation in performance, when recognizing speech under various band-limited spectral conditions. These results and the potential of the temporal domain in providing reliable linguistic information, led Hermansky (Hermansky, 1998) to hypothesize that —

> "Even though there is a strong evidence that human auditory perception does some sort of spectral analysis of the incoming acoustic signal, it may be that the main reason for frequency selectivity of human auditory system is not to derive frequency content of a given segment for phonetic classification but rather to provide means for optimal choice of high signal-to-noise (SNR) regions for deriving reliable sub-band based features by temporal analysis of the high SNR sub-bands of the signal."

Our approach investigates this hypothesis. Figure 6.1 represents our approach and contrasts it with the conventional paradigm of ASR. In contrast to doing phonetic classification based on features extracted from the entire frequency spectrum (and possibly augmented with dynamic features), we propose phonetic classification based on features

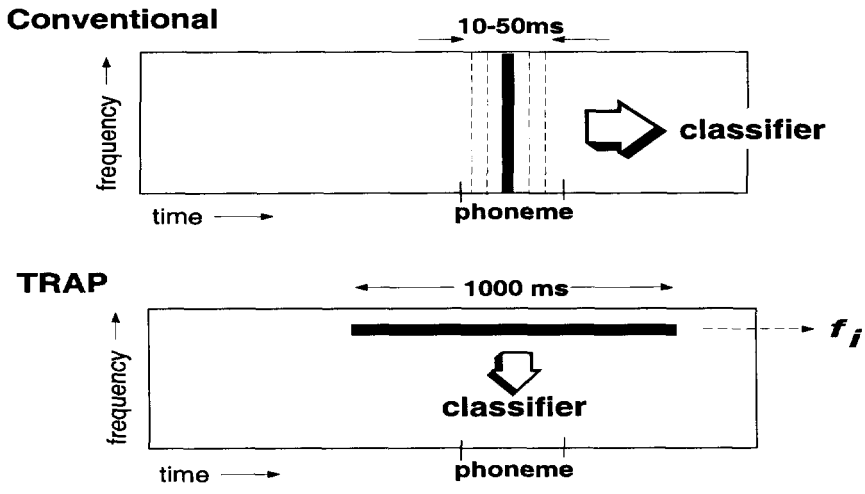


Figure 6.1: Temporal Paradigm for ASR.

extracted from 1 sec long temporal trajectory of logarithmic energies from a single frequency (critical) band. The phonetic class is defined with respect to the center of this temporal trajectory, as was used in the design of RASTA-like filters using LDA (van Vuuren and Hermansky, 1997).

By using temporal vectors for classification, we aim to capture the temporal evolution of the band-limited spectral energy in the vicinity of the underlying phonetic class. Since the classifier based on temporal vectors attempts to capture the appropriate temporal pattern from the acoustic stream, we call such a temporal sub-band classifier the TRAP classifier.

6.2 Experimental Setup

We have used two databases for our work, the OGI Stories corpus and the OGI Numbers corpus (both described in Section 2.2). The OGI Stories database is used for training the temporal (TRAP) classifiers. The OGI Numbers corpus which has a 32 word vocabulary and 29 phonetic classes is used for the recognition.

The baseline system is the hybrid, hidden Markov model/multi-layer perceptron (HMM/MLP) speech recognizer (Bourlard and Morgan, 1994), described in Chapter 2.

The features used for the baseline system consist of 8 PLP cepstral coefficients (Hermansky, 1990) with utterance-based cepstral mean subtraction, along with 9 delta and 9 acceleration coefficients. The input to the MLP consists of 9 frames of context, with the current frame at the center of this context window (234 dimensional input). The hidden layer has 500 units and the output of the MLP represents the estimated posteriori probabilities for the 29 phonetic categories occurring in the Numbers corpus. The baseline system is trained on the 1.7 hours subset of the Numbers corpus. This baseline system yields 21% frame-level error and 6.5% word-level error.

6.3 Phonetic Classification using Mean TRAPs

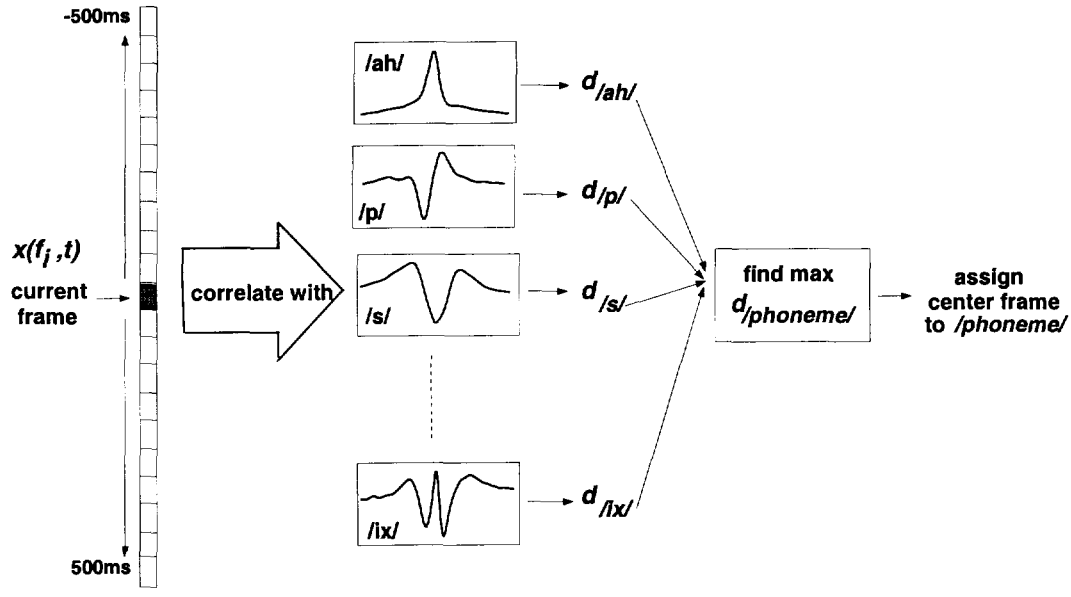


Figure 6.2: Template matching approach to phonetic classification in a frequency band, f_i, using the Mean TRAPs.

The Mean TRAPs of the phonemes in each frequency (critical) sub-band were derived in the last chapter. A simple way to use the TRAPs for phonetic classification is based on the template matching approach with a correlation-based similarity measure. Such a classifier has been referred to as the *correlation classifier* or the *matched filter classifier* in pattern recognition literature (Fukunaga, 1990).

Fig. 6.2 represents the block diagram of the Mean TRAP-based sub-band classifier.

To classify a frame in a particular critical band, a 1 sec time trajectory around that frame is matched to each of the 29 Mean TRAPs for that critical band, using the correlation measure. This measure is the same as used for the clustering of the TRAPs in Section 5.3 and is given by Equation 5.1. The frame is assigned to the phoneme associated with the TRAP which yields the maximum correlation. The availability of 1 sec of data around the current time frame allows for mean removal from each 1 sec vector, to make the features robust. Further, in order to de-emphasize the contributions of the spectral energies towards the edges of the time trajectory, each 1 sec temporal vector is weighted by a Hanning window. Mean removal and window weighting is also carried out on the Mean TRAPs.

Table 6.1: Frame-level performance of the Mean TRAP classifier in each critical band.

SYSTEM	FRAME ERROR
Baseline	21 %
Mean TRAPs	78 - 82 %

Table 6.1 gives the range of the frame errors for the template matching based Mean TRAP classifier (using 29 TRAPS) for each of the 15 critical bands, when tested on the 0.2 hours subset of the Numbers database. It is encouraging that the performance in each critical band is approximately 80% error even for the simple TRAP classifier. This is significantly better than chance (96.5% error for 29 classes), in spite of the fact that none of the TRAPs have access to any information about spectral correlations between neighboring bands.

Table 6.2: Frame errors of the Mean TRAP classifiers as a function of the size of set k, where k corresponds to the set of classes having the top k similarity scores for a given frame.

SYSTEM	FRAME ERROR
Baseline	21 %
Mean TRAPs	
top 1	80 %
top 5	50 %
top 10	28 %

The recognition performance reported in Table 6.1 is based on comparing the class of the current frame with only the class corresponding to the top similarity score. Table 6.2 gives the recognition performance, when we consider the input frame as correctly classified, if its class belongs to a set of classes having the top 5 and top 10 similarity scores. The frame error of the TRAPS in each critical band drops significantly from around 80% error to 50% error, by considering only the top 5 similarity scores, and to 28% when considering the top 10 similarity scores. This indicates that there is considerable amount of discriminative phonetic information even in a single time trajectory.

6.3.1 Similarity scores obtained from the TRAP classifiers

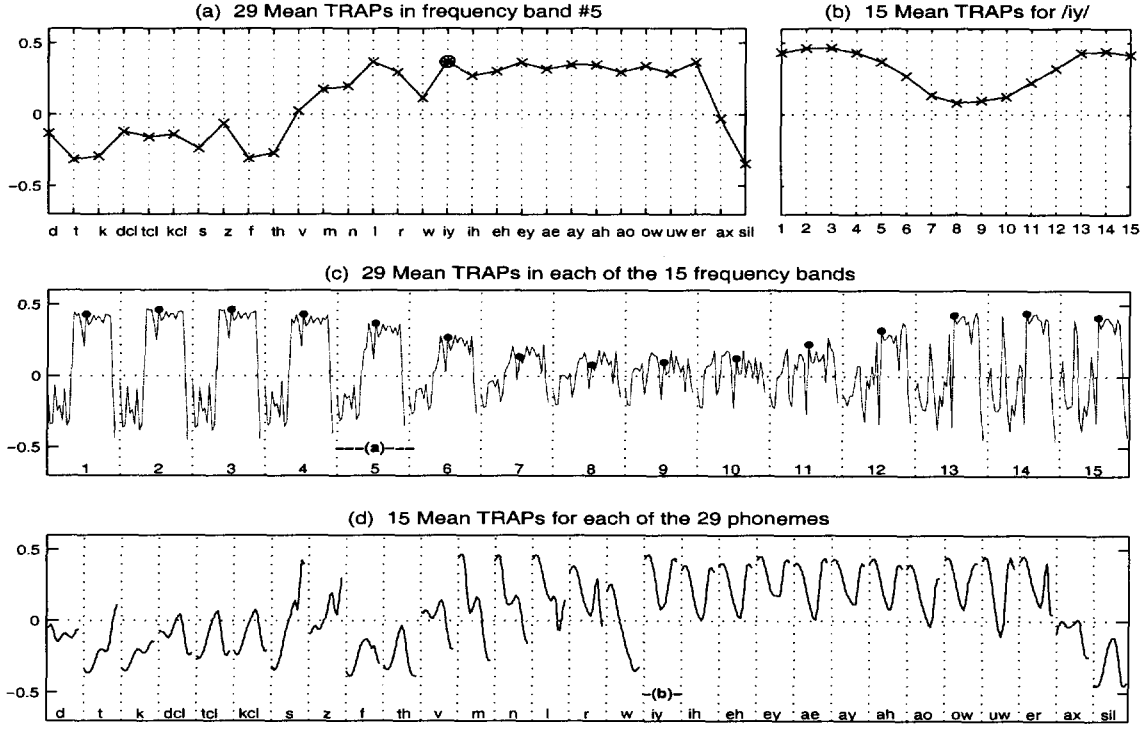


Figure 6.3: Average similarity scores for examples of /iy/ with (a), (b), (c) and (d).

For every input frame, each Mean TRAP classifier outputs a similarity vector. The vector has 29 elements corresponding to the similarity scores obtained from the 29 phoneme Mean TRAPs. Since, there are 15 critical band Mean TRAP classifiers, we obtain 15 output similarity vectors, one from each classifier. Hence, for every input frame, there are

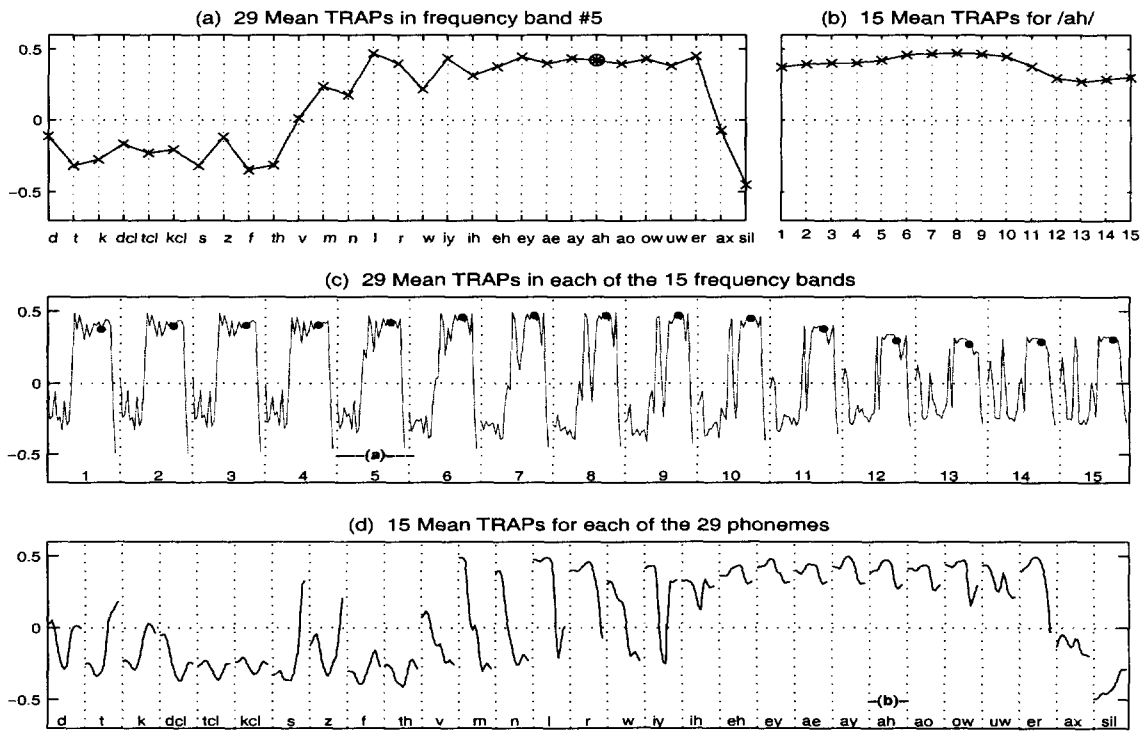


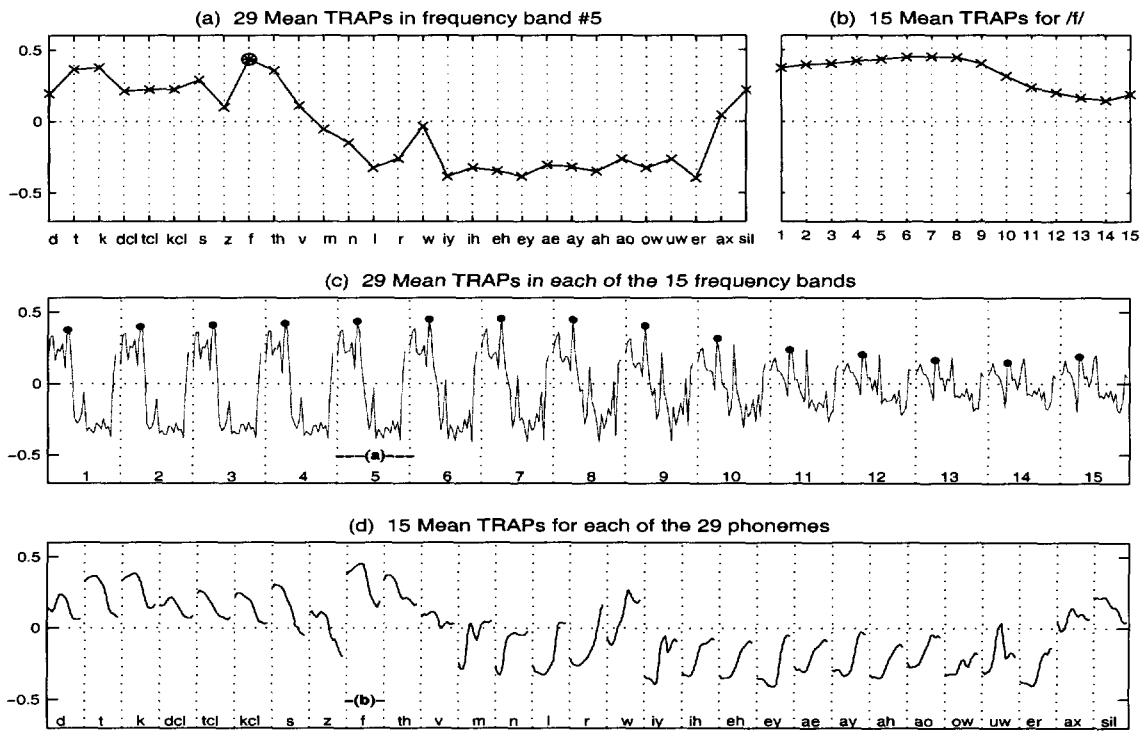
Figure 6.4: Average similarity scores for examples of /ah/ with (a), (b), (c) and (d).

