Temporal patterns of frequency-localized features in ASR

Pratibha Jain
B.Tech., Govt. Engineering College, Jabalpur, 1995
M.Tech., Indian Institute of Technology, Kanpur, 1997

A dissertation submitted to the faculty of the
OGI School of Science and Engineering
Oregon Health and Science University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Electrical and Computer Engineering

July 2003

The dissertation "Temporal patterns of frequency-localized features in ASR" by Pratibha Jain has been examined and approved by the following Examination Committee:

Hynek Hermansky
Professor
OGI School of Science and Engineering, OHSU
Thesis Research Adviser

Paul Hosom
Assistant Professor
OGI School of Science and Engineering, OHSU

Jont B. Allen
Prof. ECE,
Beckman Inst.
University of IL, Urbana Champaign

Brian Kingsbury
Research Staff member
Human Language Technologies Dept.
IBM T. J. Watson Research Center

Misha Pavel
Professor
OGI School of Science and Engineering, OHSU

Dedication

To my parents and Suyog

Acknowledgments

My doctoral study was successful because of help and motivation from many people. First, I would like to thank my thesis adviser, Prof. Hynek Hermansky, for the discussions and feedback on this work. I also want to thank him for giving me the opportunity to work with him and for providing financial support throughout this work. His advice regarding about presentation skills, writing skills, and research methodology were of great help.

I want to thank my thesis committee members – Jont B Allen, Brian Kingsbury, Paul Hosom, and Prof. Misha Pavel. Their feedback on the thesis has significantly improved the quality of this work.

During the doctoral work, I was fortunate to learn about various topics from different professors – Hynek Hermansky, Michael Macon, Paul Hosom, Robert Jaffe, Todd Leen, and Andrew Fraser . I want to thank them for broadening my perspective of research.

I would like to thank Mukund Padmanabhan who gave me a wonderful opportunity to do summer-internship in IBM T.J. Watson in year 2001. He was constantly a source of encouragement and support during my work in IBM. My special thanks to Brian Kingsbury, who mentored me during my internship in IBM T. J. Watson. He was of great help in carrying out many important experiments and in many useful discussions. I always enjoyed working with him.

I was fortunate to work with wonderful people in my group and also in Center for Spoken Language Understanding. First i want to thank Sangita for reading the very first lousy draft of my thesis. She was quite patient in reading each chapter very carefully in spite of her work-load. She was always there to help and support me during my stay in OGI. I would like to thank Ravi for his useful advices for a good presentation. My special thanks to Sachin for not only for technical discussions but also for bearing with me during our stay at Apt 26 for almost three years. I would like to thank Sunil for being caring and

helpful during my stay in OGI. I also thank Pavel and Alena for their help in computing and administration facilities. I want to thank Andre for reviewing the abstract of my thesis and being a great office-mate. Alex Nelson needs a special mention for his help in correcting the style and written English.

I would like to thank my ex-labmates Lukas, Franta, and Petr for making my stay at OGI memorable. I will never forget going to Taco Bell late in the nights with them. I also would like to thank Honza (Jan Cernoscky) for useful technical discussions and for his invitation to visit his univ of Brno in Czech Republic. I would also like to thank Akiko who visited our lab for an year, with whom i enjoyed many moments in OGI.

I would like to thank my friends, Nitu Choudhary and Abhijit Bhattacharaya who were of great emotional support during initial hard-ships in USA. They always inspired me for being remain cheerful even in hard circumstances.

I want to thank my brother and my family for supporting my decision to pursue doctoral studies. They have always been a constant source of help and motivation.

My great gratitude to Suyog, my soulmate, who's constant encouragement and support over the years kept me remain enthusiastic in many down-turns. I really appreciate his waiting so long for me. I would like to thank him for keeping a great patience for my finishing thesis work.

Contents

Dedication iv

Acknowledgments v

Abstract xv

1 Introduction 1
- 1.1 General framework for Automatic Speech Recognition Systems 2
- 1.2 Feature extraction Process 2
  - 1.2.1 Spectral processing for short-term spectral features 3
  - 1.2.2 Temporal processing of short-term spectral features 5
- 1.3 The TempoRAl Patterns based (TRAPS) approach for feature estimation . 6
- 1.4 The Tandem approach to speech recognition 7
- 1.5 Contributions of this work 8
- 1.6 Organization of the thesis 12

2 Parametric representation of temporal patterns of features 14
- 2.1 Temporal patterns as Markov-1 vector sequence 15
- 2.2 Linear transforms for parameterization 16
- 2.3 Databases 19
  - 2.3.1 Testing conditions at different SNRs 19
- 2.4 Performance of linearly transformed temporal patterns 21
- 2.5 On low-pass filtering the Modulation spectra 22
- 2.6 Conclusion 22

3 Band-specific broad-phonetic category features in ASR 24
- 3.1 Earlier work on broad-phonetic features 25
- 3.2 Broad-phonetic classes from temporal patterns 26
- 3.3 Broad-phonetic classes from confusion matrix of phones 28
- 3.4 Estimation of broad-phonetic category features 29
- 3.5 Analyzing confusion matrix of broad-categories 29

3.6 Phone classification using broad-category features . . . 30
3.7 Broad-phonetic category features in ASR . . . 30
3.8 Conclusions . . . 34

4 Broad-phonetic features : Applications . . . 36
4.1 Broad-phonetic features for Voice Activity Detection (VAD) . . . 36
4.1.1 Voice activity detection (VAD) . . . 37
4.1.2 Frequency-localized temporal patterns for VAD . . . 37
4.1.3 GMM/HMM based VAD . . . 38
4.1.4 The Database and Task description . . . 39
4.1.5 Results and Conclusions . . . 41
4.2 Broad-phonetic features in cellular telephony for DSR . . . 42
4.2.1 Distributed Speech Recognition (DSR) for cellular telephony speech 42
4.2.2 Handset-side processing . . . 43
4.2.3 Server-side Processing . . . 44
4.2.4 Aurora Dataset . . . 44
4.2.5 Broad phonetic features estimated from temporal patterns in DSR . 46
4.2.6 Conclusions . . . 49

5 Universal TempoRAl PatternS (UTRAPS) based system . . . 50
5.1 Clustering the temporal patterns to define speech events . . . 51
5.2 A Universal TemPoRAl PatternS (UTRAPS) system . . . 53
5.3 Speech-event for phone and broad-phonetic category classification . . . 54
5.4 Speech-event features in ASR . . . 55
5.5 Summary and Conclusion . . . 58

6 Incorporating broader-frequency context . . . 62
6.1 On incorporating broader frequency context . . . 63
6.1.1 Evidence of across-channel processing . . . 63
6.1.2 Computation of joint PCAs . . . 64
6.2 Experiment and Results . . . 65
6.3 Replacing Joint PCAs by DCT bases . . . 67
6.4 Summary . . . 68

7 Summary . . . 72
7.1 Summary . . . 72
7.2 Future Directions . . . 74
7.3 Final Comments . . . 75

A First-order Markov sequences **76**

B Mean Temporal patterns **78**

C Phone Set **83**

D Confusion matrices for phones **84**

E Clusters of speech events **89**

F Code for generating Speech events **92**

G Linear Transforms of temporal patterns **94**

G.1 Linear transform from Principal Component Analysis (PCA) 94

G.2 Linear transform from Linear Discriminant Analysis (LDA) 95

H Computation of PLP critical-bands and Mel critical-bands **97**

Bibliography **100**

Biographical Note **106**

List of Tables

2.1 % Word Error Rate (Recognition Performance) on using different linear transforms on temporal patterns of log critical-band energies at clean condition 22