(15 x 29) 435 similarity scores. We examined the average similarity scores obtained for the occurrences of each of the 29 phonemes, on the cross-validation set of the OGI-Numbers database.

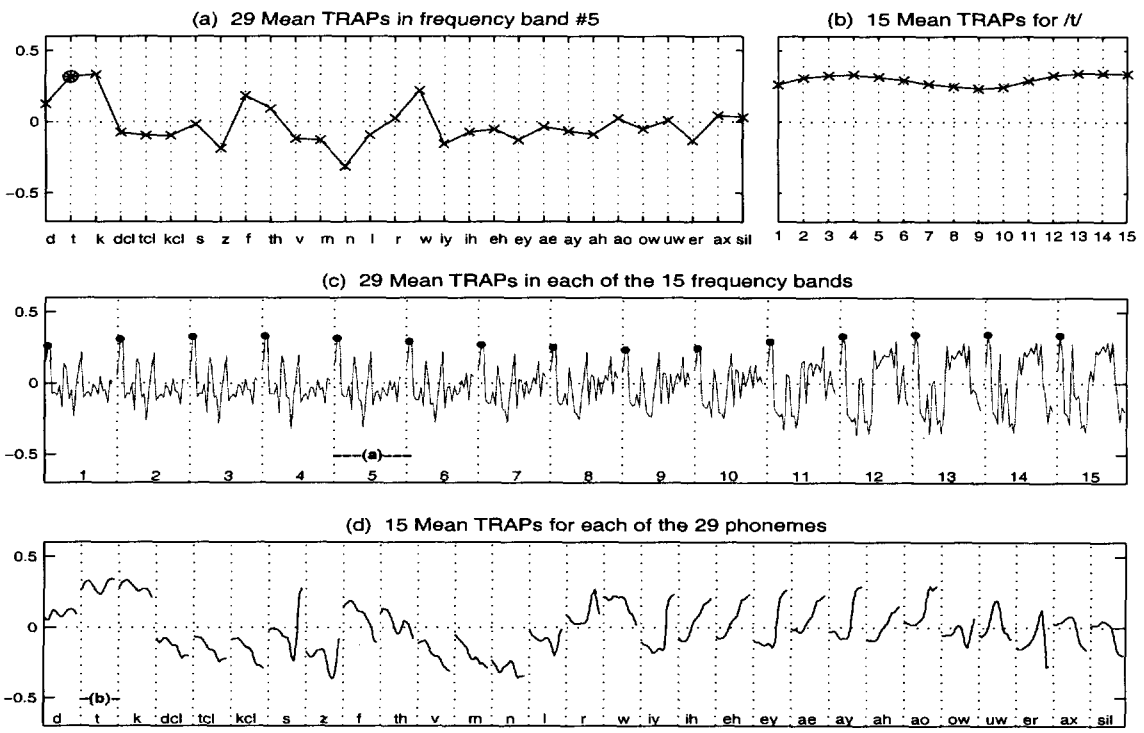
Similarity scores for phoneme /iy/

Fig. 6.3 shows the average similarity scores obtained for the phoneme /iy/. Fig. 6.3(a) shows the 29 similarity scores obtained from the fifth critical band TRAP classifier. The Mean TRAP corresponding to /iy/ produces a high score, as desired (marked by a circle in the figure). However, most of the other vowel Mean TRAPs also yield high similarity scores because of the similarity in their temporal patterns, as noted in Section 5.2.1. The Mean TRAPs corresponding to the consonants yield lower similarity scores for occurrences of /iy/.

Fig. 6.3(b) shows the similarity scores obtained from the 15 critical band Mean TRAPs of /iy/. The similarity scores are higher in the lower and higher frequency regions. Since



(I) Average similarity scores for examples of /f/ with (a), (b), (c) and (d).



(II) Average similarity scores for examples of /t/ with (a), (b), (c) and (d).

Figure 6.5: Average similarity scores obtained for the 2 classes, /f/ and /t/, from the Mean TRAP classifiers.

/iy/ is a front vowel, these frequencies also roughly correspond to the regions of high energy in the /iy/ spectrum. This suggests that the higher spectral SNR regions yield more reliable scores.

Fig. 6.3(c) shows all the 435 similarity scores obtained from the 15 Mean TRAP classifiers. The regions bounded within the vertical dotted lines show the 29 scores obtained in the corresponding critical band (marked along the x-axis). Thus, Fig. 6.3(a) represents a blow-up of the fifth critical band in this figure, as shown. There is considerable variation in the similarity scores across frequency bands, especially if we compare the lower, mid and high frequency similarity scores. The scores corresponding to the mid-frequency regions do not show a systematic confusion pattern as shown by the lower and higher frequency regions (and described above for Fig. 6.3(a)). This indicates that the lower and higher frequency sub-bands should yield more reliable information about the occurrence of /iy/ as compared to the mid-frequency bands.

Fig. 6.3(d) shows the same 435 similarity scores, as shown in Fig. 6.3(c), except that the order in which the scores are displayed has been changed. In this figure, the scores are arranged by the 29 phonetic classes, to explicitly show the distribution of scores of the correct class /iy/, with respect to the distribution of scores of the other 28 classes. The regions bounded by the vertical dotted lines represent the 15 scores obtained from the 15 frequency band Mean TRAPs, corresponding to the phoneme labeled along the x-axis. Fig. 6.3(b) is thus a blow-up of the region labeled /iy/ in this figure, as shown. The vowels tend to have a similar distribution of scores, as noted for Fig. 6.3(a). If we compare Fig.(d) obtained for occurrences of the other phonemes (discussed below), we see that this distribution of scores appears to be distinct for each of the phonemes.

Similarity scores for phoneme /ah/

Fig 6.4 shows the average similarity scores for the phoneme /ah/. In this figure, Fig 6.4(a) is similar to the corresponding Fig 6.3(a) obtained for /iy/. The similarity scores are higher for the vowels as compared to the consonants. Fig 6.4(b) however is different, since /ah/ is a back vowel and is characterized with higher energy at the lower frequencies. The similarity scores appear to follow this characteristic of the spectrum of /ah/. Consequently,

the next figure, Fig 6.4(c), appears to have more reliable information in the lower frequencies while the higher frequency scores show a different confusion pattern. If we compare Fig 6.4(d) with Fig. 6.3(d), we observe that though the vowel similarity scores are high in both these cases, the two phonemes exhibit a different relation among the similarity scores of the 29 phonemes.

Similarity scores for phonemes /f/ and /t/

Figs. 6.5(I) and 6.5(II) represent the average similarity scores obtained for the fricative /f/ and stop-consonant /t/, respectively. The figures show that the vowel TRAPs yield lower similarity scores for these phonemes as compared to the other TRAPs. For the phoneme /f/, the Mean TRAPs corresponding to /f/ appear to have different reliability in different frequency bands, with higher similarity and reliability (fewer confusions) at lower frequencies (Figs. 6.5(I)(b) and 6.5(I)(c)). The stop /t/ attains a high similarity score, irrespective of the frequency band. The similarity scores obtained for these two phonemes also show distinctly different relations with the similarity scores obtained from the other phoneme TRAPs, as is evident from the figures Fig. 6.5(I)(d) and 6.5(II)(d).

Average similarity scores for the other phonemes are given in Appendix B.

6.3.2 Combination of the Mean TRAP classifiers

In the previous section, we examined the 435 similarity scores obtained for the phonemes, on the cross-validation data of the OGI-Numbers database. We observed that the scores have a different distribution across frequency bands and also across the phonemes, depending on which phoneme data are considered. The combination of scores, so as to take into account these different distributions, should yield better performance than that of any single TRAP classifier.

We consider the combination of scores in the multi-stream framework, where each sub-stream corresponds to a critical band Mean TRAP classifier. Each TRAP classifier yields a similarity vector (comprising 29 elements) per input frame. The 15 similarity vectors thus obtained are then combined to obtain the combined frame classification. Fig. 6.6 represents the block diagram of such a system.

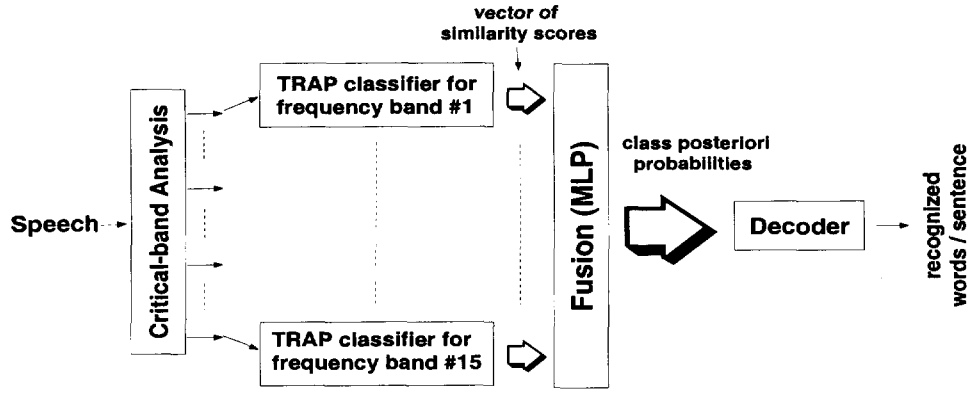


Figure 6.6: Mean TRAP combiner block diagram. Each TRAP classifier has the form shown in Fig. 6.2 up to the decision-making block.

As in the multi-band ASR system, we use a single hidden layer, feed-forward MLP for combining the outputs obtained from each of the 15 Mean TRAP classifiers. The input to the combining network is the concatenated vector of the scores for the 29 phonetic classes from each of the corresponding 15 TRAPs (435 dimensional input). The network has a hidden layer of 300 units and 29 outputs which represent the merged estimate of the class posteriori probabilities. The combination network thus has 139200 parameters which is comparable to the 131500 parameters of the baseline system.

Table 6.3: Performance of the Baseline system and the Mean TRAP-based combined system on the OGI Numbers corpus.

System	Frame error	Word error
Baseline	21 %	6.5 %
Mean TRAP-based	22 %	11.5 %

Table 6.3 compares the frame errors and word errors of the baseline system and the Mean TRAP-based combined system. The frame level performance of the baseline system and the Mean TRAP combined system is quite comparable. On the word level, the simple Mean TRAP-based recognizer yields about twice the error of the baseline system.

Comparison of the frame-level errors of the baseline system to the TRAP-based combined system shows that although the TRAP-based system improves the frame performance on some fricatives and silence as compared to the baseline system, it suffers in its

performance on the vowels. This trend in errors is consistent with the notion that the vowels are better characterized by their spectrum, and hence, a spectral based system should classify them better. Also, since most ASR systems (specifically the decoder) rely heavily on vowel recognition to achieve good word-level accuracy (Rabiner and Juang, 1993), this could account for the lower performance of the Mean TRAP-based system on the word level as compared to the baseline system.

6.3.3 Combination of the Baseline and TRAP-based systems

Table 6.4: Analysis of the frame-level decisions of the Baseline and the Mean TRAP-based combined system on the cross-validation set.

Nature of decisions	BASELINE	MEAN TRAP-BASED	NO OF FRAMES
same decisions	correct	correct	43,833
	wrong	wrong	8,296
complementary errors	correct	wrong	5,509
	wrong	correct	4,844

As noted above, though the baseline system and the TRAP-based combined system yield similar performance on the frame-level, they make errors on different frames. Table 6.4 gives the exact analysis of the decisions made by the two systems on the frame-level. For approximately 16% ((5509+4844)/62482) of the frames, both the systems make complementary errors. Considering only the errors, out of the total errors made by the TRAP-based system (8296+5509 frames), the baseline system gives correct decisions for around 40% (5509 frames) of these frames. Similarly, out of the total errors (8296+4484) made by baseline system, the TRAP-based system gives correct decisions for around 37% (4844 frames) of these frames. This indicates that the two systems make a significant number of complementary errors and hence, yield a significant amount of complementary information.

This indicates that a combination of the outputs of the two systems can yield an improvement in performance (Sharma et al., 1998). We combine the outputs of the baseline system and the TRAP-based system at the frame level (Fig. 6.7) using a simple average

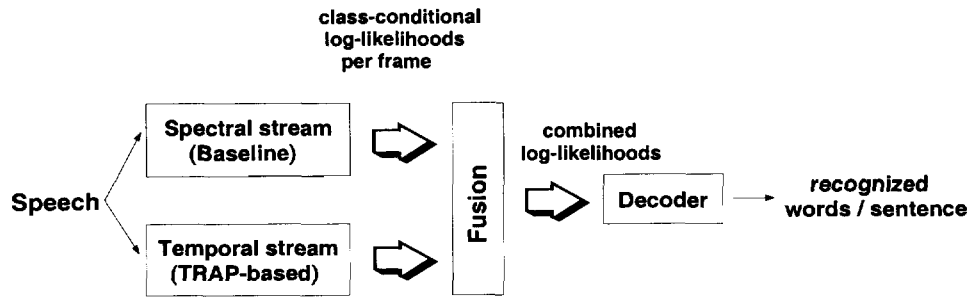


Figure 6.7: Block diagram of combination of baseline (spectral-based) system with TRAP-based (temporal) system.

of the corresponding log-likelihoods from the two systems. Table 6.5 shows that the combination yields improved word level performance as compared to the baseline system.

Table 6.5: Performance of the Baseline system, the Mean TRAP-based combined system and a combination of these two systems.

SYSTEM	WORD ERROR
Baseline	6.5 %
Mean TRAP-based	11.5 %
Combined	5.7 %

6.4 Neural TRAPs

The Mean TRAP-based system uses simple linear correlation classifiers, which compare the incoming trajectory of spectral energies with the Mean TRAP templates derived for each phoneme. Use of a more sophisticated classifier could improve performance. Hence, we explored a neural network based non-linear classifier, in order to provide a generalization of the Mean TRAP classifiers.

Fig. 6.8 shows a single Neural TempoRAl Pattern (TRAP) classifier. A feed-forward multi-layer perceptron (MLP) is used to classify the center frame of a 1 sec long (101 points) temporal trajectory, in each critical band. The temporal trajectory comprises of the logarithmic energies in the particular critical band. Each Neural TRAP classifier has 300 hidden units and 29 outputs. It is trained on 2 hours of the OGI Stories corpus for

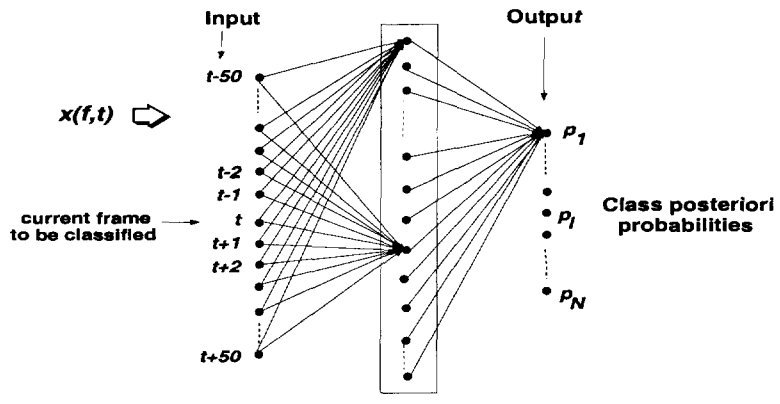


Figure 6.8: Neural TRAP.

the 29 phonetic classes. The trained TRAPs are tested on the OGI Numbers corpus. The Table 6.6 shows that the performance of the sub-band Neural TRAP classifiers is better than the performance of the sub-band Mean TRAP classifiers.

Table 6.6: Frame-level performance of different TRAPs on OGI Numbers corpus. Robust Neural TRAPs refers to the case when the input time trajectories are mean subtracted, variance normalized and windowed.

System	Frame error for each critical band
Mean TRAPs	78 - 82 %
Neural TRAPs	65 - 69 %
Robust Neural TRAPs	66 - 74 %

6.4.1 Robust Neural TRAPs

To make the input representation robust to convolutive channel distortions, the mean can be subtracted from the available 1 sec temporal trajectory vector. To compensate for the decrease in the variance of the vector, commonly observed in the presence of additive environmental noise conditions (Moreno et al., 1995; Tibrewala and Hermansky, 1997), each input 1 sec time-trajectory can be normalized to have unity variance. In order to de-emphasize the contributions of the spectral energies towards the edges of the time trajectory, each input pattern can be further weighted by a Hamming window, as in the case of Mean TRAP classifiers. Fig. 6.9 shows the effect of mean and variance normalization,

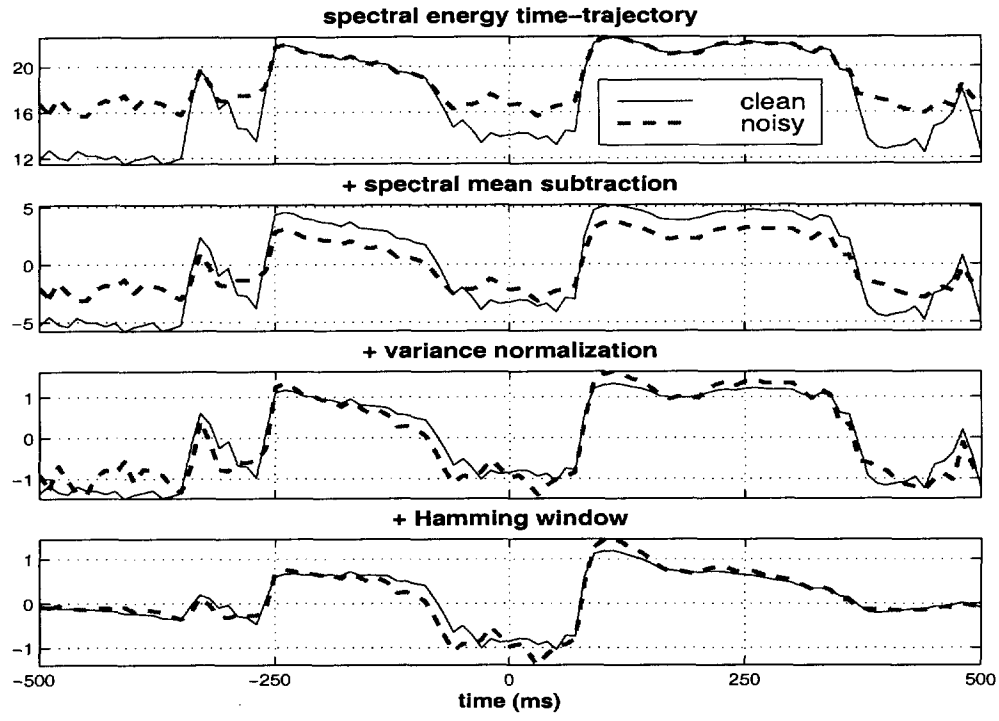


Figure 6.9: Effect of mean subtraction, variance normalization and windowing on a 1 sec spectral time trajectory of clean speech and speech corrupted with additive white noise. Spectral trajectory shown is at 500 Hz.

and windowing on a particular 1 sec vector. The normalization attempts to reduce the mismatch between clean speech case and the case with white noise added. Though this normalization is expected to improve performance in mismatched conditions, it is seen from the Table 6.6 that the additional mean and variance normalization (referred to as *Robust Neural TRAPs*) degrades the performance, as compared to the Neural TRAPs, in matched conditions. However, the performance is still better than that of the Mean TRAPs.

6.4.2 Combination of Neural TRAPs

Each input speech frame is classified by the 15 Neural TRAPs corresponding to the 15 critical bands. To obtain a single classification result for a frame, we use a MLP for combining the outputs obtained from each of the 15 TRAPs, as used in the case of Mean TRAP classifiers (Section 6.3.2). The input to the combining network is the concatenated

vector of the estimated class conditional log-likelihoods of the 29 phonetic classes, from each of the 15 TRAPs (435 dimensional input). The network has a single hidden layer of 300 units and 29 outputs which represent the merged estimate of the class posteriori probabilities. The combiner network is trained on the 1.7 hours subset of the Numbers corpus.

Table 6.7: Performance of the combined TRAPs on the OGI Numbers corpus.

SYSTEM	FRAME ERROR	WORD ERROR
Baseline	21 %	6.5 %
Mean TRAP-based	22 %	11.5 %
Neural TRAP-based	18.7 %	7.6 %
Robust Neural TRAP-based	20.0 %	8.8 %

Table 6.7 compares the frame errors and the word errors of the baseline system and the Neural TRAP-based combined system. The performance of the Neural TRAP-based combined system on the frame-level is better than that of the baseline system and the Mean TRAP-based combined system. On the word level, the Neural TRAP-based system performs significantly better than the Mean TRAP-based system. However, it is still worse as compared to the baseline system. On comparing the frame-level errors it is seen that the Neural TRAP-based system still suffers in the recognition of some vowels as compared to the baseline system. The Robust Neural TRAP-based system results in slightly degraded performance as compared to the Neural TRAP-based system.

6.4.3 Combination of the Baseline and Neural TRAP-based system

An analysis of the frame errors of the baseline and Neural TRAP-based system shows that approximately 40% of the errors made by the two systems are complementary. This indicates that both the systems yield significant complementary information. Such a situation makes both systems good candidates for merging (Sharma et al., 1998), as was observed in the case of Mean TRAPs-based system (Section 6.3.3).

We combined the outputs of the baseline system and the Neural TRAP-based system at the frame level, using a simple averaging of the corresponding log-likelihoods obtained

Table 6.8: Performance of the Baseline system, and combination of the baseline system with the TRAP-based systems.

System	Word error
Baseline	6.5 %
Combined with Baseline	
Mean TRAP-based	5.7 %
Robust Neural TRAP-based	4.7 %

from the two systems. Table 6.8 shows that the combination results in a significant improvement (at the 0.001 level using a binomial significance test) in performance as compared to the baseline system.

6.5 Broad TRAPs

The Broad TRAPs were derived in Section 5.3, by clustering the Mean TRAPs in each critical band, using the correlation similarity measure. The clustering results in 5 broad categories, which also correspond to the broad phonetic categories in speech - i.e., vowels, stops, fricatives, schwas and silence. For every incoming 1 sec time trajectory vector, we can compute the similarities between the vector and the corresponding frequency band Broad TRAPs. These similarity scores will represent the degree of vowel-like, stop-like, fricative-like, schwa-like and silence-like qualities, contained in the time trajectory of the phonetic class represented by the center frame.

Fig. 6.10 shows the average similarity scores for all occurrences of the 4 phonetic classes, /iy/, /ah/, /f/ and /t/ in the cross-validation set of the OGI-Numbers database, with respect to the 5 Broad TRAPs. The scores within each broad category are plotted for all the 15 critical band Broad TRAPs, resulting in (5 x 15) 75 scores per phoneme. The figure shows that the vowels /iy/ and /ah/ have a high degree of vowel-like quality. However, the scores representing the vowel-like quality vary across critical bands depending on the specific vowel. For example, for the vowel /iy/ the scores are higher in the lower and higher frequency bands of the vowel TRAP. This conforms with the fact that /iy/ is a front vowel. While, for the vowel /ah/, which is a back vowel, the scores have a higher

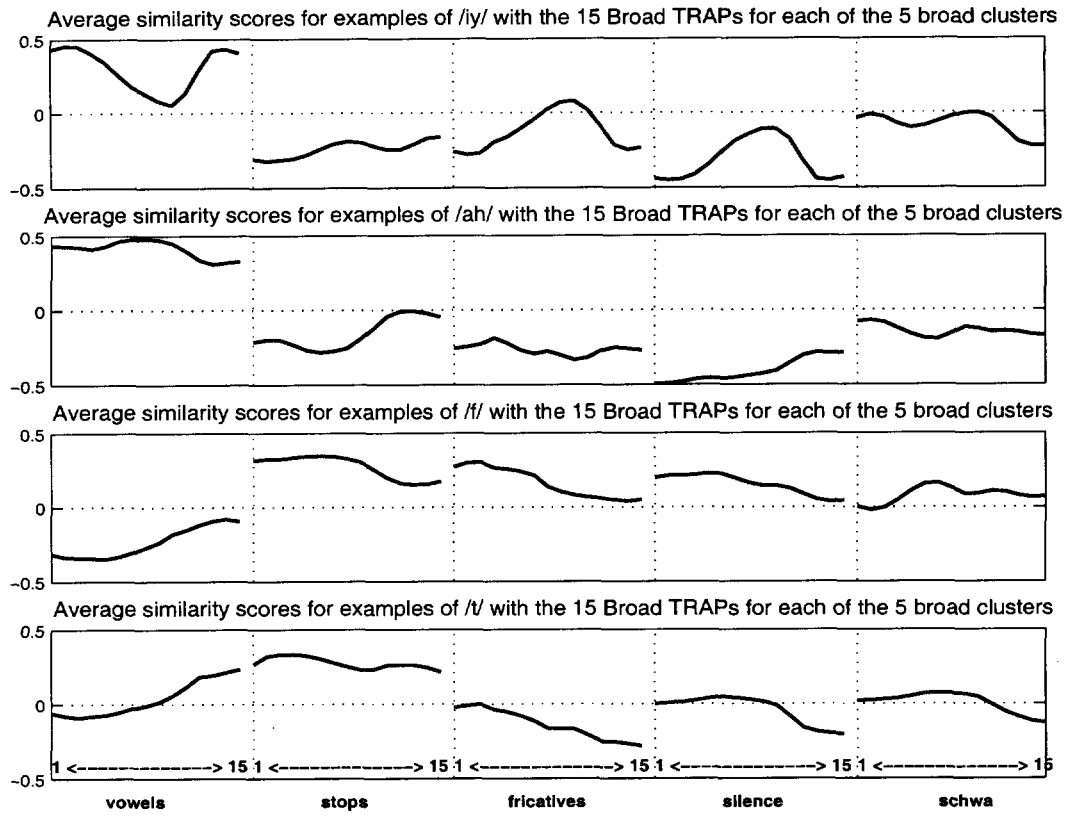


Figure 6.10: Average similarity scores for each of the phonemes /iy/, /ah/, /f/ and /t/. The regions bounded by dotted lines represent the 15 frequency bands within each of the 5 Broad TRAP categories, as labeled on the x-axis.

value in the low and mid frequency regions. The fricative /f/ has a high degree of fricative and stop-like characteristic, while the stop /t/ has a high degree of stop-like quality.

If we combine the 75 similarity scores obtained per frame, so as to exploit the variations in the degrees of the five qualities, we should obtain good discrimination among the phonetic categories.

6.5.1 Classification using Broad TRAPs

The Broad TRAPs can be used for the classification of incoming time trajectories into the corresponding broad phonetic categories, using the template matching technique described in Section 6.3. The frame-level error for such a classification in each critical band, is in the range of 32% - 40%. The 5 correlation based similarity scores obtained in each critical

band can further be used for phonetic classification, by using a MLP for combination. The scores obtained from each of the 15 critical bands are concatenated to yield a 75 dimensional vector. This vector forms the input to a combiner which has 500 hidden units and 29 outputs.

Table 6.9: Performance with the Broad TRAPs.

System	Frame error	Word error
Mean TRAP-based	22 %	11.5 %
Broad TRAP-based	24.6 %	12.8 %

Table 6.9 shows that the performance using Broad TRAPs is only slightly worse as compared to the performance yielded by the Mean TRAPs. This indicates the possibility of significant dimensionality reduction in the TRAP-based system. It also suggests that the full phoneme classification on each sub-band temporal energy pattern may not be necessary.

6.6 Performance of the TRAP-based System in Noise

To assess possible advantages of the TRAP-based recognizer in degraded environments, we tested it on speech artificially degraded by various types of noise. We use only the Robust Neural TRAP-based system for testing under mismatched conditions.

6.6.1 TRAPs in convolutive noise

The baseline system uses utterance-based cepstral mean subtracted features, which are known to be robust to convolutive noise. TRAPs should also be robust to such a distortion because of the local (1 sec) input mean removal. To simulate convolutive distortion, the test data was pre-processed by a pre-emphasis filter $(1 - 0.95z^{-1})$.

Table 6.10 shows that the performance of the baseline system without cepstral mean subtraction degrades rapidly from 21.8% frame error and 8% word error on clean test data, to 33.3% frame error and 16% word error on pre-emphasized data. On the other hand, both the baseline system with mean subtraction and the TRAP-based system show

Table 6.10: Comparison of the degradation in performance from clean test condition to condition corrupted by convolutive distortion.

SYSTEM	FRAME ERROR		WORD ERROR	
	Clean	Noise	Clean	Noise
Baseline without CMS	21.8 %	33.3 %	8.0 %	16 %
Baseline	21 %	22.5 %	6.5 %	7 %
TRAP-based	22 %	21.1 %	8.8 %	10.2 %
Baseline + TRAP			4.7 %	5.3 %

only a slight degradation in performance to such convolutive distortion, as compared to the clean test case. This demonstrates an inherent robustness of the TRAP-based system to convolutive channel distortion.

6.6.2 TRAPs in additive sinusoidal noise

Table 6.11: Frame and Word errors (%).

SYSTEM	FRAME	WORD
Baseline	42.57 %	36.9 %
TRAP	29.15 %	18.1 %

The performance of the TRAP recognizer on additive sinusoidal noise at 1 kHz and SNR 10dB is shown in Table 6.11. The TRAP-based system results in half the error rate as compared to the baseline system. This result is consistent with the results of the multi-band system on narrowband noise.

6.6.3 TRAPs in realistic additive noise

Realistic noises (white, pink, factory and destroyer-engine) from the NOISEX-92 database were added to the data. Table 6.12 compares the performance of the baseline, TRAP-based and combined systems in the presence of these noise conditions. It is seen that the TRAP-based system consistently gives reduced frame error as compared to the baseline

Table 6.12: Frame and word errors (%) for the additive noises from the NOISEX-92 database.

System	White		Pink		Factory		Engine	
	Frame	Word	Frame	Word	Frame	Word	Frame	Word
Baseline	41.59	23.4	49.39	33.5	42.15	24.3	41.56	24.9
TRAP	37.52	25.1	45.81	33.9	39.21	26.6	34.25	21.6
Combined		17.8		24.6		19.0		15.5

system and gives quite comparable performance on the word level. The combined baseline and TRAP system results in significant improvement in the word level performance. Specifically, the combined system results in around 25% reduction in word error (average reduction on the four noise conditions) as compared to the baseline system.

6.7 Summary

In this chapter, we investigated an extension of the multi-band model to incorporate medium-term temporal information. The resulting TRAP-based system represents a generalization of the multi-band system. Each sub-band consists of only one critical band, hence any across-frequency correlation is ignored. Each sub-band uses a 1 sec time trajectory of spectral energies per input frame for phonetic classification.

Our experiments show that it is possible to classify phonemes with a reasonable accuracy (around 30%) based on rather long (longer than a single phoneme) temporal pattern of spectral energy, in a single critical band alone. We also show that by combining classification results from the individual critical bands, we can achieve recognition performance quite competitive with the current state-of-the-art spectral-based ASR systems. The resulting temporal (TRAP-based) system also yields significant amount of complementary information, as compared to the conventional (spectral-based) system. Hence, a combination of these two systems yields a significant improvement in performance (around 25% reduction in word error rate) over the baseline system, in both matched and mismatched test conditions.

This work thus represents a further development of the Fletcher-Allen model (Allen,

1994) of speech recognition. It supports the notion of moving away from the conventional *across spectrum processing* approach towards an *across time processing* approach to speech recognition.

Chapter 7

Conclusions and Future Work

This dissertation presents an approach to address the problem of robustness of speech recognition systems to environmental noise. The approach taken is to incorporate two concepts into the ASR system - 1) the partial recognition of phonemes in several frequency sub-bands, and 2) syllable-length temporal processing of speech information. Psychoacoustic evidence suggests that the human auditory system is capable of the processing implied by these two concepts (Fletcher, 1953; Allen, 1994; Hermansky, 1998; Greenberg, 1996), which we believe accounts to some extent for its robustness. We have used a multi-stream framework to incorporate the two concepts.

In this work, we have proposed and developed two multi-stream models for speech recognition. We have also shown the advantages of these models in dealing with environmental noise. The first model is the *multi-band model*, which processes different regions of the speech spectrum independently, in contrast to the conventional across-spectrum processing model. The phone probability estimates from each of the frequency sub-bands are then merged for final decoding purposes. The second multi-stream model is the *TempoRAl Pattern Classifier (TRAP) model*, which is an extension of the multi-band model. In addition to using independent frequency processing as in the multi-band model, it also incorporates medium-term temporal dependencies (around 200-500ms) in each frequency channel.

In Section 7.1, we present the conclusions from this work. Section 7.2 suggests directions for future work. Section 7.3 concludes with some final remarks.

7.1 Conclusions

The main conclusions from the work presented in this dissertation are:

Independent frequency processing:

We have shown, using the multi-band and the TRAP models, that it is feasible to independently estimate phonetic probabilities from different regions of the frequency spectrum and then merge these partial results. The resulting system, which overlooks possible across frequency correlations, performs comparably to the conventional system, which uses correlations across frequency regions.

Medium-term temporal processing:

We have shown, using the TRAP model, that is it viable to use an *across-time processing* approach which processes medium time-spans (around 200-500ms), as opposed to an *across-frequency approach* to speech recognition. We used a 1 sec time trajectory in each frequency channel, instead of the conventional 10-100ms spectral vector, to classify the phonetic category in the center of the vector. Our experiments show that it is possible to classify phonemes with reasonable accuracy (30%) in the relatively narrow (critical band) sub-bands. On a continuous numbers task, the system obtained by the combination of the temporal sub-band classifiers yields performance close to that of the conventional system.

Complementary nature of spectral and temporal systems

We have shown that the temporal-based (TRAP) system yields considerable complementary phonetic information as compared to the conventional spectral based system (40% of the errors of the two systems are complementary). A combination of these two systems yields around 25% reduction in word errors.

Robustness to environmental degradations

We have shown that merging independently processed frequency sub-bands makes the system robust to narrow-band noise, for e.g., we showed using an isolated digit task, that

the multi-band model is inherently robust to additive noise which corrupts some frequency sub-bands more than the others. If we estimate the sub-band reliability and merge only the more reliable sub-bands, a further improvement in performance is obtained.