3.1 Seven broad-phonetic categories used for feature estimation from temporal patterns in individual critical-bands 28
3.2 Frame-level broad-phonetic categories error rate (%) on TIMIT dataset. . . 33
3.3 Frame-level phone classification error (%) and WER on OGI connected digits task using $6 * 15$ broad-phonetic category features estimated from each band. 33
3.4 Word Error Rates (%) on OGI-Digits, using cepstral features alone, or using cepstral augmented with temporal patterns estimated features. Band-specific posterior estimators (MLPs) were trained on OGI-Stories and the posterior estimator used in the Tandem system was trained on the OGI-Numbers. The dimensionality of final feature vector was kept at 29 33
3.5 Word Error Rates (%) on OGI-Digits task, using cepstral features alone or cepstral augmented with temporal patterns estimated features. The band-specific posterior estimators (MLPs) and the posterior estimator used in Tandem system were trained on OGI-Stories, an independent dataset. The dimensionality of final feature vector was kept at 29 33

4.1 % Recognition performance for each segmentation of the SPINE-2 evaluation data and each pass of decoding. The word error rate is broken down in terms of substitutions, deletions, insertions and gap insertions. Gap insertions are caused by segmentation errors: the recognizer hypothesizes words during non-speech segments. For all three decoding passes, the best accuracy is obtained with the TRAPS+Viterbi segmentation. 42
4.2 SAT-n recognition performance for SPINE-2 evaluation data, by noise condition and segmentation. The table is divided into two parts corresponding two speaker groups in the data. Because there are only sixteen speakers per group, meaningful comparisons can only be made within a speaker group. 43

4.3 SAT-n recognition performance for SPINE-2 evaluation data, by noise condition and segmentation. The table is divided into two parts corresponding two speaker groups in the data. Because there are only sixteen speakers per group, meaningful comparisons can only be made within a speaker group. 44
4.4 Word error rates (%) on Aurora-2, TIDIGITS data, by training data (multi-style vs. clean) and by feature set (noise-robust MFCCs alone,or robust MFCCs augmented with TRAPS-estimated broad-phonetic features, context window = 190 ms). 48
4.5 Word error rates (%) on Aurora-3, SpeechDat-car data, by testing conditions by feature set (noise-robust MFCCs alone,or robust MFCCs augmented with TRAPS-estimated broad-phonetic features, context window = 1 s). 48
4.6 Word error rates (%) on Aurora-2, TIDIGITS data, by training data (multi-style vs. clean) and by feature set (noise-robust MFCCs alone,or robust MFCCs augmented with TRAPS-estimated broad-phonetic features, context window = 1 s). 48

5.1 % Frame-level phone classification error rates obtained at the output of the Tandem MLP, using the UTRAPS features and using broad-phonetic TRAPS features estimated independently from individual critical-bands. Results are obtained on the TIMIT dataset. The minimum error is shown in bold type 56
5.2 % Frame-level broad-phonetic classification error rates obtained at the output of the Tandem MLP, using the UTRAPS features and using broad-phonetic features estimated independently from individual critical-bands. Results are obtained on the TIMIT dataset 59
5.3 Word Error Rates (%) on OGI-Digits : PLP cepstral features, the UTRAPS features, or the broad-phonetic TRAPS features. 59
5.4 Word Error Rates (%) on OGI-Digits : PLP cepstral features alone, PLP cepstra augmented with UTRAPS features, or cepstra augmented with various TRAPS features. 59
5.5 Word Error Rates (%) on OGI-Digits : PLP cepstral features alone, PLP cepstra augmented with the UTRAPS features, or PLP cepstra augmented with various TRAPS features. In this case, MLPs training was entirely done on an independent dataset, OGI-Stories. 59

5.6 Word error rates (%) on the Aurora-2, TIDIGITs data and by feature set (noise-robust MFCCs alone, robust MFCCs augmented with broad-phonetic TRAPS features or the UTRAPS features), with context window = 1 s. 60
5.7 Word error rates (%) on the Aurora-3, by testing conditions and by feature set (noise-robust MFCCs alone, robust MFCCs augmented with broad-phonetic TRAPS features or robust MFCCs augmented with the UTRAPS features), with context window = 1 s. 60

6.1 % Word Error Rate (Recognition Performance) on using joint PCAs on time-frequency plane of speech representation for feature estimation at clean condition 67
6.2 % Word Error Rate on using 76 joint DCTs on the temporal patterns of three adjacent critical-bands, 75 joint 3-bands PCAs, 50 1-band DCT in clean, and in noisy condition 67
6.3 % Frame-level phone error rates obtained using $1-5$ critical-bands of frequency context. The minimum error is shown in bold type. 68

C.1 The set of 56 phones in TIMIT. 83
C.2 The set of 29 phones in OGI-Numbers. 83

E.1 Clusters in individual critical-bands 91

H.1 Cut-off frequencies of the 1-Bark spacing critical band filters used in PLP feature representation for 8 kHz sampling frequency. 97
H.2 Cut-off frequencies of the used Mel critical band filters for 8 kHz sampling frequency. 98

List of Figures

1.1 Block diagram of speech recognition system 3
1.2 The TempoRAl PatternS (TRAPS) approach for feature estimation 7
1.3 Block diagram of Tandem speech recognition system 9
1.4 Combining multiple feature streams in Tandem speech recognition system . 9

2.1 Correlation matrix of 101 point long temporal pattern of log critical-band energies 16
2.2 Principal components basis vectors 16
2.3 % of total variability in number of retained PCA basis vectors 16
2.4 Effect of low-pass filtering the modulation spectrum at different SNRs by limiting DCT bases in White noise 16
2.5 Effect of low-pass filtering the modulation spectrum at different SNRs by limiting DCT bases in Pink noise 17
2.6 Effect of low-pass filtering the modulation spectrum at different SNRs by limiting DCT bases in Babble noise 17
2.7 Word error rates from different linear transforms at different SNR conditions in White noise 17
2.8 % Word error rates from different linear transform at different SNR conditions in Pink noise 17
2.9 % Word error rates from different linear transforms at different SNR conditions in Babble noise 18
2.10 Block diagram of temporal patterns based feature estimation 20

3.1 Confusion matrix in 1st band 31
3.2 Confusion matrix in 5th band 31
3.3 Confusion matrix in 15th band 31
3.4 Confusion matrix at the Tandem MLP output 31
3.5 Mean Temporal Patterns of seven broad categories in 1st critical-band . . . 35
3.6 Mean Temporal Patterns of seven broad categories in 5th critical-band . . 35
3.7 Mean Temporal Patterns of seven broad categories in 15th critical-band . . 35

4.1 Frequency-localized temporal patterns for VAD 38

4.2 Distributed Speech Recognition (DSR) system 45

5.1 Mean temporal patterns of nine speech-events 53
5.2 Block diagram of the UTRAPS system 54

6.1 Effect of broadening frequency context on ASR using TRAPS features . . . 69
6.2 Effect of broadening frequency context on ASR using TRAPS features . . . 69
6.3 % Effect of broadening frequency context on ASR using TRAPS features . 69
6.4 Effect of broadening frequency context on ASR using TRAPS features, At 0 dB SNR conditions 69
6.5 First 8 PCAs over 3 critical-bands and 1 s (101-samples) temporal context . 70
6.6 Some of the next 6 PCAs over 3 critical-bands and 1 s (101-samples) temporal context 71

B.1 Mean Temporal Patterns for a front vowel /iy/ 79
B.2 Mean Temporal Patterns for a fricative /sh/ 79
B.3 Mean Temporal Patterns for silence /sil/ 80
B.4 Mean Temporal Patterns for a glide /l/ 80
B.5 Mean Temporal Patterns for a vowel /ae/ 81
B.6 Mean Temporal Patterns for a plosive /b/ 81
B.7 Mean Temporal Patterns for a schwa /axr/ 81
B.8 Mean Temporal Patterns for a glide /m/ 82

D.1 For 56 TIMIT phones in the 1st critical-band, 115 – 265 Hz 85
D.2 For 56 TIMIT phones in the 5th critical-band, 4.538 – 629 Hz 86
D.3 For 56 TIMIT phones in the 15th critical-band, 2962 – 3769 Hz 87
D.4 For 56 TIMIT phones at the Tandem outputs 88