We have shown that medium-term temporal processing also improves noise robustness. The TRAP-based system is robust to convolutive distortions due to the 1 sec mean removal from the time trajectories of critical band energies. On a continuous numbers task, the combined spectral and TRAP-based (temporal) system yields around 25% reduction in word errors in additive noise conditions.

7.2 Suggestions for Future Work

The systems developed in this dissertation can be considered as a proof of concept that - 1) correlations across frequency regions are less important to speech recognition than assumed by most state-of-the-art systems, and 2) temporal correlations can play a significant role in improving noise robustness of speech recognizers. Hence, further studies on the issues of model implementation such as sub-band definitions could improve performance. We have used cepstral-based features in the sub-bands, as used in most state-of-the-art systems. However, research on features in the conventional systems has been on-going for several decades. Likewise, the multi-band model could also benefit from efforts to improve the sub-band features.

Other areas of future work could include:

Extending the scope to multi-input framework

In this work we focussed on the single-input, multi-stream framework. However, the techniques discussed for sub-stream combination and reliability estimation can be applied to the generalized multi-stream framework (discussed in Section 1.3) which uses multiple inputs. For example, visual input in addition to audio input or audio input from multiple microphones. Future work could investigate the multi-input framework.

Merging techniques

We have used hybrid (HMM/MLP) based recognizers for all the experiments reported in this dissertation. However, HMM based recognizers are widely used in speech research. The definition of sub-bands can be easily used in the HMM framework. It is also easy to combine the HMM-based sub-bands, provided the sub-streams are independent. The estimated sub-band log-likelihoods can then be combined linearly, as given in Eq. 3.10. This technique has been used in (Tomlinson et al., 1997) and (Okawa et al., 1998) with some success. However our experiments on merging techniques, as discussed in Section 3.7.3, indicate that the non-linear merging of the log-likelihoods yields an improvement in performance, as compared to a linear merging (or weighted linear merging) technique. Use of non-linear merging techniques should be investigated in the HMM framework.

Sub-band reliability estimation in large vocabulary tasks

We have shown in Section 4.4.1 that combination of the more reliable sub-bands in the multi-band system improves performance in noisy conditions. The selection strategy which uses sub-band SNR estimation can be applied to the large-vocabulary recognition task. However, most of the classifier-based techniques (discussed in Section 4.2.2) resort to majority voting among the top N% of the merging classifiers which have high reliability estimates. Majority voting is feasible in case of an isolated word recognition task. Extension of the classifier-based techniques to large vocabulary tasks can be investigated.

Features in TRAP classifier

We have used 1 sec time trajectories of log critical band energies as features in each TRAP classifier. We found that mean removal and variance normalization of each 1 sec input vector improved the robustness of the TRAP classifier. Other techniques for improving robustness, such as time trajectory smoothing using low-pass filters, could be investigated.

7.3 Final Remarks

This work represents a small step in the direction of moving away from the commonly accepted paradigm of *across-frequency processing* approach to speech recognition. The work in this dissertation demonstrates that speech recognizers can gain by incorporating some concepts that are appealing from the engineering point of view, as well as that appear to account for the robustness of the human speech perception system. We have shown that even a simple engineering modeling of these concepts can improve system performance, especially in mis-matched conditions, without the need for re-training the system.

It is our hope that this work will encourage new approaches to acoustic modeling in speech recognition.

Bibliography

Adjoudani, A. and Benoit, C. (1996). On the integration of auditory and visual parameters in an HMM-based ASR. In *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 461–471, Berlin. Springer Verlag, Berlin.

Allen, J. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577.

Allen, J. B. (1996a). Harvey fletcher's role in the creation of communication acoustics. *Journal of the Acoustical Society of America*, 99(4):1825–1839.

Allen, J. B. (1996b). How do humans process and recognize speech? In Ramachandran, R. and R.Mammone, editors, *Modern Methods of Speech Processing*, pages 251–275. Kluwer Academic Publishers, Boston.

Arai, T., Hermansky, H., Pavel, M., and Avendano, C. (1996). Intelligibility of speech with filtered time trajectories of spectral envelopes. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, volume 4, pages 2490–2493, Philadelphia.

Arthanari, T. and Dogde, Y. (1981). *Mathematical Programming in Statistics.* John Wiley & Sons, New York.

Avendano, C., van Vuuren, S., and Hermansky, H. (1996). Data-based RASTA-like filter design for channel normalization in ASR. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, volume 4, pages 2087–2090, Philadelphia.

Bourlard, H. and Dupont, S. (1996). A new ASR approach based on independent processing and re-combination of partial frequency bands. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 426–429, Philadelphia.

Bourlard, H. and Dupont, S. (1997). Sub-band based speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume 2, pages 1251–1254, Munich, Germany.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition — A Hybrid Approach.* Kluwer Academic Publishers, Massachusetts.

Bregman, A. (1990). *Auditory Scene Analysis.* MIT Press.

Clark, J. J. and Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing Systems.* Kluwer Academic Publishers, Massachusetts.

Cohen, J. (1989). Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, 85(6):2623–2629.

Cole, R., Noel, M., and Lander, T. (1994). Telephone speech corpus development at CSLU. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'94*, volume 4, pages 1815–1818, Yokohama, Japan.

Cole, R. A., Noel, M., Lander, T., and T.Durham (1995). New telephone speech corpora at CSLU. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'95)*, volume 1, pages 821–824, Madrid, Spain.

Cook, G., Waterhouse, S., and Robinson, A. (1997). Ensemble methods for connectionist acoustic modelling. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, 4:1959–1962.

Cooke, M., Green, P., and Crawford, M. (1994). Handling missing data in speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'94)*, volume 3, pages 1555–1558, Yokohama, Japan.

Cooke, M., Morris, A., and Green, P. (1996). Recognising occluded speech. In *Proceedings of the European Speech Communication Association (ESCA) Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, pages 1–8, Keele, England.

Davis, S. and Mermelstein, P. (1980). Comparisons of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366.

Dodd, D. and Campbell, R. (1987). *Hearing by Eye: The Psychology of Lipreading.* Lawrence Erlbaum Press.

Drucker, H., Schapire, R., and Simard, P. (1993). Improving performance in neural networks using a boosting algorithm. In S. Hanson, J. C. and Giles, C., editors, *Advances in Neural Information Processing Systems 5*, pages 42–49. Morgan Kaufmann.

Drullman, R., Festen, J., and R.Plomp (1994). Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95:2670–2680.

Duchnowski, P. (1992). A new structure for automatic speech recognition. *Ph.D. Thesis, MIT.*

Fletcher, H. (1953). *Speech and Hearing in Communication.* Krieger, New York.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Academic Press Inc.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 29:254–272.

Ghitza, O. (1987). Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer, Speech and Language*, 1(2):109–130.

Ghosh, J., Beck, S., and Chu, C.-C. (1992). Evidence combination techniques for robust classification of short-duration oceanic signals. *Proceedings of SPIE Conference on Applications of Artificial Neural Networks*, 1706:266–276.

Gong, Y. (1995). Speech recognition in noisy environments: a survey. *Speech Communication*, 16:261–291.

Green, D. M. and John, S. A. (1974). *Signal Detection Theory and Psychophysics.* Peninsula Publishing, Los Altos, California.

Green, P., Cooke, M., and Crawford, M. (1995). Auditory scene analysis and hidden markov model recognition of speech in noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, volume 1, pages 401–404, Detroit, MI.

Greenberg, S. (1996). Understanding speech understanding – Towards a unified theory of speech perception. In *Proceedings of the European Speech Communication Association (ESCA) Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, pages 1–8, Keele, England. ESCA.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

Hashem, S. (1994). Optimal linear combination of neural networks. *Neural Networks*, 10(4):599–614.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation.* Prentice Hall, New Jersey.

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

Hermansky, H. (1995). Exploring temporal domain for robustness in speech recognition, invited paper. *Proceedings of the 15th International Congress on Acoustics*, 3:61–64.

Hermansky, H. (1998). Should recognizers have ears? *Speech Communication*, 25(1-3):3–27.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

Hirsch, H. (1993). Estimation of noise spectrum and its applications to snr estimation and speech enhancement. *Technical report, TR-93-012, International Computer Science Institute, Berkeley, CA.*

Ho, T. K. (1992). A theory of multiple classifier systems and its application to visual word recognition. *PhD Thesis,State University of New York at Buffalo.*

Honathan, T. K. H., Hull, J., and Srihari, S. N. (1994). Decision combination in multiple classfier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75.

Huang, Y. and Suen, C. (1994). A method of combining multiple classifiers - a neural network approach. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, volume II, pages 473–475.

Jacob, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixture of local experts. *Neural Computation*, 3(1):79–87.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixture of experts and the em algorithm. *Neural Computation*, 6:181–214.

Kohonen, T. (1984). *Self-Organization and Associative Memory.* Springer-Verlag, Berlin.

Kryter, K. (1960). Speech bandwidth compression through spectrum selection. *Journal of the Acoustical Society of America*, 32(5):547–556.

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

Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Lippmann, R. and Carlson, B. (1997). Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, volume 1, pages 37–40.

Luettin, J., Hermansky, H., and van Vuuren, S. (1997). RASTA HMM: Relative spectra hidden markov model. In *Proceedings of the 1997 CLSP/JHU Summer Workshop on Innovative Techniques for Large Vocabulary Conversational Speech Recognition.*

Luo, R. and Kay, M. (1988). Multisensor integration and fusion: Issues and approaches. In *Proceedings of SPIE on Sensor Fusion*, volume 931, pages 19–25.

Lyon, R. F. and Yaeger, L. S. (1996). On-line hand-printing recognition with neural networks. In *Proceedings of the Fifth International Conference on Microelectronics for Neural Networks and Fuzzy Systems*, pages 201–212.

Massaro, D. and Cohen, M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5):753–771.

Miller, G. and Nicely, P. (1995). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27(2):338–352.

Mirghafori, N. and Morgan, N. (1998). Transmissions and transitions: A study of two common assumptions in multi-band ASR. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, volume II, pages 713–716, Seattle, Washington.

Moore, B. C. (1989). *An Introduction to the Psychology of Hearing.* Academic Press, San Deigo, California.

Moreno, P., Raj, B., Gouvea, E., and Stern, R. (1995). Multi-variate-gaussian-based cepstral normalization for robust speech recognition. *Proc. ICASSP-95*, 1:141–144.

Morgan, N. and Bourlard, H. (1995). An introduction to hybrid HMM/connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42.

Okawa, S., enrico Bocchieri, and Potamianos, A. (1998). Multi-band speech recognition in noisy environments. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, volume 2, pages 641–644.

O'Shaughnessy, D. (1987). *Speech Communication - Human and Machine.* Addison-Wesley Publishing Company, Reading, MA.

Pavel, M. and Hermansky, H. (1997). Information fusion by human and machine. *Proceedings of the First European Conference on Signal Analysis and Prediction*, pages 350–353.

Perrone, M. P. and Cooper, L. N. (1993). When networks disagree: Ensemble methods for hybrid neural networks. In Mammone, R., editor, *Artificial Neural Networks for Speech and Vision*, pages 126–142. Chapman-Hall,London.

Petajan, E. (1987). An improved automatic lipreading system to enhance speech recognition. *Technical Report,11251-871012-111TM, AT&T Bell Labs.*

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey.

Rabiner, L. R. (1997). Applications of speech recognition in the area of telecommunications. *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 501–510.

Richard, M. and Lippmann, R. (1991). Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3(4):461–483.

Riener, K., Warren, R., and Jr., J. B. (1992). Novel findings concerning intelligibility of bandpass speech. *Journal of the Acoustical Society of America*, 91(4):S2339.

Robinson, T., Almeida, L., Boite, J., Bourlard, H., Fallside, F., Hochberg, M., Kershaw, D., Kohn, P., Konig, Y., Morgan, N., Neto, J., Renals, S., Saerens, M., and Wooters, C. (1993). A neural network based, speaker independent, large vocabulary, continuous speech recognition system, the WERNICKE project. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'93)*, volume 3, pages 1941–1944, Berlin, Germany.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.

Seneff, S. (1985). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):55–76.

Shamma, S. A. (1985). Speech processing in the auditory system ii: lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, 78(5):1662–1632.

Sharma, R. K., Leen, T. K., and Pavel, M. (1999). Probabilistic image sensor fusion. In Kearns, M. S., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems 11*, pages 824–830. MIT Press.

Sharma, S., Vermeulen, P., and Hermansky, H. (1998). Combining information from multiple classifiers for speaker verification. *Proceedings of the Speaker Recognition and its Commercial and Forensic Applications Workshop (RLA2C)*, pages 115–119.

Shimshoni, Y. and Intrator, N. (1996). Classifying seismic signals by integrating ensembles of neural networks. *IEEE Transactions on Signal Processing*, 46(5):1194–1201.

Silsbee, P. and Bovik, A. (1996). Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351.

Stork, D., Wolff, G., and Levine, E. (1992). Neural network lipreading system for improved speech recognition. *Proceedings of the International Joint Conference on Neural Networks*, pages 285–295.

Thomson, D. L. (1997). Ten case studies of the effect of field conditions on speech recognition errors. *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 511–518.

Tibrewala, S. and Hermansky, H. (1997). Multi-band and adaptation approaches to robust speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, volume 5, pages 2619–2622.

Tomlinson, M., Russell, M., Moore, R., Buckland, A., and Fawley, M. (1997). Modeling asynchrony in speech using elementary single-signal decomposition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume II, pages 1247–1250.

van Vuuren, S. and Hermansky, H. (1997). Data-driven design of RASTA-like filters. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, pages 409–412.

Varga, A. and Moore, R. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'90)*, volume 2, pages 845–848.

Varga, A. and Steeneken, H. J. (1993). Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251.

Waibel, A., Sawai, H., and Shikano, K. (1989). Modularity and scaling in large phonemic neural networks. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 37(12):1888–1898.

Wang, K. and Shamma, S. A. (1995). Spectral shape analysis in the central auditory spectrum. *IEEE Transactions on Speech and Audio Processing*, 3(5):382–395.

Warren, R., Riener, K., Jr., J. B., and Brubaker, B. (1995). Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception & Psychophysics*, 57(2):175–182.

Waterhouse, S. and Cook, G. (1997). Ensemble methods for phoneme classification. In M.C. Mozer, M. J. and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, pages 800–806. MIT Press.

Wolpert, D. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Xu, L., Kryzak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435.

Yang, H., van Vuuren, S., and Hermansky, H. (1999). Relevancy of the time-frequency features for phonetic classification measured by mutual information. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, volume 1, pages 225–228.

Appendix A

TIMIT Phoneme Set

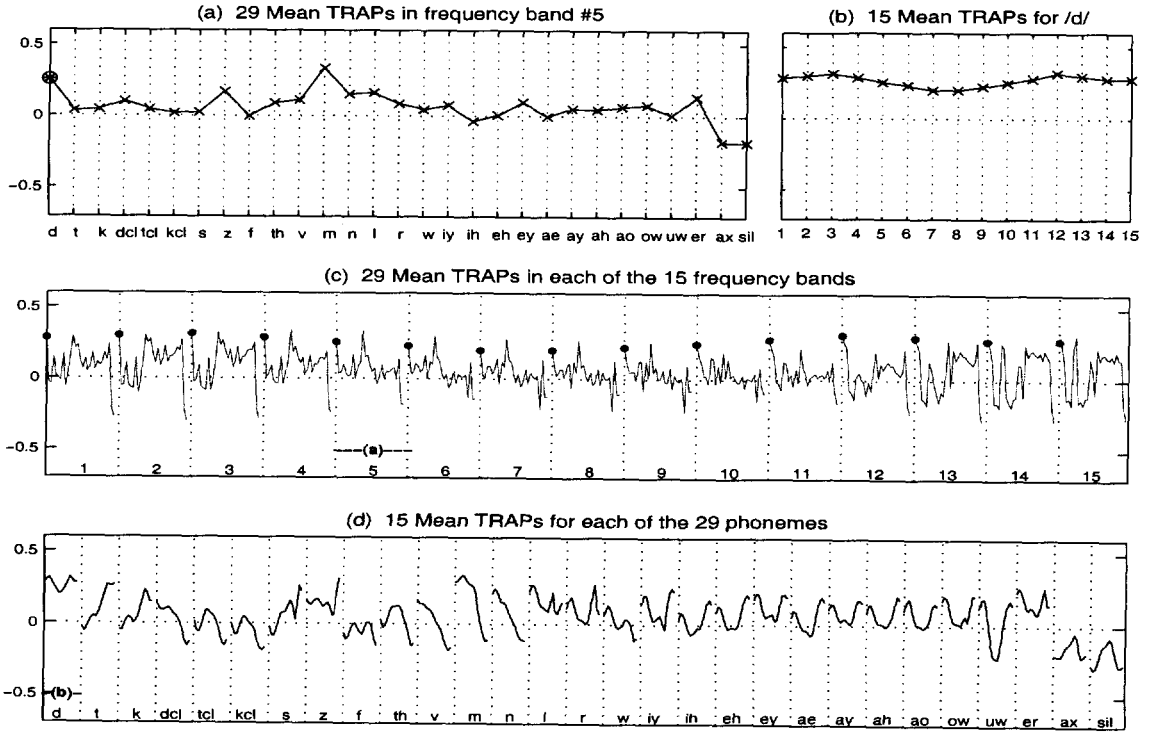
TIMIT	Worldbet	Example		TIMIT	Worldbet	Example
b	b	**b**ee		l	l	lay
d	d	**d**ay		r	9r	**r**ay
g	g	**g**ay		w	w	**w**ay
p	ph	**p**ea		y	y	**y**acht
t	th	**t**ea		hh	h	**h**ay
k	kh	**k**ey		hv		a**h**ead
dx	d_(	dirty		el	l=	bott**le**
q	_?*	glottal onset		iy	i:	b**ee**t
jh	dZ	**j**oke		ih	I	b**i**t
ch	tS	**ch**oke		eh	E	b**e**t
bcl	bc	(b closure)		ey	ei	**ai**t
dcl	dc	(d closure)		ae	@	b**a**t
gcl	gc	(g closure)		aa	A	b**o**tt
pcl	pc	(p closure)		aw	aU	b**ou**t
tcl	tc	(t closure)		ay	ay	b**i**te
kcl	kc	(k closure)		ah		b**u**t
s	s	**s**ea		ao	¿	b**ough**t
sh	S	**sh**e		oy	¿i	b**oy**
z	z	**z**one		ow	oU	b**oa**t
zh	Z	a**z**ure		uh	U	b**oo**k
f	f	**f**in		uw	u	b**oo**t
th	T	**th**in		ux	u_x	t**oo**t
v	v	**v**an		er	3r	b**ir**d
dh	D	**th**en		ax	&	**a**bout
m	m	**m**om		ix	I_x	deb**i**t
n	n	**n**oon		axr	&r	butt**er**
ng	N	si**ng**		ax-h		su**s**pect
em	m=	bott**om**		pau		pause
en	n=	butt**on**		epi		epenthetic silence
eng	N=	washing**ton**		h#		begin/end marker
nx		win**n**er				

Table A.1: 61 TIMIT phonemes used in the multi-band experiments.