Abstract

Temporal patterns of frequency-localized features in ASR

Pratibha Jain

Supervising Professor: Hynek Hermansky

This work investigates the use of frequency-localized temporal patterns of the speech signal for developing robust front-end for Automatic Speech Recognition (ASR). Various linear transforms are investigated for parameterization of the frequency-localized temporal patterns. We show that temporal patterns closely follow the properties of a first-order Markov process, which results in the PCA transforms being very close to the DCT transform. Better recognition performance is achieved on using the DCT components of temporal patterns as opposed to directly using temporal patterns for feature estimation. Other linear transforms such as Linear Discriminant Analysis (LDA) are also studied for the parameterization. The parameterized TempoRAl PatternS (TRAPS) are used to estimate broad-phonetic class-posteriors independently in each critical-band. These class-posteriors are combined and used as the features for word recognition. Our work shows that broad-phonetic features generalize better than other conventional features and yield considerable complementary information with respect to short-term cepstral features in ASR. Two practical applications are proposed for the broad-phonetic TRAPS features : 1) Distributed Speech Recognition (DSR) in cellular telephony, 2) Voice Activity Detection (VAD) tasks. These features yield a significant improvement in the performance for these applications. New band-independent categories are proposed which represent distinct

speech-events in the frequency-localized temporal patterns of the speech signal. These categories are obtained by clustering the mean temporal patterns of context-independent phones using an agglomerative hierarchical clustering technique. A Universal TempoRAl PatternS (UTRAPS) system is proposed for the speech-event class-posteriors estimation. Combining UTRAPS features with cepstral features achieves a significant improvement in the recognition performance under noisy conditions. Finally, this work studies the effect of broadening the frequency-context on TRAPS features and ASR. This study shows that combining temporal patterns from more than one critical-band is important to achieve higher recognition rates.

Chapter 1

Introduction

The goal of automatic speech recognition (ASR) is to interact with machines via voice commands and provide a mean of natural interface. The major problem with the current ASR systems is their limited ability to deal with the diversity and variability of real-world environments. For example, the performance of an ASR system degrades drastically in the presence of different speakers, different speech-context, communication channels, and background noises, i.e., in environments not seen by the system during its training [9]. Several studies have been carried out to develop noise-robust ASR systems that can perform well in different real-world environments [22]. This includes extraction of various robust front-end features from the speech signal that are relatively insensitive to operational environments [2, 25, 28, 49]. Another approach that has been studied is the use of various adaptation techniques that can adapt an ASR system (at various levels such as at the feature-level or the model-level) to real-world operating conditions [32, 43, 70].

The objective of this dissertation is to investigate frequency-localized temporal patterns of the speech signal for developing noise-robust front-end for ASR systems.

This chapter is organized as follows. Section 1.1 gives a general frame-work of ASR systems. The first step in ASR systems is the estimation of useful features or measurements from the speech signal for speech recognition. The different approaches for feature extraction are described in detail in Section 1.2. Section 1.3 describes a recently proposed approach for robust feature estimation. The work proposed in this dissertation mainly expands, and explores upon on this approach for feature estimation. In this technique temporal patterns of frequency-localized features are used for robust feature detection for ASR. The chapter concludes with contributions of the thesis in Section 1.5 and an

overview of the thesis in Section 1.6.

1.1 General framework for Automatic Speech Recognition Systems

Automatic speech recognition is a process which maps the speech signal to a sequence of speech sub-units such as phones, syllables, words. The mapping is based on statistical pattern recognition techniques [18]. The general framework of such a system is shown in Figure 1.1. The recognition process comprises three main processing steps - feature extraction, likelihood computation, and search for the most likely sub-unit sequence such as word sequence. In the feature extraction step, the incoming signal is typically divided into short-term segments (typically 20 to 32 ms) equally spaced in time, and a feature vector is extracted from each segment. This is followed by the computation of log-likelihoods or posterior probabilities of feature vectors with respect to the speech sub-units. The speech sub-units are modeled as a sequence of hidden states. The states corresponds to a particular part of speech sub-units. Each state is modeled as mixture of Gaussians. Neural Networks (NN) also have been used for modeling the states [24]. The mathematic structure known as Hidden Markov Model (HMM) [38, 42] is utilized to map sequence of acoustic features to the sequence of speech sub-unit log-likelihoods. Given the sequence of log-likelihoods, finally the recognition process performs a search over all possible sub-unit sequences to obtain the the highest likely or probable sequence. The Viterbi algorithm is commonly used to perform the search [42].

1.2 Feature extraction Process

The major problem for high-accuracy recognition is the large variability in the speech signal characteristics. This includes variability due to different pitch, variability due to different speakers, the effect of background noises, and the transmission channel (e.g., microphone, telephone) etc. In ASR, the goal of the feature extraction process is to reduce these aspects of signal variability and extract a robust representation of the speech signal. The feature extraction process can be divided into two parts : 1) short-term spectral

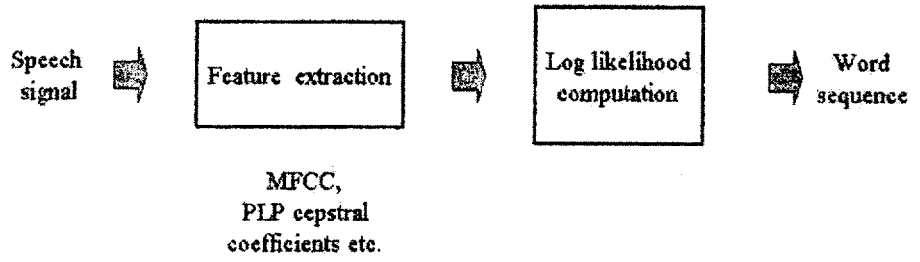

Figure 1.1: Block diagram of speech recognition system

processing, and 2) long-term temporal processing. The spectral processing extracts short-term (typically $20 - 32$ ms) spectral features from the speech signal. The short-term spectral features are temporally processed using various techniques. In this section, we first describe conventional spectral processing techniques. Next we describe some of the commonly used techniques for temporal processing.

1.2.1 Spectral processing for short-term spectral features

The speech signal is divided into overlapping short (typically $20-32$ ms or $160-256$ samples at 8 kHz) segments. The adjacent segments overlap by $10-20$ ms ie $80-160$ samples. The segments are multiplied by a Hamming window to minimize the discontinuities at the edges of the short segments. A short-time Fourier transform is applied to convert these segments to the frequency domain. Only the magnitude of Fourier transformed segments is kept while phase is ignored. For a speech signal sampled at $f_s = 8$ kHz this operation results in around 129 parameters. The resulting parameter vector corresponds to frequencies that are uniformly spaced from $0 - 4000$ Hz. The parameter vector is referred to as the short-term spectral density of the speech signal. This operation can be expressed mathematically as,

$$X_m(k) = |\sum_{n=-\infty}^{n=\infty} s_m(n) e^{-j\frac{2\pi k}{K}n}|$$

Where $s_m(n)$ is given by,

$$s_m(n) = h(mM - n) * x(n)$$

Where $x(n)$ refers to the speech signal, $s_m(n)$ refers to windowed speech signal centered at mM sample, h refers to the Hamming window which acts as a sliding analysis window, M refers to the amount of frame shift in samples, $X_m(k)$ refers to the short-term spectral density (sampled in frequency at frequencies $w_k = \frac{2\pi k}{K}, k = 0, 1, \cdots, K-1$) at time instant $n = mM/f_s$ sec, and m is an integer variable. The analysis of short-term spectrum of speech signal is more discussed in [3, 7, 6, 58].

Human perceptual studies indicate that human hearing has a non-uniform frequency resolution. To emulate the human processing, non-uniform filterbanks based on non-uniform frequency scales such as the Mel [63, 66] or Bark scales [15, 25] have been proposed to warp the short-term spectrum into a non-uniform frequency spectrum. The magnitude coefficients of the short-term spectrum are binned by correlating them with each non-uniform filter. Here binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filter. This operation can be expressed mathematically as,

$$P_m(i) = \sum_{k=0}^{k=Bandwidth} X_m(k) FB(i,k)$$

Where i refers to the filterbank index, $FB(i,k)$ is the values of k^{th} filter coefficient of the ith filter in the frequency domain, and $P_m(i)$ refers to the value of ith filter energy at time instant $n = mM/f_s$ sec.