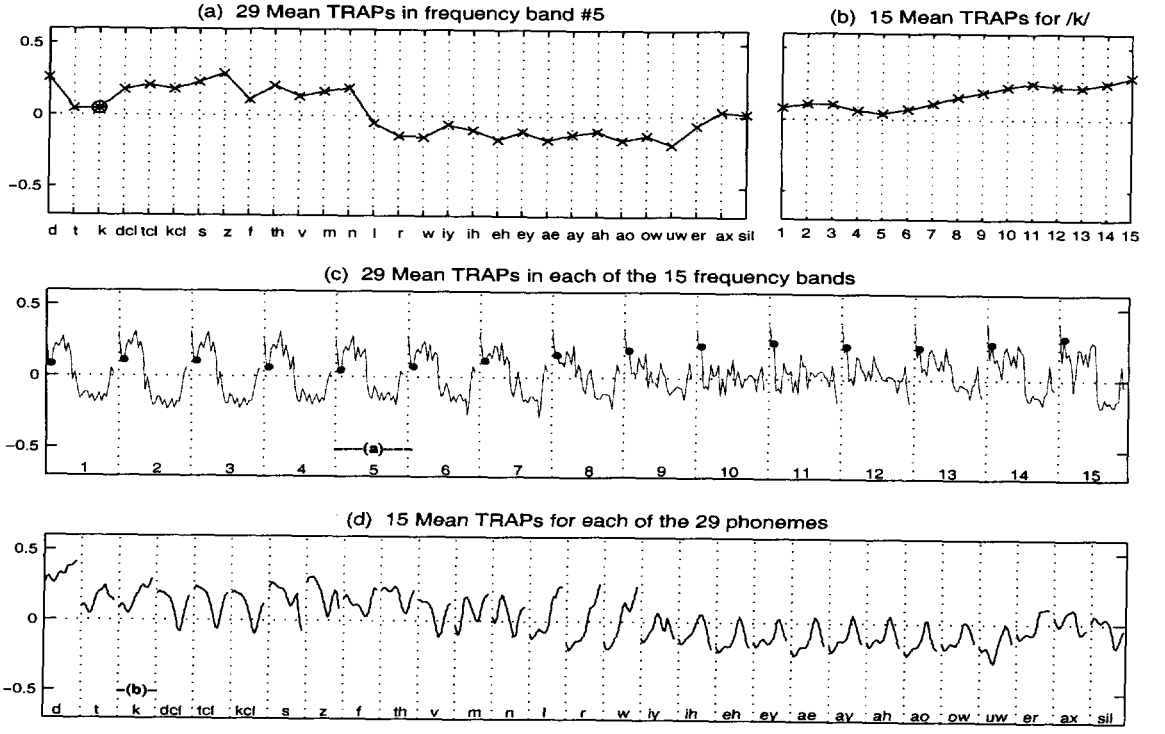
Appendix B

Mean TRAPs

The figures in this section show the average similarity scores obtained for all occurrences of 24 phonemes in the cross-validation portion of the OGI-Numbers database, as explained in Section 6.3.1.

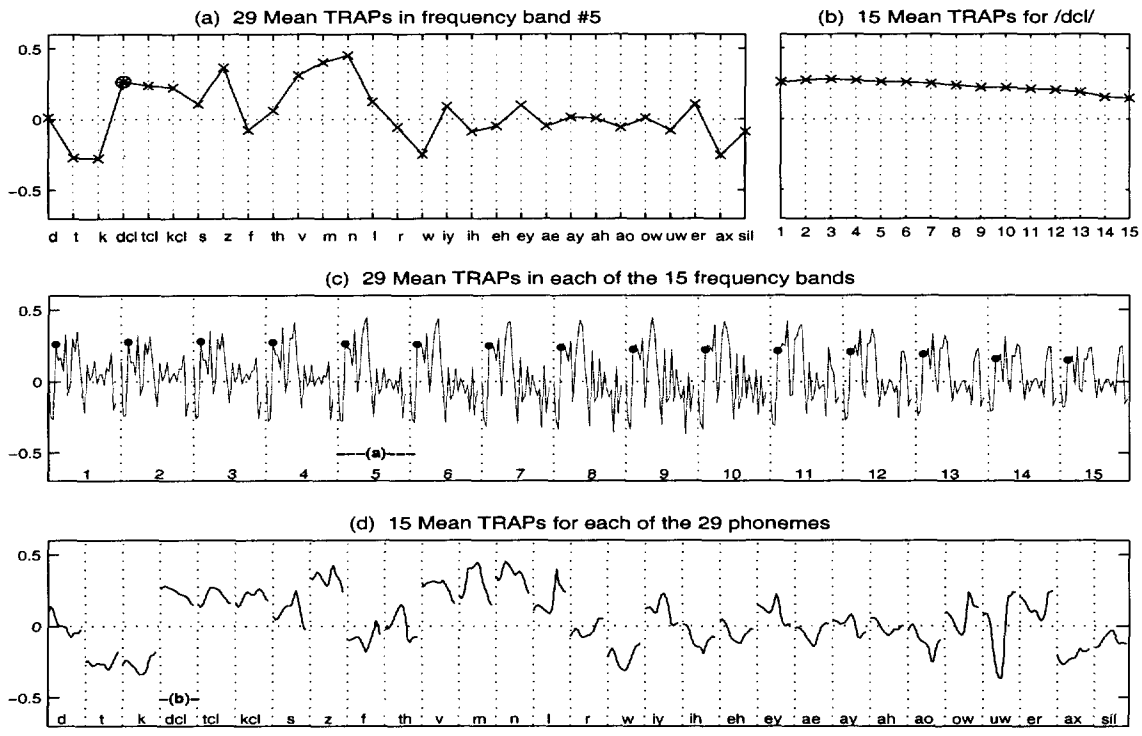


(I) **Average similarity scores for examples of /d/ with (a), (b), (c) and (d).**

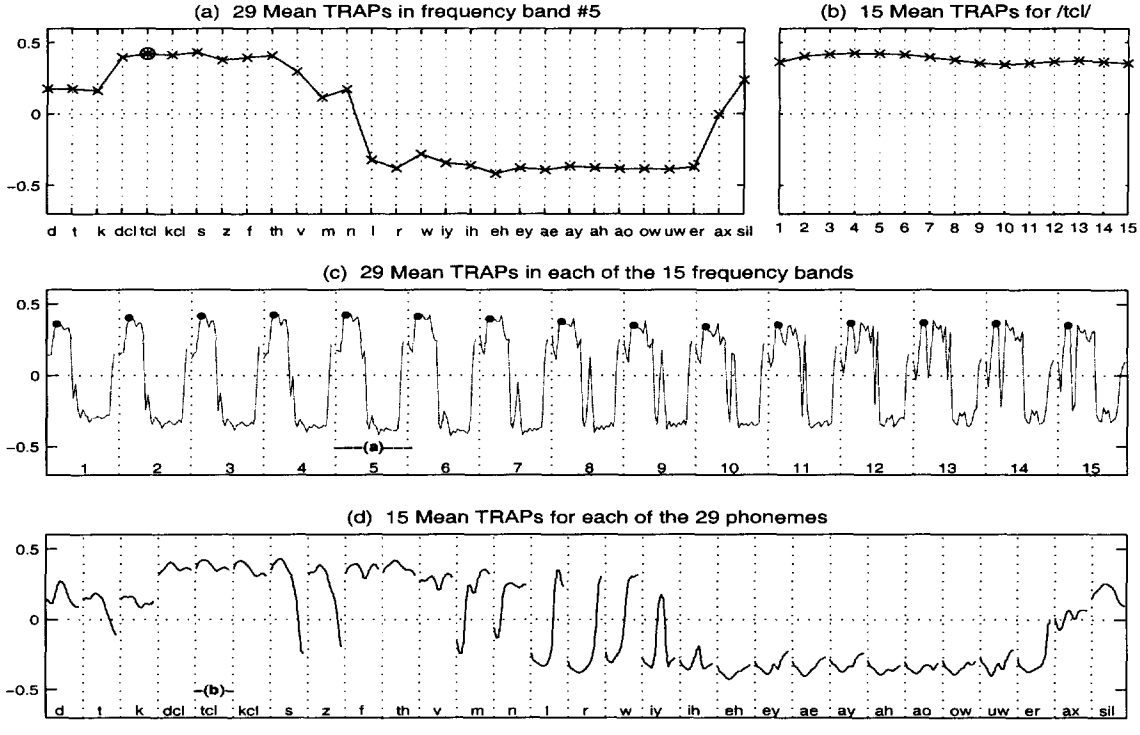


(II) **Average similarity scores for examples of /k/ with (a), (b), (c) and (d).**

Figure B.1: Comparison of the average similarity scores obtained for the 2 classes, /d/ and /k/, from the Mean TRAP classifiers.

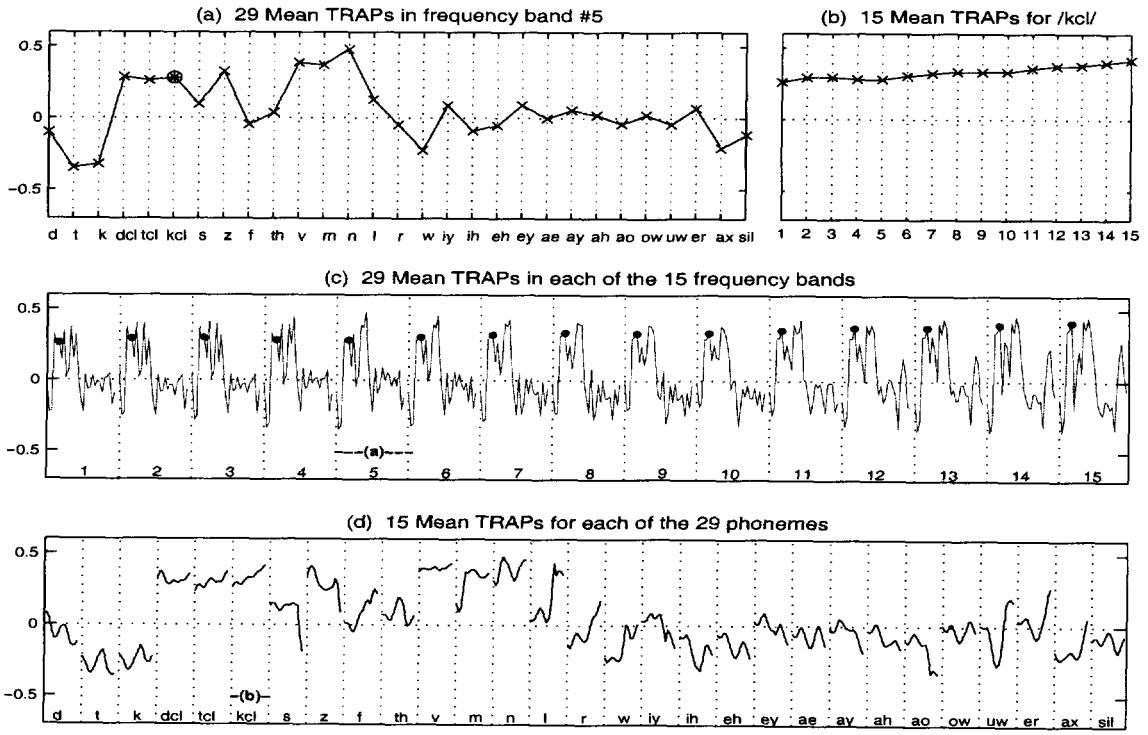


(I) **Average similarity scores for examples of /dcl/ with (a), (b), (c) and (d).**

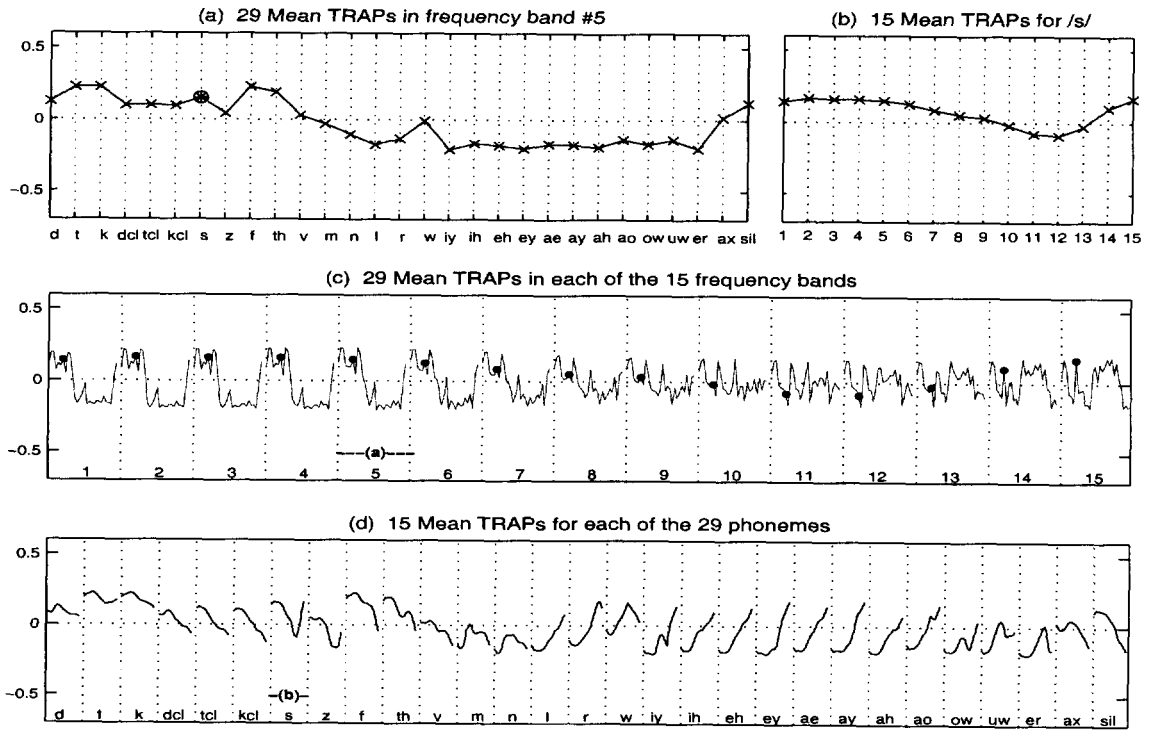


(II) **Average similarity scores for examples of /tcl/ with (a), (b), (c) and (d).**

Figure B.2: Comparison of the average similarity scores obtained for the 2 classes, /dcl/ and /tcl/, from the Mean TRAP classifiers.

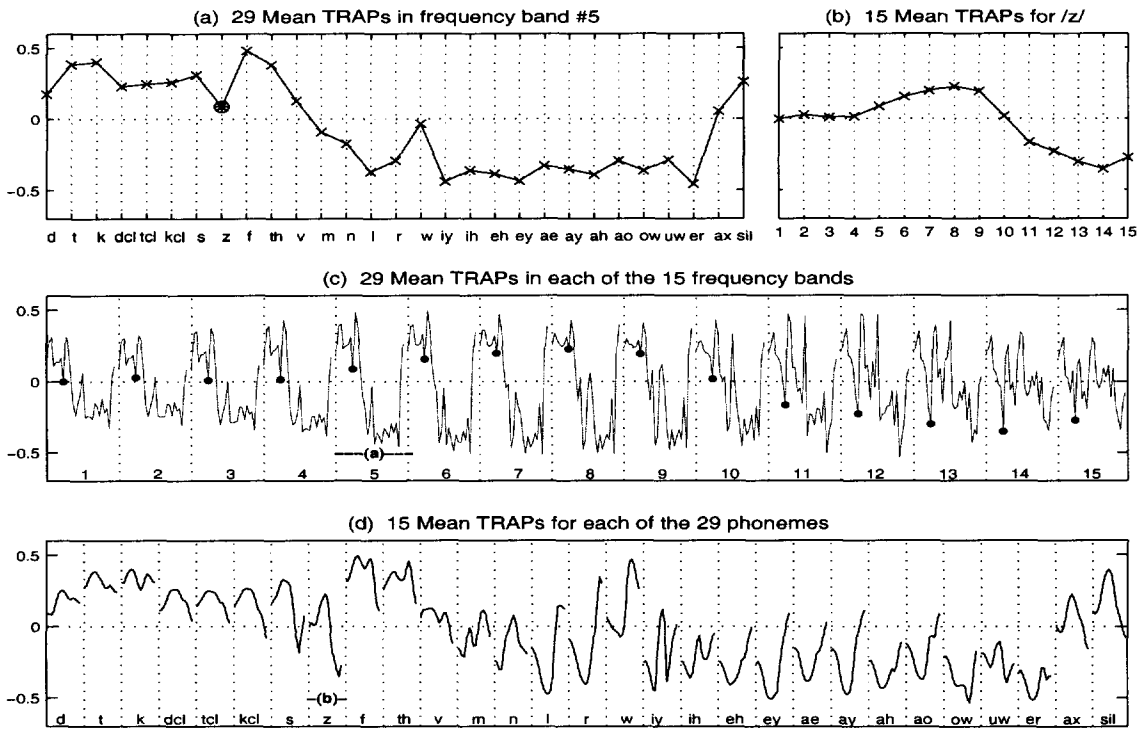


(I) Average similarity scores for examples of /kcl/ with (a), (b), (c) and (d).

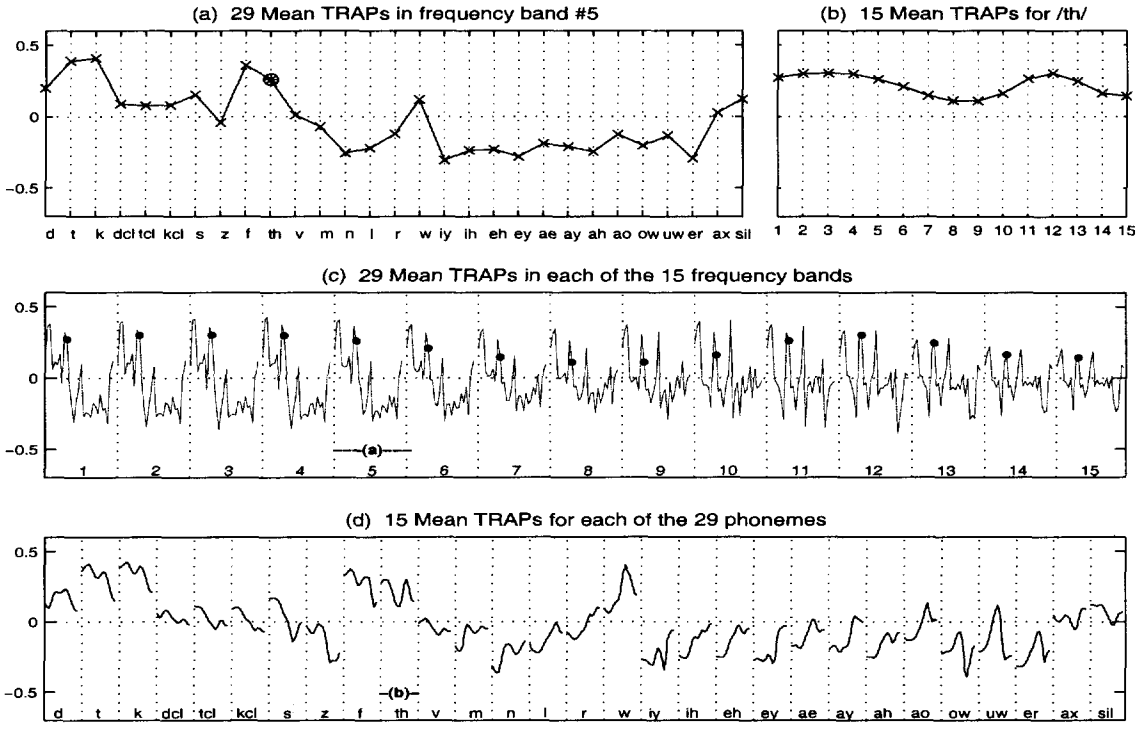


(II) Average similarity scores for examples of /s/ with (a), (b), (c) and (d).

Figure B.3: Comparison of the average similarity scores obtained for the 2 classes, /kcl/ and /s/, from the Mean TRAP classifiers.

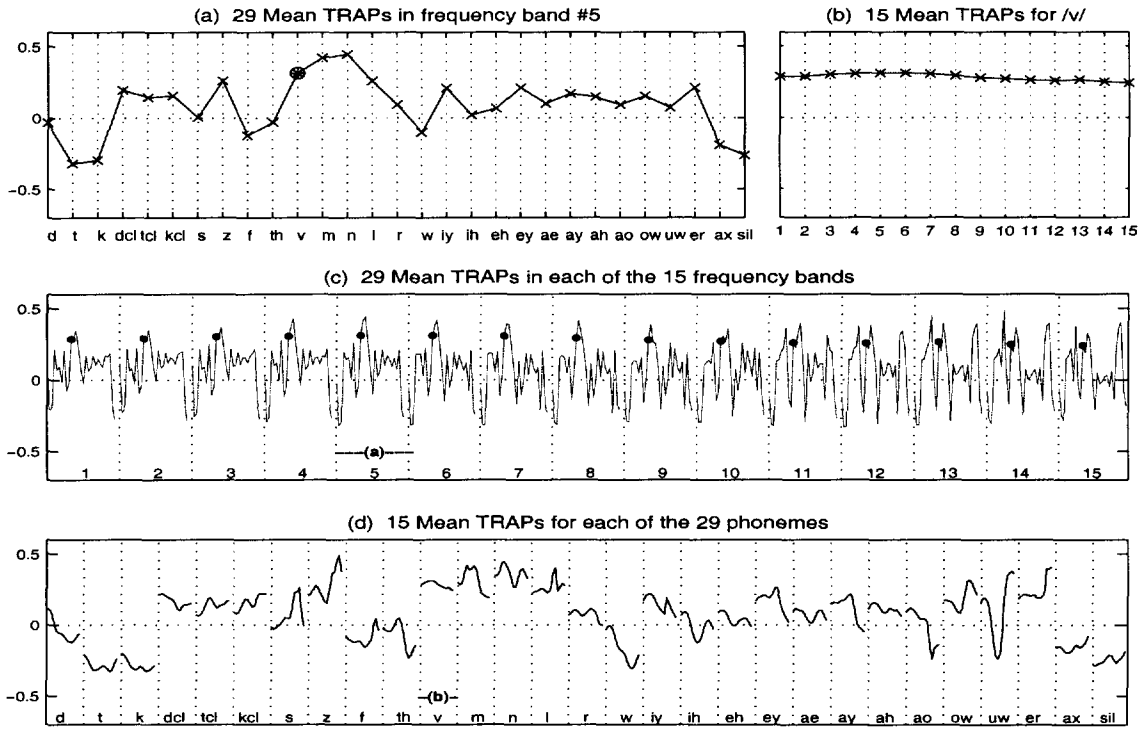


(I) **Average similarity scores for examples of /z/ with (a), (b), (c) and (d).**

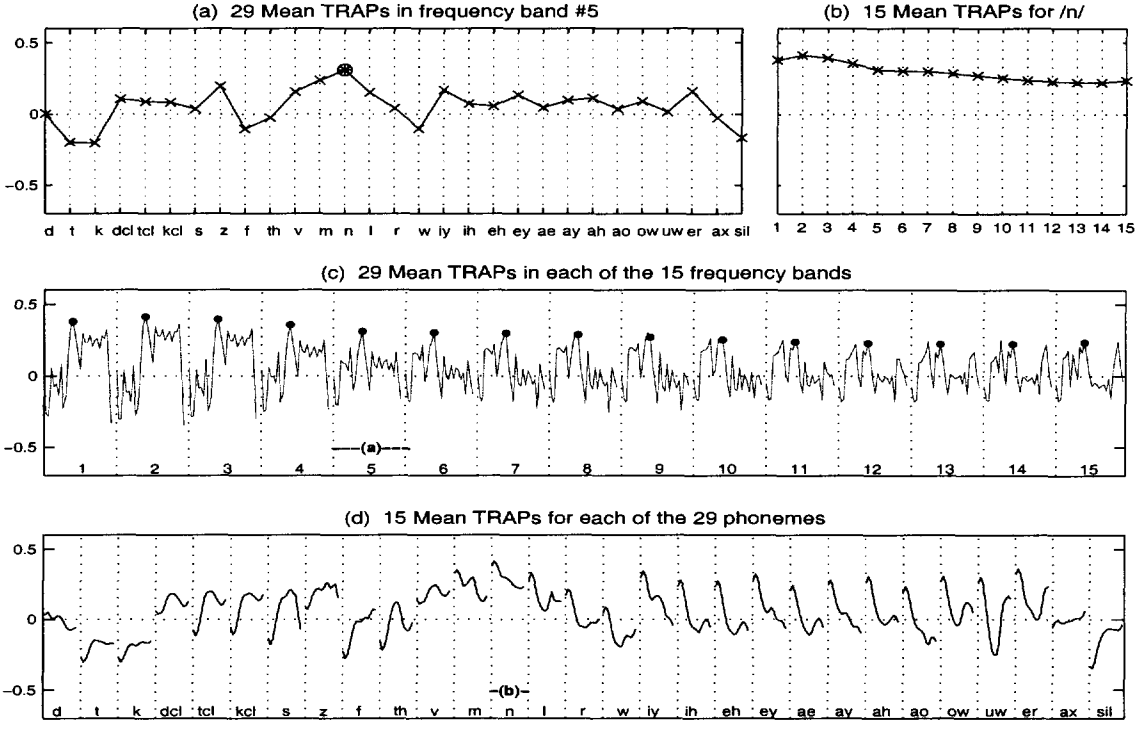


(II) **Average similarity scores for examples of /th/ with (a), (b), (c) and (d).**

Figure B.4: Comparison of the average similarity scores obtained for the 2 classes, /z/ and /th/, from the Mean TRAP classifiers.

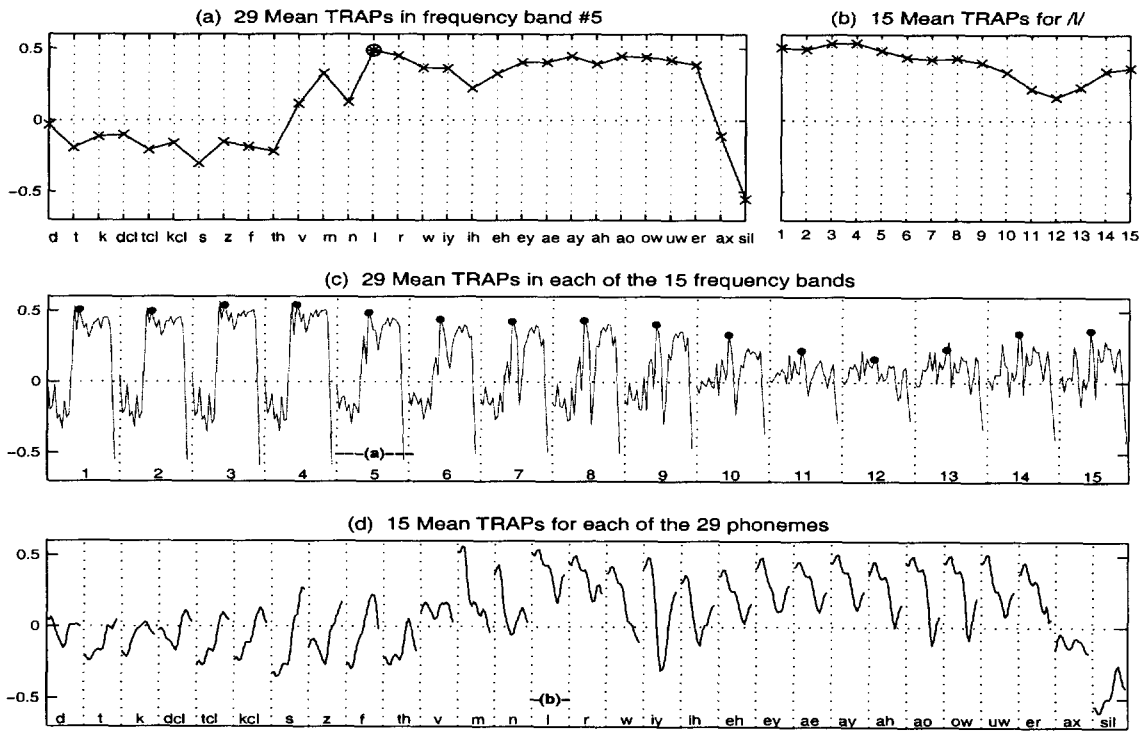


(I) Average similarity scores for examples of /v/ with (a), (b), (c) and (d).

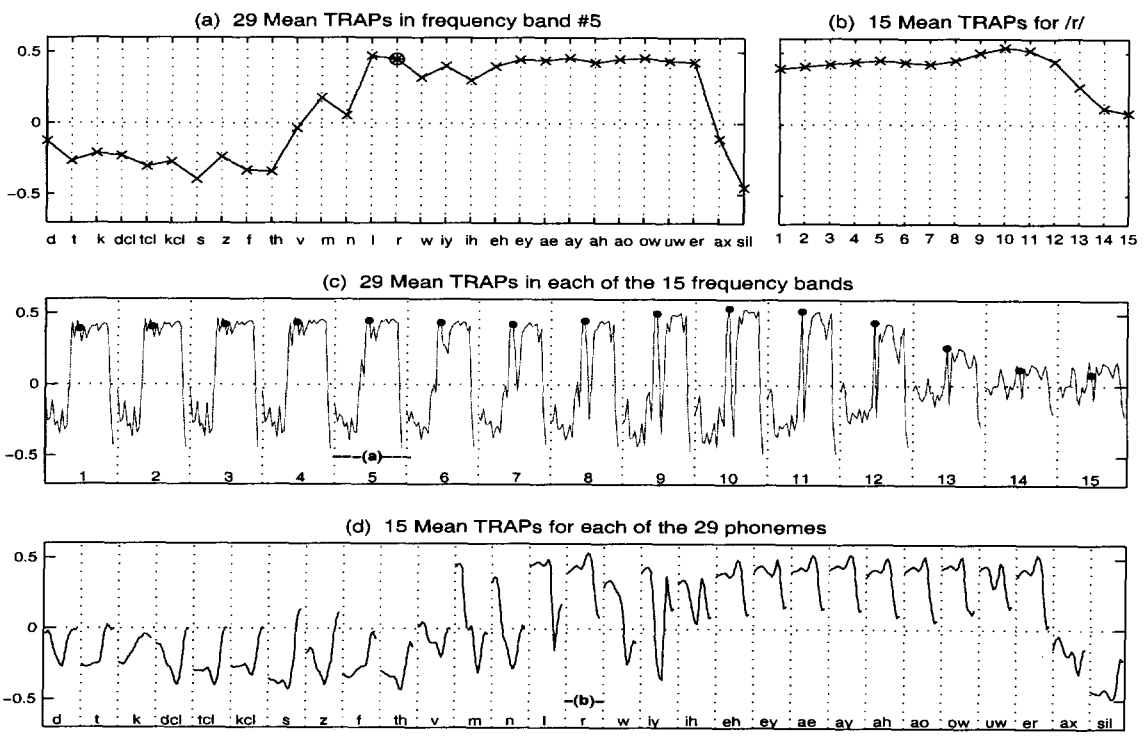


(II) Average similarity scores for examples of /n/ with (a), (b), (c) and (d).

Figure B.5: Comparison of the average similarity scores obtained for the 2 classes, /v/ and /n/, from the Mean TRAP classifiers.