Next, a compressive non-linear transformation such as log or cube-root is applied to the filterbank energies, and finally these energies are projected on a discrete cosine basis. These features form the basis of the conventional front-end used in ASR systems. Example of features based on this technique includes Mel Cepstral Coefficients (MFCC) [40].

1.2.2 Temporal processing of short-term spectral features

The short-term spectral features have the drawback that they are sensitive to changes in the communication environment, such as characteristics of different channels or background noise. As a result, the performance of recognizers that uses only these features rapidly degrades in realistic communication environments. In ASR, the success of dynamic features [21] , cepstral mean subtraction (CMS), and RelAtive SpecTrAl (RASTA) techniques, all of which involve temporal processing of short-term spectral features over 50 ms to several hundred milliseconds, indicates that temporal processing is required for improving robustness of ASR systems. Psychoacoustic studies also suggest that the peripheral auditory system in humans integrates information from much larger time spans than the temporal duration of the short segments used in spectral processing. One piece evidence is forward masking [34, 50, 51] which is an auditory process that operates over a relatively longer time span ($200 - 500$ ms). Recent studies based on Multivariate ANalysis Of VAriance (MANOVA) [59] and based on joint mutual information [29] also indicate that the effect of a phone lasts for approximately 500 ms. Several attempts were made to incorporate information from longer time spans of the speech signal into features [26, 62].

One recent approach applies temporal processing in the multi-band framework [26]. In this approach, acoustics and environment characteristics are estimated independently from narrow frequency regions of speech signal and later combined for final classification. This approach is motivated by the Fletcher and Stewart multi-channel model of phone recognition in human auditory system, which was brought up into attention later by Allen [4]. Their work suggests that the recognition errors in a given frequency band are independent of the errors in the other frequency bands, thus each frequency channel should be processed independently for building reliable acoustic models [19]. The independent processing of each frequency channel allows for selection of reliable (with high signal-to-noise ratio (SNR)) feature estimates when combining information from narrow frequency regions. Several benefits have been reported on using the multi-band approach for robust recognition [12, 67].

1.3 The TempoRAl Patterns based (TRAPS) approach for feature estimation

With the previous studies as motivation, the temporal characteristics of individual phones were analyzed to understand the nature of the linguistic information available in the temporal structure of speech. For this analysis, frequency-localized features such as cube-root or log compressed critical-band energies were used. Results showed that each phone has a unique pattern of energy evolution in each frequency band [64]. Utilizing the information present in the temporal patterns of the phones in each band, a TRAPS based system was developed for feature estimation [26, 61]. The basic building blocks of this system are shown in Figure 1.2. This system consists of the following processing steps :

Class-posteriors estimation from individual frequency-bands

The temporal patterns of critical-band energies are used to estimate the posterior probabilities of sub-word classes (phones) in each frequency band. Multilayer perceptrons (MLPs) are used as the class (phone) posterior estimators. The input features to each MLP (posterior estimators) are cubic-root or log compressed Bark or Mel-scaled critical-band energies. Each MLP makes a local estimate of the posterior probability of sub-word (phone) classes from a 101−sample window ($\pm$50 samples around the current frame) of compressed critical-band energies at every 10 ms. MLPs are trained using error backpropagation algorithm with a cross-entropy error criterion. Each 101−sample window (covers 1 s of context) is normalized to have zero mean and unit variance before using it as input to the posterior estimators (MLPs). The normalized 101−sample vector of compressed critical-band energies is also referred to as temporal pattern which essentially captures temporal dynamics around the current frame.

Merging information from frequency bands

The information about sub-word classes (phones) gained from narrow frequency bands is then combined for final recognition. For this, the estimated class-posterior probabilities from individual bands are concatenated and converted into class-conditional log-likelihoods

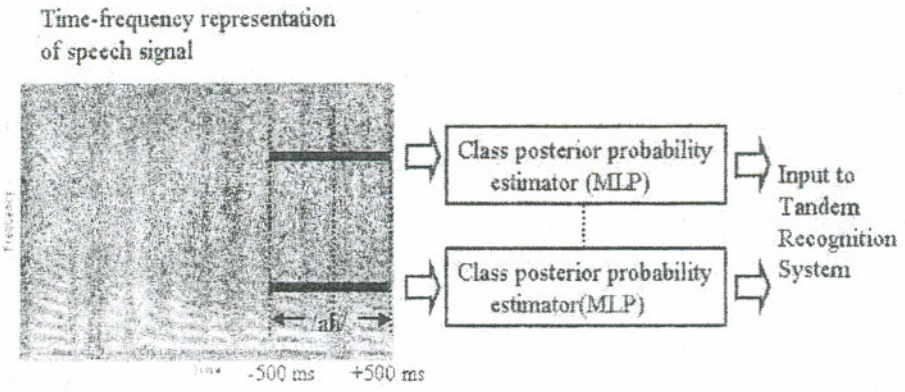

Figure 1.2: The TempoRAl PatternS (TRAPS) approach for feature estimation

by taking log of their estimates and scaling by class priors. The concatenated vector of log-likelihoods is then treated as the input feature vector to a Tandem speech recognition system [27]. The details of this recognizer are described in the following section.

1.4 The Tandem approach to speech recognition

The basic idea behind this approach is to use a Neural Network (NN) to map given features into class-posterior (e.g. phone posteriors) probabilities. A Neural Network is discriminantly trained using a labeled dataset to estimate posterior probabilities of the phone classes given the observations. The observations are nothing but a set of features (e.g. 9 frames of MFCC , class-posteriors estimated from individual critical-bands etc.). At a given framerate (e.g. 10−ms) NN gives an estimate of a set of posterior probabilities given a observation or a feature vector. The distribution of these posteriors is heavily skewed and these posteriors are correlated with each other. This makes these posteriors, if used directly as input features, not suitable for successive Gaussian Mixture Modeling (GMM) in a HMM-based recognizer. Due this reason, class-posterior probabilities are first processed and then used further as input features for a conventional back-end HMM-based recognizer. The posteriors are converted to log likelihoods by taking a log to their

values. The other alternative is to collect outputs from NN without the final non-linearity (e.g. softmax) which gives a close estimate of log of the posterior probabilities. The distribution of the log likelihoods of classes is more Gaussian-like, which is required for the successive GMM. This step is referred to as Gaussianization of features. The log likelihoods values of the classes can be correlated with each other. Since most commonly GMM uses diagonal covariances assuming feature vectors are decorreleted, a KLT (Karhunen-Loeve Transform) is applied on the log of posterior probabilities to decorrelate them. This processing step is referred to as Diagonalization or De-correlation of features. The KLT transform is estimated a priori using the training dataset or using an independent dataset. KLT transform is also known as Principal Component Analysis (PCA) transform. Finally the processed posteriors that are estimated at every $10-$ms are used as input features to the back-end HMM-based recognizer [27, 13].

A Tandem system adds a layer of discriminability to the feature extraction process and provides a convenient mechanism for combining multiple feature sets [61]. The general block diagram of such a system is shown in Figure 1.3.

Combining different feature streams in a Tandem system

In a Tandem system, information from different feature streams can be combined for final classification. The block diagram of such a system is shown in Figure 1.4. A set of estimated features are mapped to class-posteriors (e.g. phone posteriors) using MLPs and different feature streams are combined at the level of class-posterior probabilities. The combination techniques can be as simple as averaging the class-posteriors or concatenating the posteriors (followed by decorrelation) estimated from different feature streams. A gain in recognition performance has been seen when two or more feature streams make complementary errors [61].