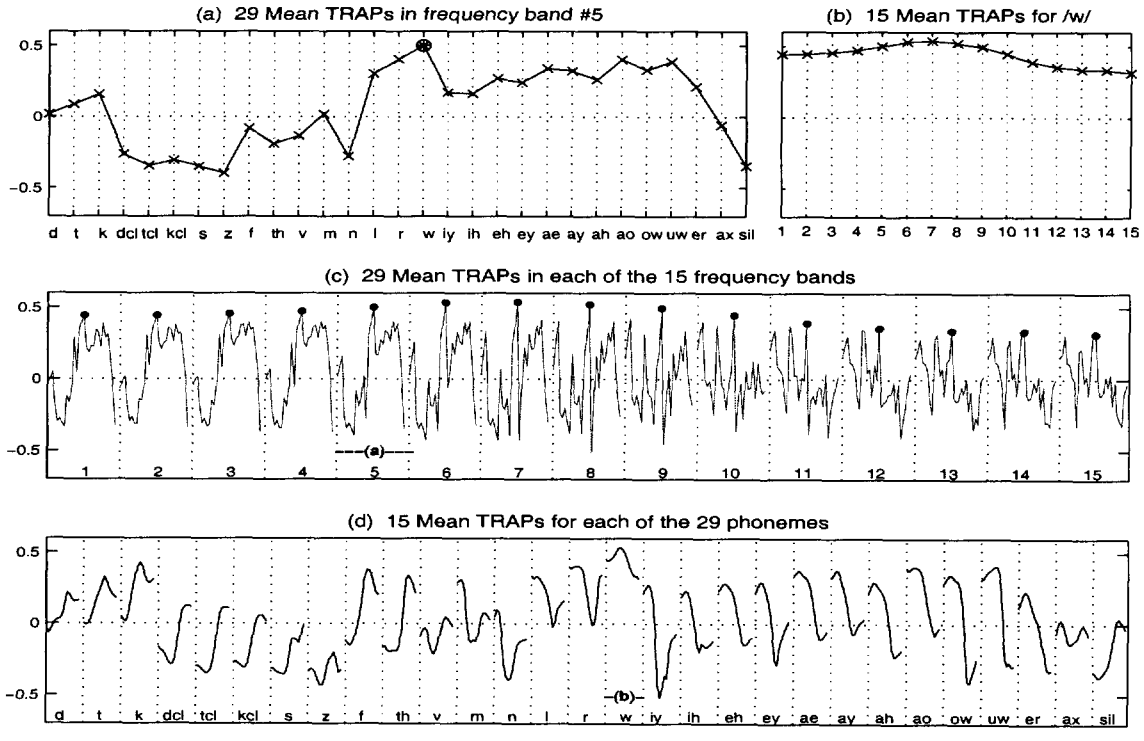


(I) **Average similarity scores for examples of /l/ with (a), (b), (c) and (d).**

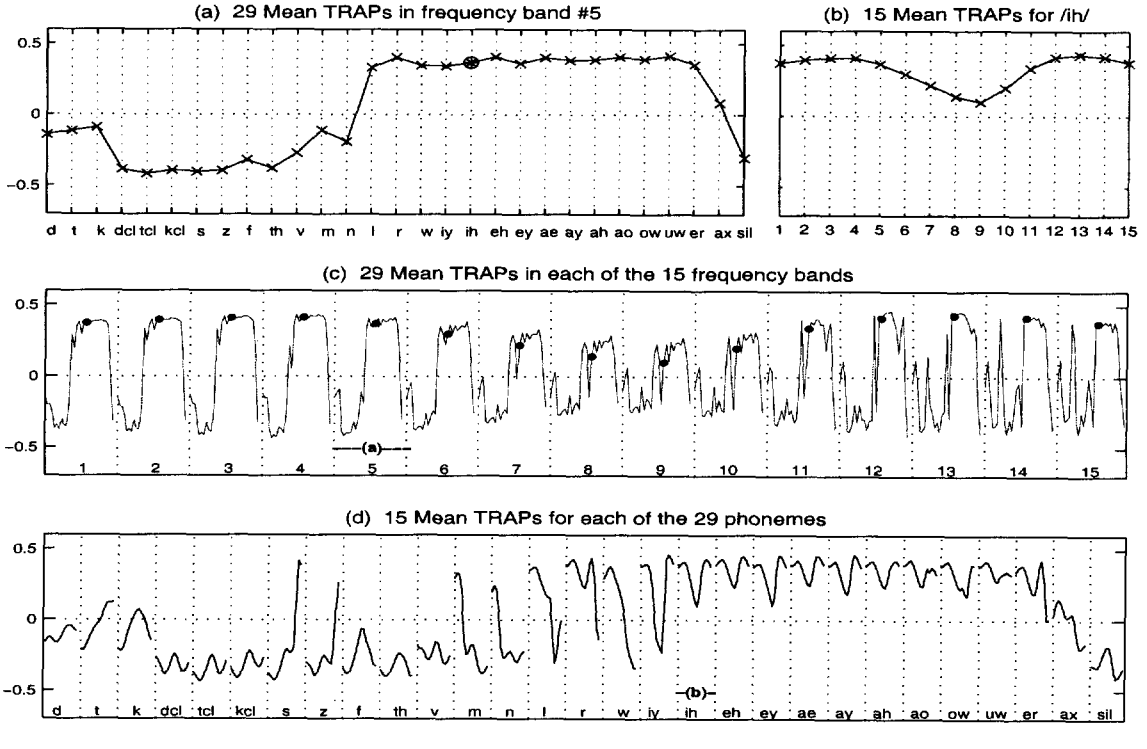


(II) **Average similarity scores for examples of /r/ with (a), (b), (c) and (d).**

Figure B.6: Comparison of the average similarity scores obtained for the 2 classes, /l/ and /r/, from the Mean TRAP classifiers.

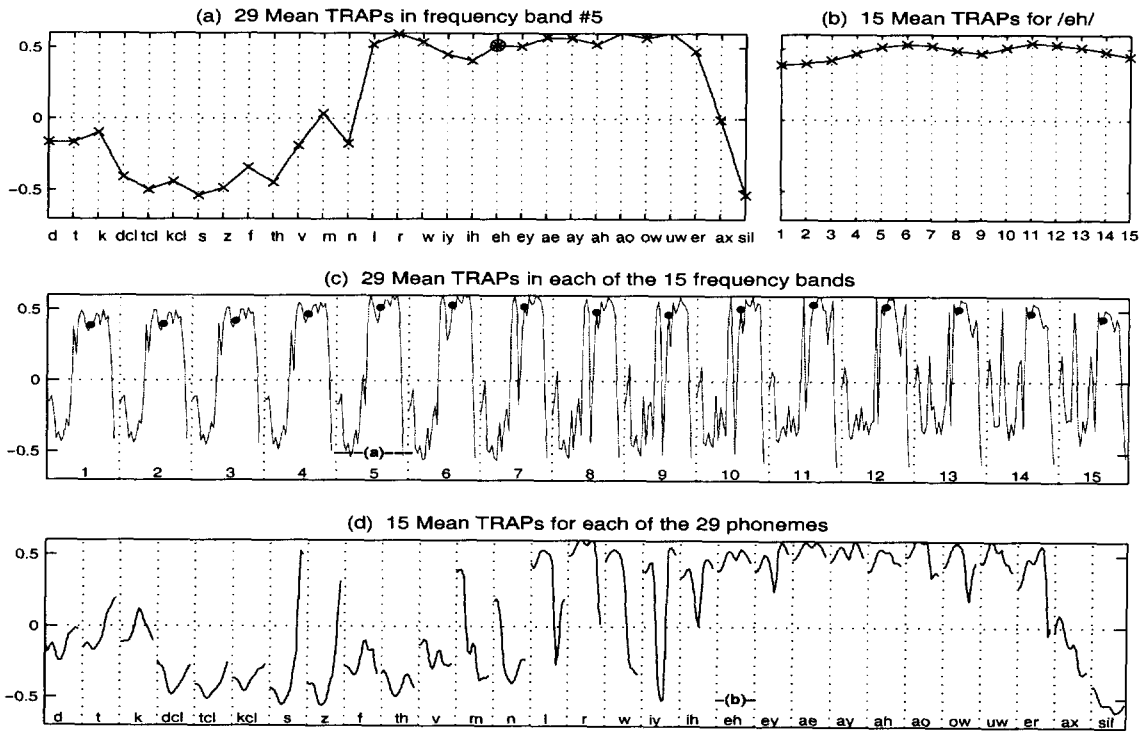


(I) **Average similarity scores for examples of /w/ with (a), (b), (c) and (d).**

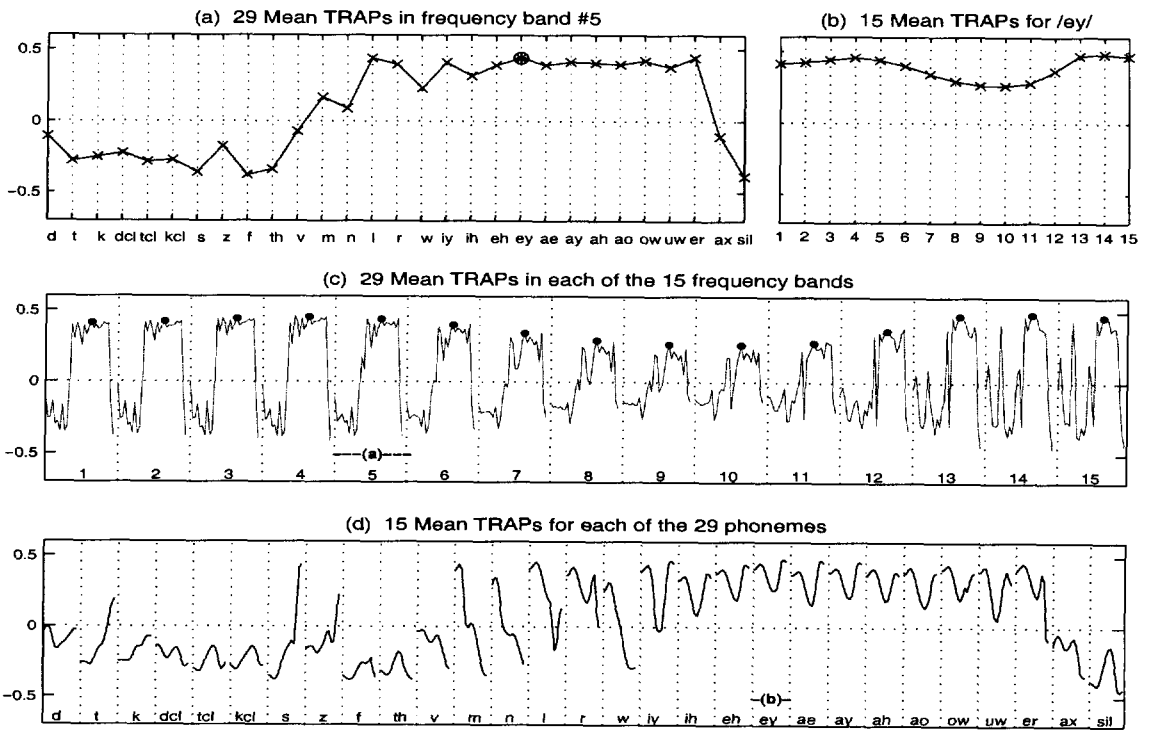


(II) **Average similarity scores for examples of /ih/ with (a), (b), (c) and (d).**

Figure B.7: Comparison of the average similarity scores obtained for the 2 classes, /w/ and /ih/, from the Mean TRAP classifiers.

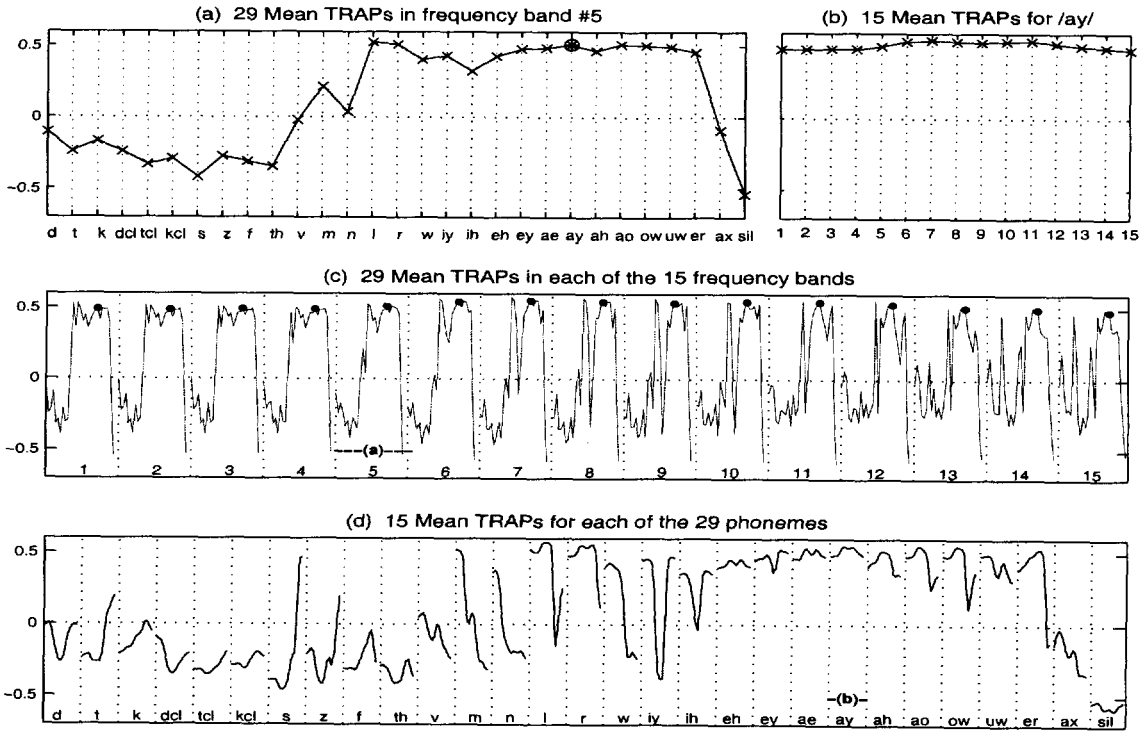


(I) **Average similarity scores for examples of /eh/ with (a), (b), (c) and (d).**

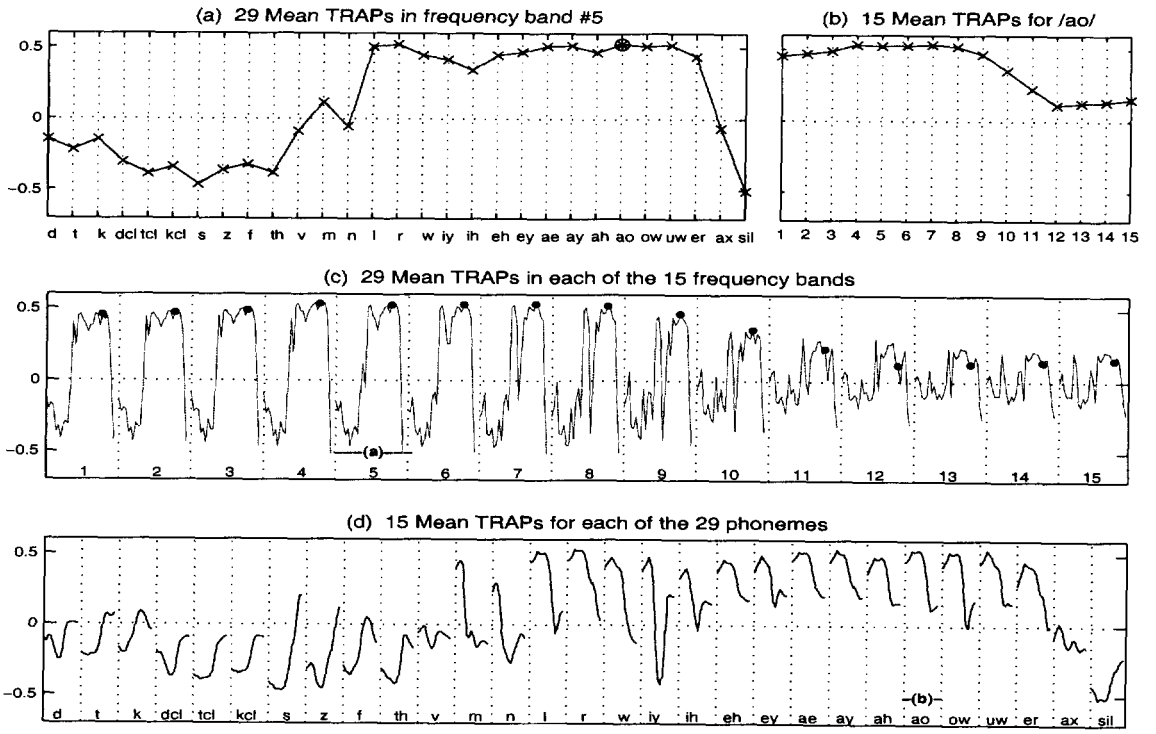


(II) **Average similarity scores for examples of /ey/ with (a), (b), (c) and (d).**

Figure B.8: Comparison of the average similarity scores obtained for the 2 classes, /eh/ and /ey/, from the Mean TRAP classifiers.