1.5 Contributions of this work

This thesis work further investigates the use of frequency-localized temporal patterns of features in ASR. First, it investigates various techniques for parameterization and

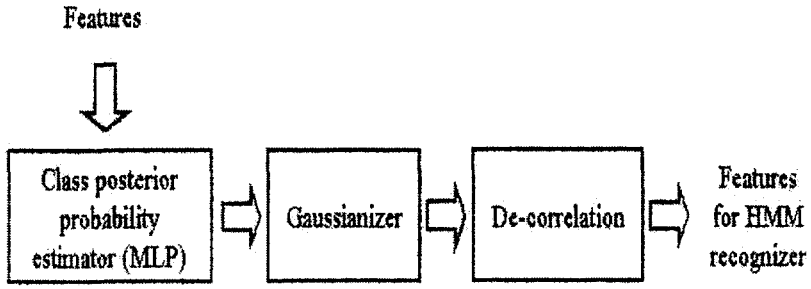

Figure 1.3: Block diagram of Tandem speech recognition system

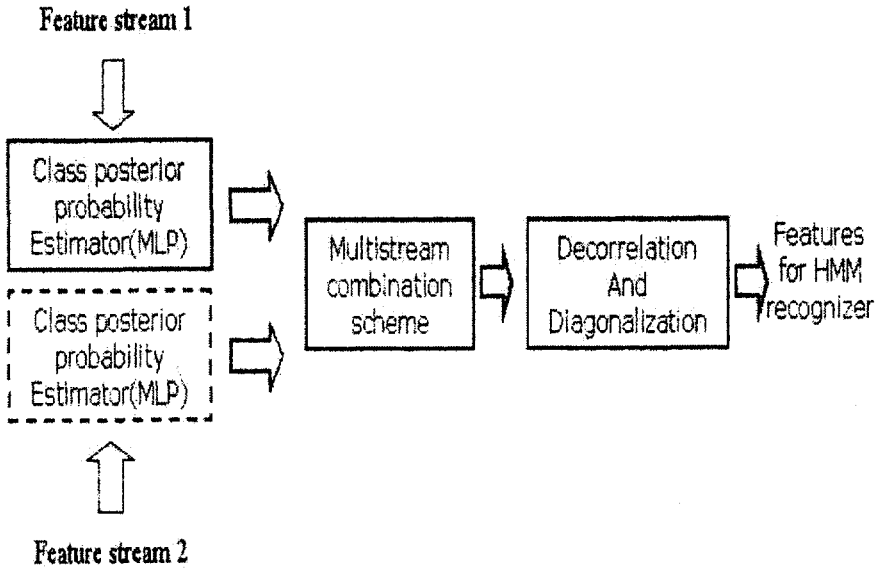

Figure 1.4: Combining multiple feature streams in Tandem speech recognition system

representation of the temporal patterns in each band. Then it investigates the use of broad-phonetic categories based on, for example nasality, affrication, voicing, and plosive nature of the speech sounds for feature estimation. It shows that broad phonetic features generalize better across different conditions and are robust to noisy environments. The estimated broad-phonetic category features are evaluated for two practical applications : Voice Activity Detection (VAD) and Distributed Speech Recognition (DSR) in cellular telephony. This thesis further presents a new approach for defining band-independent broad-phonetic categories for feature estimation. These categories characterize distinct speech events manifested in the frequency-localized temporal patterns of the speech signal. We also show that by using these new speech-activity based categories, band-specific class-posterior estimators can be replaced by a universal class-posterior estimator. Finally, this thesis work studies the effect of broadening the frequency-context presented to the class-posterior estimators.

The contribution of this thesis can be summarized as follows

- **Investigated various parametric representations for frequency-localized temporal patterns in ASR** Several linear transforms such as DFT, DCT, and LDA are investigated for the parameterization of temporal patterns. We show that the temporal patterns closely follow the properties of a first-order Markov process which results in the PCA transforms being very close to the DCT transform. Better recognition performance is achieved on using linear components of temporal patterns as opposed to directly using patterns as input to the posterior estimators for feature estimation. Frequency components of frequency-localized temporal patterns are referred to as modulation spectral components. These components capture the temporal dynamics of spectral envelopes. The 101-sample (1 s) temporal trajectory covers $0 - 50$ Hz of modulation frequency components. The effect of cutting off the components of higher modulation spectra on ASR is also studied by limiting the higher order DCT components during parameterization. We show that on cutting off modulation frequency components beyond 8 Hz, recognition performance severely degrades in different testing environments. With around 32 DCT components (covers upto 16 Hz range of modulation spectrum) better recognition performance can

be achieved across different testing environments.

- **Investigated broad-phonetic feature estimation from temporal patterns in ASR** Various broad-phonetic categories are investigated for feature estimation. The broad-phonetic categories are obtained independently in each band by clustering the mean temporal patterns of phones using a similarity measure. This results in broad-phonetic classes based on vowels, plosives, fricatives, flaps, nasals, schwa, and silence like groups of phones in each band. The broad-phonetic features are evaluated for ASR. Work this part shows that broad-phonetic category based features generalize better than other conventional features and yield considerable complementary information with respect to short-term spectral features.

- **Proposed broad-phonetic features for VAD and DSR** The estimated broad-phonetic category based features are evaluated for two practical applications : Voice Activity Detection (VAD) and Distributed Speech Recognition (DSR) in cellular telephony. The proposed features show significant improvements in performance for these applications.

- **Proposed Universal Temporal Patterns (UTRAPS) based approach for feature estimation** We found that mean temporal patterns are quite similar not only across similar phones but also across different frequency bands. We propose new band-independent categories obtained by clustering temporal patterns from all the critical-bands. The resulting clusters represent distinct speech-activities manifested in frequency localized temporal patterns of the speech signal. The class-posterior probabilities for these new, band-independent categories can be estimated using a universal posterior estimator. We show significant improvement in recognition performance in various testing conditions on using UTRAPS for feature estimation.

- **Studied the effect of broadening frequency-context for feature estimation on ASR** This study shows that combining temporal patterns from more than one critical-band is important to achieve better recognition rates. This work indicates

modeling the interaction among several adjacent critical-bands is important for obtaining higher recognition performance.

1.6 Organization of the thesis

Chapter 2 describes different linear representations for the parameterization of temporal patterns of log critical-band energies. The effect of parametrization of temporal patterns on ASR is also studied. This work proposes Discrete Cosine Transform (DCT) or DFT components from $1 - 16$ Hz frequency range for efficient representation of information carried by temporal patterns.

Chapter 3 investigates band-dependent broad-phonetic categories for feature estimation using the parameterized temporal patterns. This work shows that broad-phonetic category features generalize better across different testing conditions. It shows that six broad-phonetic category features carry a significant amount of complementary information to short-term spectral features such as cepstral features. The effect of choosing different sets of broad-phonetic categories for feature estimation on ASR is also investigated.

Chapter 4 investigates further the use of broad-phonetic category based features in two practical applications : Voice Activity Detection (VAD) and Distributed Speech Recognition (DSR) in cellular telephony. Results shows significant improvements in the recognition performance for both applications.

Chapter 5 investigates a new set of band-independent categories for feature estimation. These categories are based on distinct speech-activities localized in time-frequency regions of the speech signal. As opposed to earlier work that used band-specific class posterior estimator, a universal class-posterior estimators is proposed for feature estimation.

Chapter 6 studies the effect of broadening the frequency-context by using temporal patterns from several adjacent critical-bands for feature estimation on ASR. The work in this chapter proposes that temporal patterns from several adjacent critical-bands should be used jointly for feature estimation. A significant improvement in recognition performance is achieved on using more than one critical-band frequency-context for feature estimation, as opposed to using frequency-context spanning just a single critical-band.

Chapter 7 summarizes the work and suggests future directions.

Chapter 2

Parametric representation of temporal patterns of features

As described earlier in section 1.3, the TRAPS based system was developed to incorporate information from syllable-length time spans of the speech signal into short-term features [26]. In this system, 1 s long (101 samples at a 10 ms frame-rate) temporal trajectories of log critical-band energies were used directly as input features for sub-word unit (phones) posterior estimation. The 101-point input vectors require the training of a large number of parameters in individual class-posterior estimators (MLPs) which cannot be accurately estimated from a limited amount of available training data. We believe that by doing appropriate parameterization of the temporal feature vector, the large dimensionality of temporal patterns can be reduced without losing useful information required for estimating class-posterior probabilities.

In this work, we investigate various linear transforms for efficient representation or parameterization of frequency-localized temporal patterns. We show that temporal patterns closely follow the characteristics of a first order Markov sequence [8]. This results in the principal components of the temporal patterns being similar to Discrete Cosine Transform (DCT) components. We show that DCT components can be used to parameterize 1 s long (101-point) temporal patterns and can achieve a significant reduction in number of system parameters. The Linear Discriminant Analysis (LDA) [55] and the Fourier Transform (DFT) are also investigated for representing temporal patterns.

The frequency components of temporal patterns are referred to as modulation spectral components. They represent the change in the spectral envelope over time. The effect of

low-pass filtering the modulation spectra of temporal patterns on ASR is also studied in this chapter.

The chapter is organized as follows. Section 2.1 analyzes temporal patterns and show their association with first-order Markov sequences. Section 2.2 describes various linear transforms used for parameterization of temporal patterns. Section 2.3 describes the used databases and the experimental setup for evaluating parameterized temporal patterns in ASR. Section 2.4 shows the effect of parameterization of temporal patterns on ASR. Section 2.5 investigates the effect of low-pass filtering the modulation spectra of temporal patterns on ASR by limiting the linear transform basis.

2.1 Temporal patterns as Markov-1 vector sequence

For 101 point long, mean subtracted and variance normalized, temporal patterns of log critical-band energies, the estimated covariance matrix is shown in Figure 2.1. Its highest values lie along the diagonal and the off-diagonal terms gradually decay. This matrix has a structure close to that of $r(n)$, which is given by,

$$r(n) = |\rho|^n, \forall n,$$

where $r(n)$ represents the covariance function of a Markov-1 vector sequence [8]. For temporal patterns, the estimated value of the correlation coefficient, ρ, is around 0.85. This results in eigen vectors of the covariance matrix of the temporal patterns (Principal components i.e. PCAs) being close to the DCT bases. The more details of the covariance function of a Markov-1 vector sequences can be found in Appendix A. The PCA computation is described in detail in Appendix G. The estimated principal components are shown in Figure 2.2. Figure 2.3 shows the distribution of total variability in the estimated principal components. It can be observed that around 50 principal components cover around 98 % of the total variability. These 50 bases span $1 - 25$ Hz of modulation spectrum of temporal patterns.

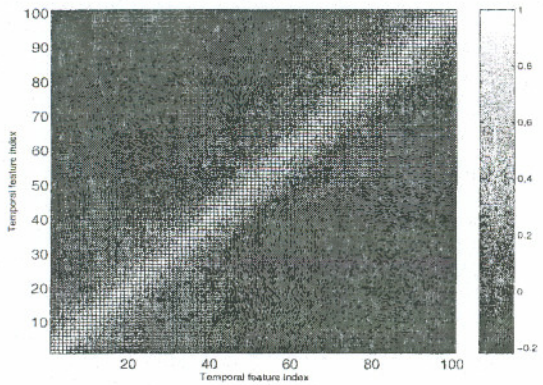

Figure 2.1: Correlation matrix of 101 point long temporal pattern of log critical-band energies

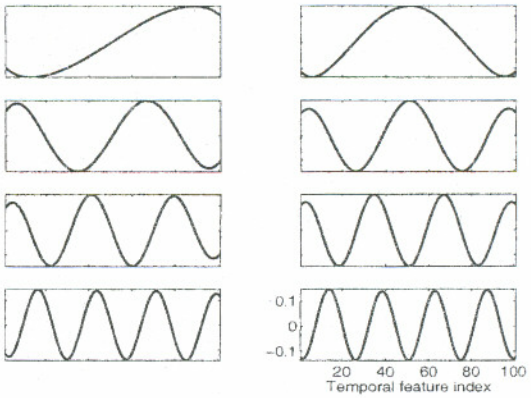

Figure 2.2: Principal components basis vectors

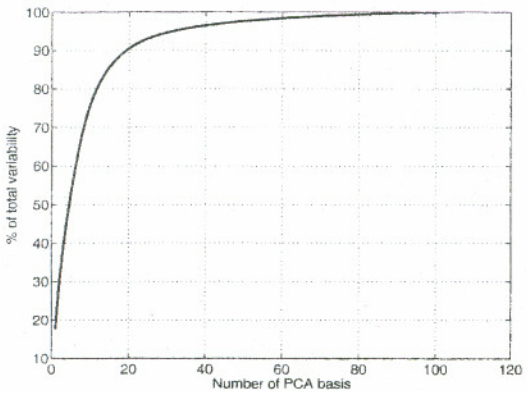

Figure 2.3: % of total variability in number of retained PCA basis vectors

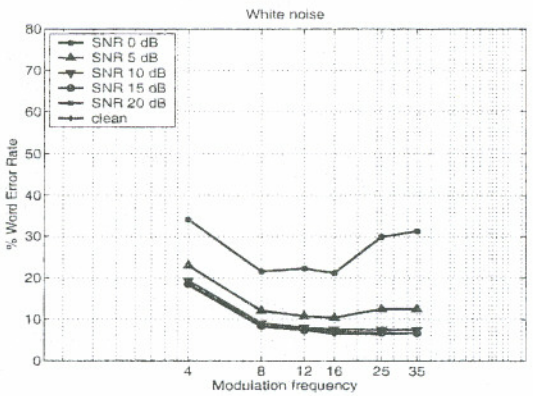

Figure 2.4: Effect of low-pass filtering the modulation spectrum at different SNRs by limiting DCT bases in White noise

2.2 Linear transforms for parameterization

In this section, we describe various linear transforms used for the parameterization of temporal patterns. The parameterized temporal patterns are used as input to class-posterior estimators (MLPs) in each critical band. Later these posteriors are combined (concatenated), converted to log-likelihoods, and used as input features in a Tandem speech recognition system [13].

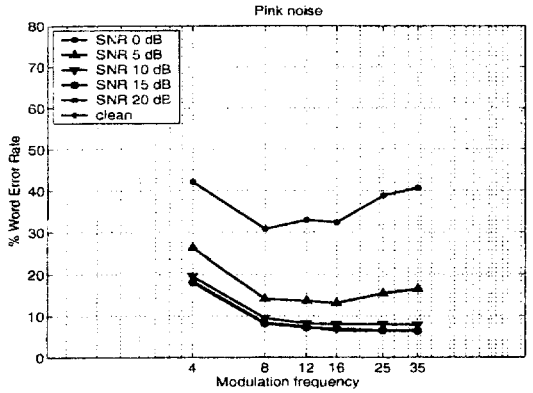

Figure 2.5: Effect of low-pass filtering the modulation spectrum at different SNRs by limiting DCT bases in Pink noise

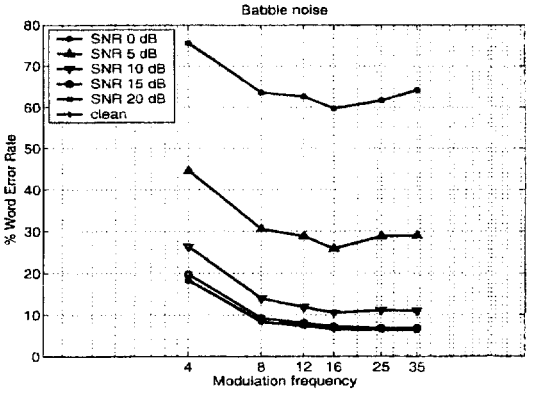

Figure 2.6: Effect of low-pass filtering the modulation spectrum at different SNRs by limiting DCT bases in Babble noise

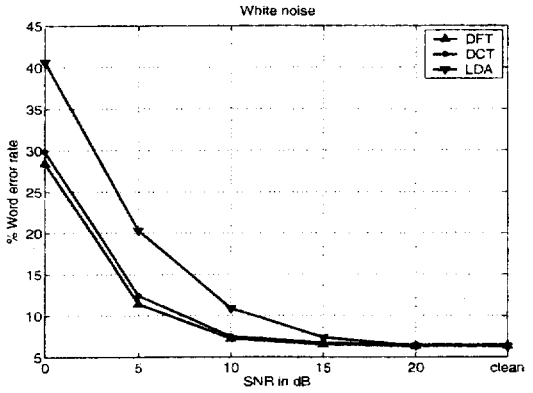

Figure 2.7: Word error rates from different linear transforms at different SNR conditions in White noise