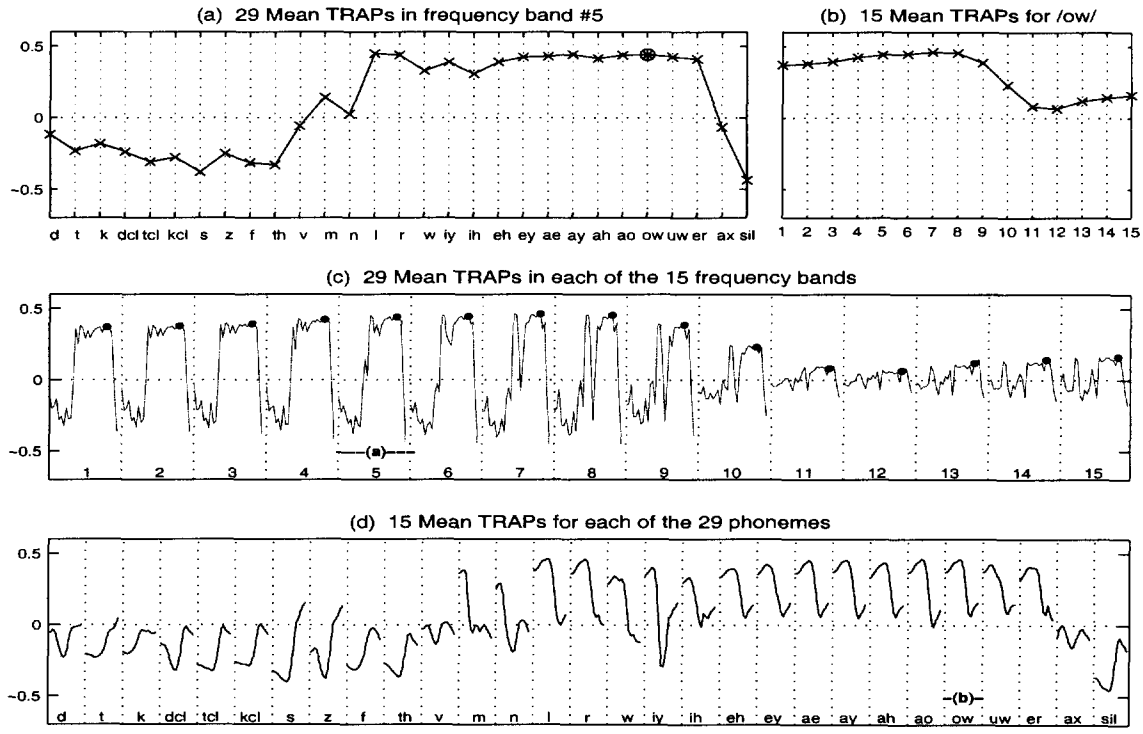


(I) **Average similarity scores for examples of /ay/ with (a), (b), (c) and (d).**

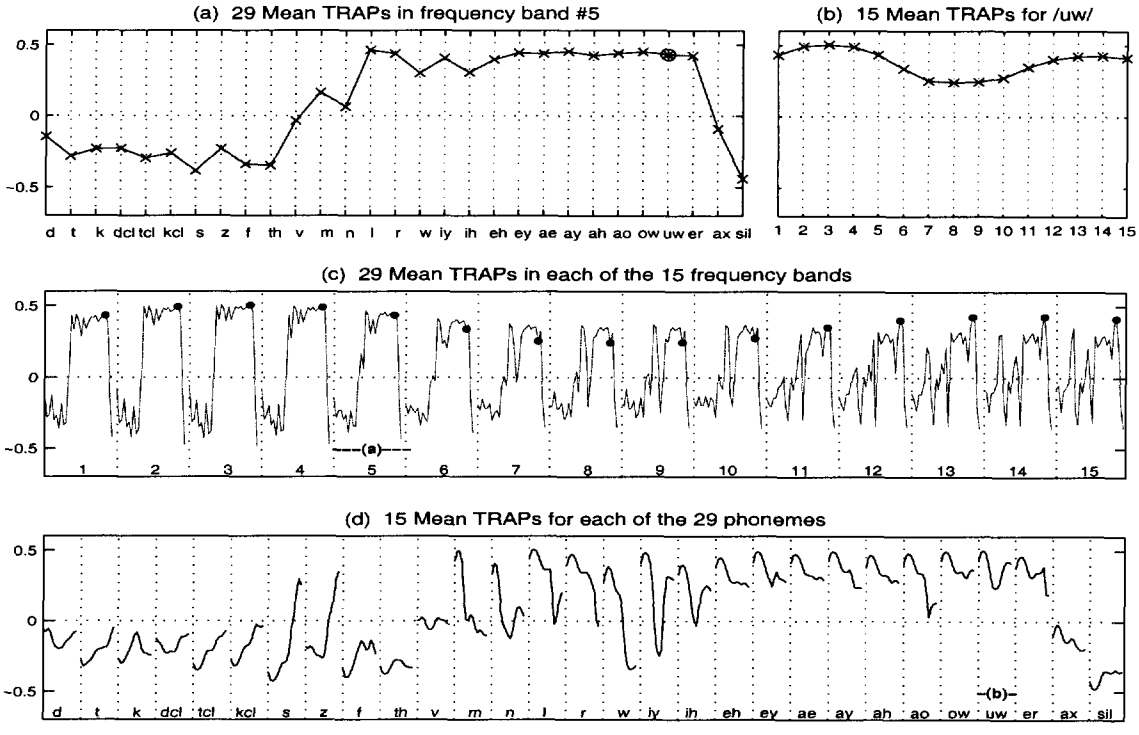


(II) **Average similarity scores for examples of /ao/ with (a), (b), (c) and (d).**

Figure B.9: Comparison of the average similarity scores obtained for the 2 classes, /ay/ and /ao/, from the Mean TRAP classifiers.

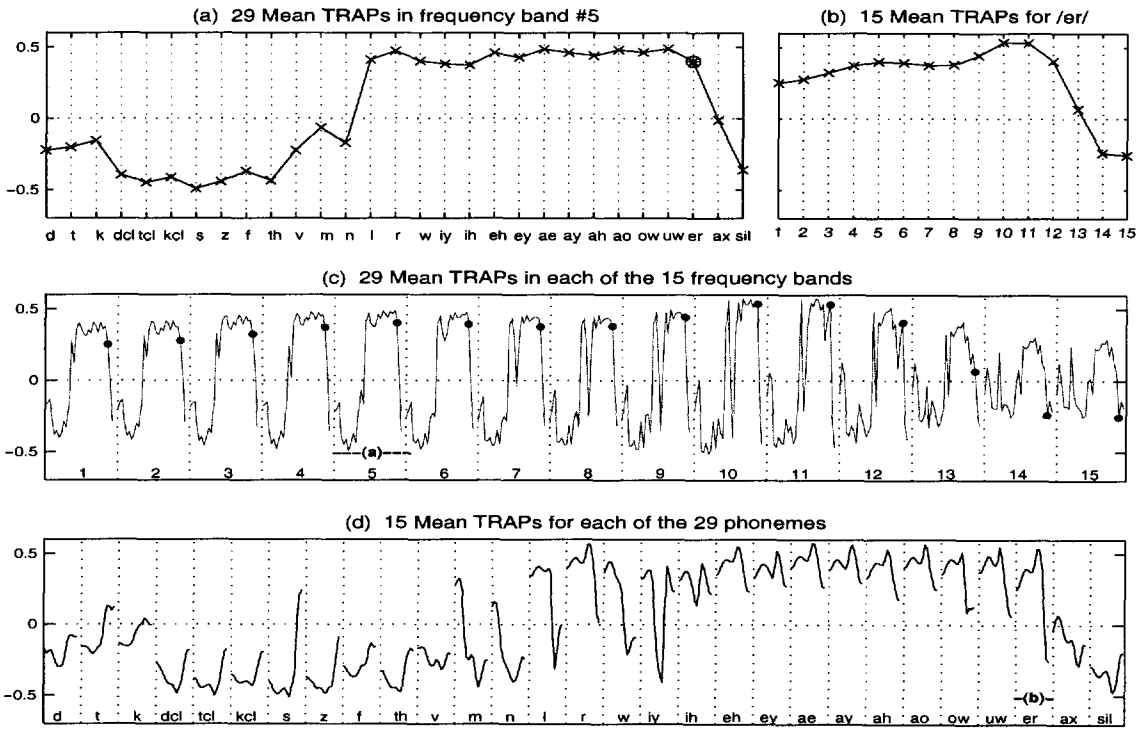


(I) Average similarity scores for examples of /ow/ with (a), (b), (c) and (d).

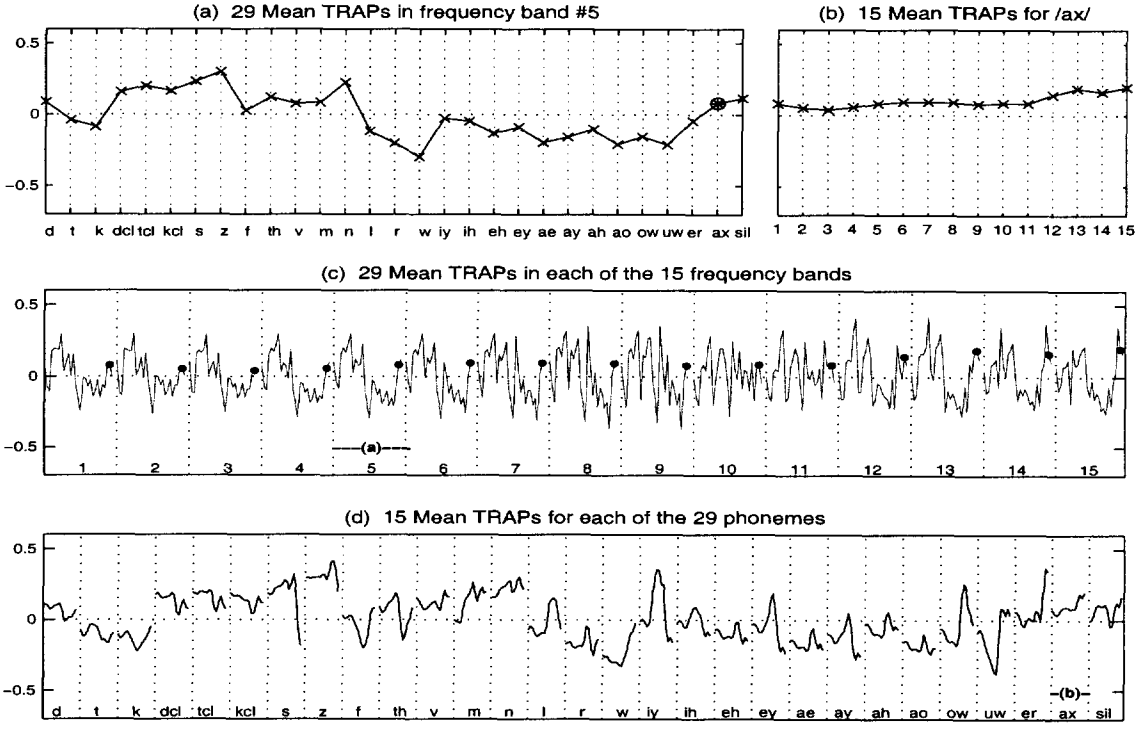


(II) Average similarity scores for examples of /uw/ with (a), (b), (c) and (d).

Figure B.10: Comparison of the average similarity scores obtained for the 2 classes, /ow/ and /uw/, from the Mean TRAP classifiers.

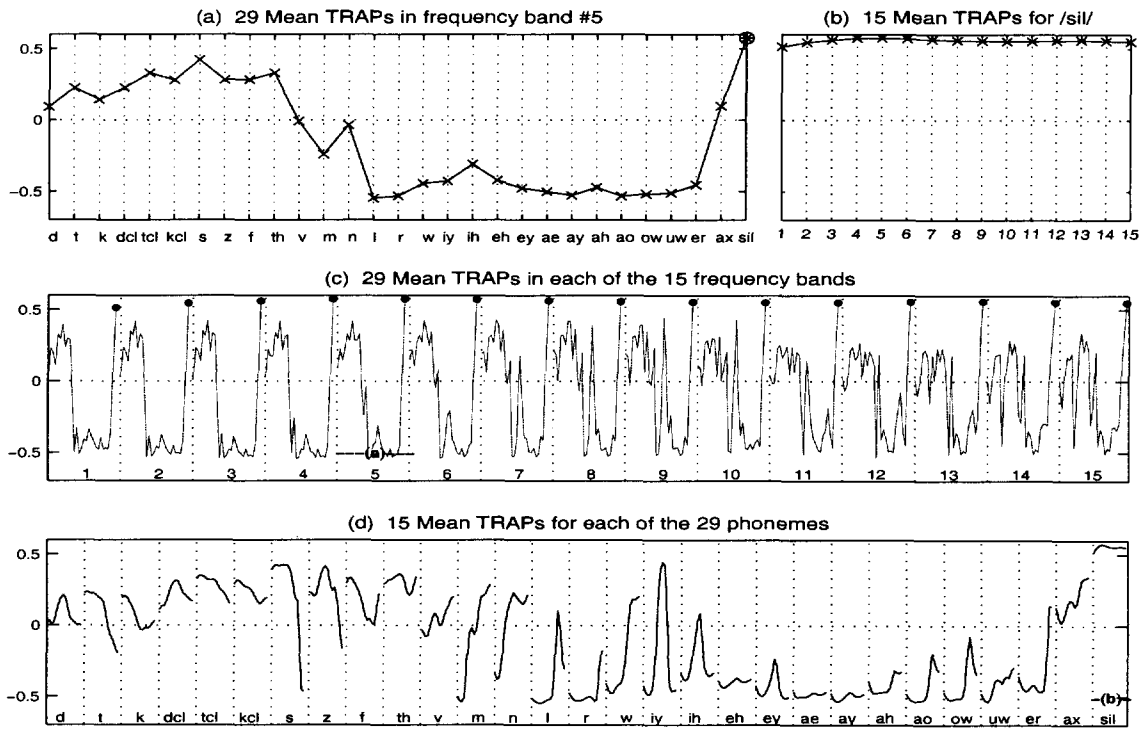


(I) Average similarity scores for examples of /er/ with (a), (b), (c) and (d).



(II) Average similarity scores for examples of /ax/ with (a), (b), (c) and (d).

Figure B.11: Comparison of the average similarity scores obtained for the 2 classes, /er/ and /ax/, from the Mean TRAP classifiers.



(I) Average similarity scores for examples of /sil/ with (a), (b), (c) and (d).

Figure B.12: Comparison of the average similarity scores obtained for the class /sil/ from the Mean TRAP classifiers.

Biographical Note

Sangita was born on May 15, 1971 in Bombay, India. She completed her Bachelor's in Electrical Engineering with honors from the University of Bombay in May 1992. Her final year undergraduate project in speech recognition as an intern at the CMS Computers Pvt. Ltd. developed her interest in the exciting field of speech processing.

She worked at Digital Equipment Corporation, India, as a Software Engineer for two years from 1992 to 1994 and learned important software development skills. However, her interest in speech persuaded her to return back to graduate studies. She enrolled in to the Ph.D. program at the Oregon Graduate Institute in Fall 1994, and has since then been a graduate research assistant under the guidance of Prof. Hynek Hermansky. She received her Master's degree in 1997.

During the summer of 1996 she was invited to participate in the summer research workshop at the Center for Language and Speech Processing, Johns Hopkins University, where she worked on new techniques for robust speech recognition, along with a group of renowned researchers from academic institutions and industry. The following year in 1997, she was a summer intern at the Media Technologies Lab at Texas Instruments where she worked on techniques for improving robustness of speech recognition in hands-free car environment, under the supervision of Dr. Yifan Gong.

During her graduate studies, Sangita has co-authored several scientific papers in international conferences. She has also applied for a United States patent. She has accepted a position in Intel Corporation.

Selected Publications

1. Hynek Hermansky and Sangita Sharma. TempoRAl Patterns (TRAPs) in ASR of noisy speech, *Proceedings of the International Conference on Acoustics,Speech and Signal Processing (ICASSP'99)*, volume 99, pages 289-292, Mar. 1999.

2. Hynek Hermansky and Sangita Sharma. TRAPs - classifiers of TempoRAl Patterns, *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, volume 3, pages 1003-1006, Dec. 1998.

3. Sangita Sharma, Pieter Vermeulen and Hynek Hermansky. Combining information from multiple classifiers for speaker verification, *Proceedings of the Speaker Recognition and its Commercial and Forensic Applications Workshop (RLA2C)*, pages 115-119, Apr. 1998.

4. Sangita Tibrewala and Hynek Hermansky. Multi-band and Adaptation approaches to robust speech recognition, *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, volume 5, pages 2619-2622, Sep. 1997.

5. Carlos Avendano, Sangita Tibrewala and Hynek Hermansky. Multiresolution channel normalization for ASR in reverberant environments, *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, volume 3, pages 1107-1110, Sep. 1997.

6. Sangita Tibrewala and Hynek Hermansky. Sub-band based recognition of noisy speech, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume II, pages 1255-1258, Apr. 1997.

7. Hynek Hermansky, Carlos Avendano, Sarel van Vuuren and Sangita Tibrewala. Recent Advances in Addressing Sources of Non-Linguistic Information, *ECSA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 103-106, Apr. 1997.

8. Hynek Hermansky, Sangita Tibrewala and Misha Pavel. Towards ASR on partially corrupted speech, *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 462-465, Oct. 1996.