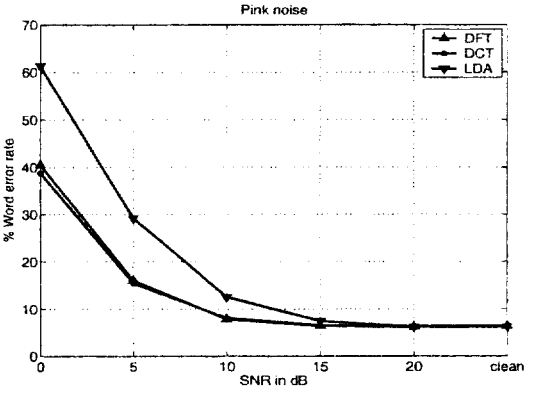

Figure 2.8: % Word error rates from different linear transform at different SNR conditions in Pink noise

DCT linear transform

In the previous section we observed that the covariance matrix of 101 point long temporal feature vectors (spanning around 1 s of context) has a structure very similar to that of first-order Markov sequences and due to this reason, the PCA transform can well be approximated by a DCT. The 50 PCA components cover around 98 % of the total variability of 101-point long temporal patterns. We used 50 DCT components of mean subtracted, variance normalized temporal patterns as input features to sub-word class (phones) posterior estimators (MLPs).

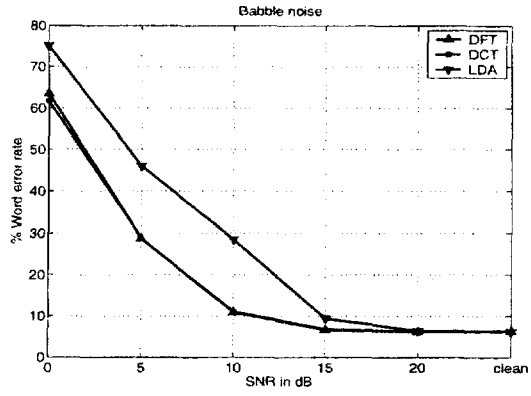

Figure 2.9: % Word error rates from different linear transforms at different SNR conditions in Babble noise

Discrete Fourier Transform (DFT)

Fourier components covering 1 – 25 Hz of modulation spectrum are used for parameterization of temporal patterns. The 50 DFT components of temporal patterns, which consist of 25 real and 25 imaginary parts, are used as input features to sub-word class (phones) posterior estimators.

Linear transform obtained from LDA

This linear transform is obtain by using Linear Discriminant Analysis (LDA) [20, 55]. The linear discriminants are derived from 101 point long temporal trajectories of log critical-band energies, which are mean subtracted and variance normalized. The 29 phones are used as the target classes for deriving 29 linear discriminant basis vectors. The LDA bases are derived on the clean, OGI-Stories corpus, a dataset described in section 2.3. The first 15 LDA components of the temporal patterns are then used as input features to sub-word class (phones) posterior estimators. We also derive the first 50 LDAs basis vectors from 554 Gaussian mixtures (from 3-state monophone HMM models), which are used as the target categories, for the second experiment. The more details of computing linear discriminants is covered in Appendix G.

2.3 Databases

Two databases are used in this work : 1) the OGI-Numbers database and 2) the OGI-Stories database. The training of band-specific MLPs is performed on OGI-Stories while the Tandem MLP is trained on OGI-Numbers.

OGI-Numbers database

The OGI-Numbers corpus [52] consists of a set of continuous, naturally spoken utterances collected from many different speakers over the telephone and sampled at 8 kHz. 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 but 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 of approximately 1.7-hour, 0.6-hour and 0.2-hour respectively have been used for our experiments. The 1.7-hour subset is the training set, the 0.2 hours subset is the cross-validation set on which the frame-level error rates are reported. We used these datasets for training of the posterior estimators (MLPs). The Numbers database consists of 29 phonetic classes which are a subset of the 61 phones from the TIMIT database. The word recognition experiments are performed on the continuous digit-part of this database. This dataset is referred to as OGI-digits. It consists of 1.3-hour training set and 1.7-hour testing set.

2.3.1 Testing conditions at different SNRs

There are five testing conditions : 1) clean, 2) babble, 3) pink, and 4) white noises artificially added at 0, 5, 10, 15, and 20 dB signal-to-noise ratio (SNR) to the clean testing dataset. We define SNR as the ratio of signal to noise energy after filtering both speech and noise signal to $0 - 4000$Hz frequency-range. The noise energy is calculated as the root mean square value (RMS) over the noise signal segment of same length as the

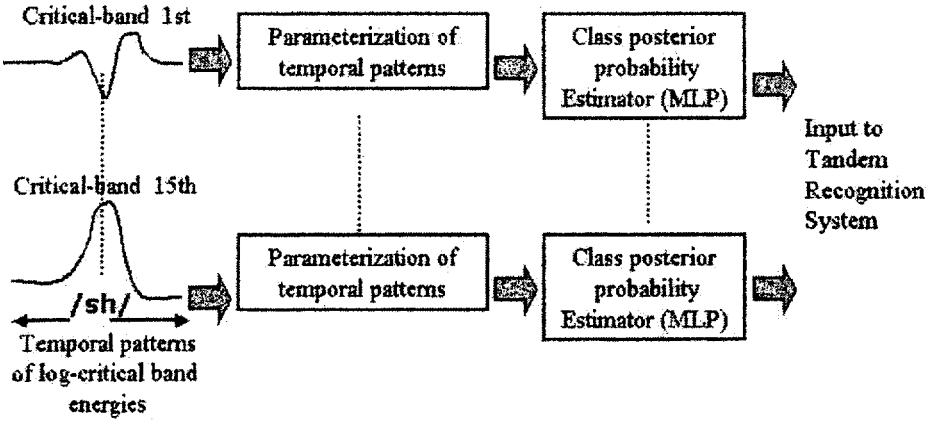

Figure 2.10: Block diagram of temporal patterns based feature estimation

speech signal is. This noise segment is selected randomly from the noise signal. We assume duration of the noise signal much longer than that of the speech signal. The level of the speech signal is not changed . Based on the desired SNR the scaling factor is calculated to multiply the noise samples before adding them to the speech signal.

OGI-Stories database

OGI-stories is the English portion of the OGI multi-lingual database [53]. The database consists of telephone quality conversational speech. A subset is phonetically hand labeled and comprises approximately 2.7 hours of extemporaneous speech from each of the 210 different speakers.

Experimental setup

Different linear transforms for parameterization of temporal patterns are evaluated on a continuous digit recognition task using OGI-digits database. In this task, the vocabulary consists of 11 words ($0 - 9$ and "oh"). Each word is modeled as a sequence of context-independent phones and each phone is modeled using a five-state, three-mixture Gaussian hidden Markov model (HMM). The block diagram of this system is shown in Figure 2.10. The training of the band-specific class-posterior estimators (MLPs) is performed on clean OGI-Stories and the posterior estimator used in Tandem system is trained on clean

OGI-Numbers. The Multilayer perceptrons (MLPs) are trained using backpropagation algorithm with a cross-entropy error criterion. They were trained to estimate 29 phones posteriors from linearly transformed temporal patterns of log critical-band energies. The used critical-bands are described in Appendix H. The band-specific MLPs are trained using 150 hidden units and 29 phones as the target categories used at the output nodes. The MLP used in the Tandem system (described in detail in section 1.4) is trained with 300 hidden units and the same 29 phone targets for final class-posterior estimation. These posteriors are gaussianized (by taking the log), decorrelated (using a KLT transform), and then used as features in an HMM-based recognizer.

2.4 Performance of linearly transformed temporal patterns

Table 2.1 shows word error rates for features estimated from parameterized temporal patterns using various linear transforms at the clean condition. The performance of these features is compared with the features obtained by directly using temporal patterns as the input to the class-posterior estimators. From the results, it can be observed that the linearly transformed temporal patterns perform significantly better than using these patterns directly for feature estimation. The gain in the performance is significant at 98 % confidence interval. Figures 2.7, 2.8, and 2.9 show the recognition performance of linearly transformed temporal patterns using DFT, DCT, and LDA transforms at different noises and SNR conditions. In this experiment, 50 linear bases were used for the projection. The features estimated from DCT-transformed and DFT-transformed temporal patterns perform significantly better than LDA-transformed temporal patterns at lower SNR noisy conditions. This result is significant at 98% confidence interval. For confidence measure, we performed a Matched Pairs Sentence-Segment Word Error (MAPSSWE) test [23]. On the average, DCT and DFT-transformed temporal patterns give very similar recognition performance across most of the noisy conditions.

2.5 On low-pass filtering the Modulation spectra

Previous studies on the intelligibility of speech with filtered spectral time-trajectories show that speech intelligibility isn't severely impaired as long as the spectral envelope has a rate of change up to 16 Hz in modulation frequency [54]. The studies such as conducted by N. Kanedera et.al. [45, 46] indicate that modulation frequency components from range between 1 to 16 Hz carry most of the linguistic information important for robust ASR. In contrast to the above findings, the work by R. V. Shannon et. al. [57] had found that recognition scores from 50, 160, and 500 Hz modulation (speech envelop) filters for vowels, consonants, and sentences were not significantly different but the results from 16 Hz modulation filter gave significantly lower recognition scores.

With these studies as motivation, we investigate the effect of eliminating higher modulation frequency components of temporal patterns on ASR at different noisy conditions and at different signal-to-noise ratios (SNR). Figures 2.4, 2.5, and 2.6 show the effect of cutting off higher modulation frequencies on the word error rate by dropping higher DCT components at different testing conditions. The recognition performance drastically reduces on cutting off the modulation frequencies beyond 8 Hz whereas it fairly remains the same for upto 12-35 of modulation spectrum for higher SNR ($10 - 20$ dB and clean) testing conditions. However, for lower SNR ($0 - 5$ dB) conditions, keeping modulation spectrum above 16 Hz reduces the recognition performance.

Table 2.1: % Word Error Rate (Recognition Performance) on using different linear transforms on temporal patterns of log critical-band energies at clean condition

101 temporal patterns	50 DFT	50 DCT	50 LDA	15 LDA
7.1	6.4	6.4	6.3	6.6

2.6 Conclusion

In this chapter, we investigated various linear transforms for the parameterization of temporal patterns. We showed that temporal patterns closely follow the properties of first

order Markov sequences and as a result the principal components (PCA) of temporal patterns can be well approximated by DCT components. Around 50 PCA (DCT) components cover 25 Hz of modulation spectrum of 101−point frequency-localized temporal patterns. The features estimated from linearly transformed temporal patterns give significantly better recognition performance than those estimated directly from temporal patterns. Using 50 linear components of temporal patterns reduces the total number of parameters in posterior estimator from 585 k to 178 k. DCT, DFT, and LDA components give very similar recognition performance in clean condition. However, DCT components and DFT components of temporal patterns outperform those obtained from LDA components of temporal patterns [62] in $0 - 15$ dB SNR conditions. The reason why LDA components of temporal patterns didn't perform well is that the posterior estimators (MLPs) play the role of non-linear discriminants that reduces the advantage of using linear discriminants for parameterizing temporal patterns. They were also derived from a clean dataset (OGI-stories corpus), so they do not generalize well to unseen conditions. We also showed the effect of low-pass filtering the modulation spectrum of temporal patterns on ASR. The results show that recognition performance drops significantly on cutting off modulation spectrum below 8 Hz. As long as upto 16 Hz of modulation spectrum is retained for the successive feature estimation, higher recognition rates can be achieved. This result holds good for noisy conditions ranging from $0 - 20$ dB SNR in our experiments.

Chapter 3

Band-specific broad-phonetic category features in ASR

In the previous chapter, we investigated various linear transforms for parameterization of frequency-localized temporal patterns (temporal trajectories of log energies of individual critical-bands) of the speech signal. The parameterized temporal patterns were used to estimate phone class-posterior probabilities, and later these posteriors were combined and used as input features to a Tandem recognizer (described in Section 1.4) for word recognition. In this work, the estimated phone-posterior probabilities are referred to as phone-category features. In this chapter, we investigate the estimation of broad-phonetic category features using parameterized temporal patterns independently from each frequency band. We further investigate the use of these features for robust speech recognition.

The perceptual studies such as those performed by Miller and Nicely [44] suggest that early in the human recognition process, speech perception is based on categorical decisions that are other than phones. They attempted to measure the confusion matrix for spoken and perceived phones in a noisy, bandlimited speech signal. They showed that unlike phones, distinction implied by certain phonetic features such as voicing, nasality, frication, etc. can be heard accurately in noisy and corrupted speech. Their studies suggested that categorical decisions could be based on these phonetic features. Their findings are summarized by Jont B. Allen [5]. The phonetic features categorize phones into broad-phonetic classes based on their articulation or acoustic properties. In this work, we refer phonetic features to broad-phonetic features. Several studies have been conducted,

which attempt to incorporate phonetic features for robust recognition [37, 36]. In our studies, We have also observed that temporal patterns of phones cluster very well into broad categories that are based on nasality, voicing, frication, etc. of the speech sounds. The phones cannot be recognized as robustly in narrow frequency bands as broad-phonetic categories. Motivated by perceptual studies and our observations, we investigate broad-phonetic category features for robust speech recognition. These features are estimated from temporal patterns of the speech signal independently in each frequency band.

We show that estimated broad-phonetic category features generalize better across unseen testing conditions than other conventional features (cepstral coefficients etc.) in ASR. We also show that broad-phonetic category features carry a significant amount of complementary information with respect to short-term spectral features such as Perceptual Linear Predictive coefficients (PLP).

The chapter is organized as follows. Section 3.1 describes earlier work on using broad-phonetic features in ASR. In Section 3.2, we analyze mean temporal patterns of subword classes (phones) and describe broad-phonetic classes which are obtained by clustering mean temporal patterns of phones. Section 3.3 analyzes the confusion matrix of phones and groups of the phones based on the degree of confusion obtained on phone classification. Section 3.4 describes the system used for broad-phonetic category feature estimation. Section 3.6 analyzes phone classification rates obtained from using broad-phonetic features estimated independently from individual bands. Section 3.7 shows the performance of broad-phonetic category features in ASR. Section 3.8 gives summary and conclusion of this work.

3.1 Earlier work on broad-phonetic features

Saul and his collaborators [39] attempted to detect [+/- sonorant] features from critical-band measurements of SNR and periodicity. This work suggested that greater understanding of narrowband phonetic cues is required for robust feature estimation for ASR. Kirchhoff [36] also demonstrated that broad-phonetic category features are quite noise-robust and can be used along with conventional features for robust recognition. Unlike

our method, the whole spectrum (9 frames of cepstral coefficients) of speech was used for detecting broad-phonetic cues. Other related work can be found in [37, 41, 47, 4].

3.2 Broad-phonetic classes from temporal patterns

For understanding the nature of information available in the frequency-localized temporal domain of speech signal, temporal patterns of log critical-band energies , given a phone, were analyzed [64]. We used Mel filterbank energies for our analysis. The computation of the filterbank is described in Appendix H. The mean temporal patterns of phones are computed as described earlier in [64]. The details of this computation can be found in Appendix B. The derived mean temporal trajectories show distinct patterns of temporal evolution for different sounds. They indicate that the effect of the center phone lasts for about syllable-length duration (250 ms) around it. The estimated mean temporal patterns of some of the TIMIT phones are shown in Appendix B. The TIMIT phone-set is listed in Appendix C. The following observations are made by analyzing the mean temporal patterns of different phones.

1. Mean temporal patterns of the phones spoken in a similar manner share very similar temporal characteristics. For example, the vowel sounds such as /ah/ and /ae/ are very similar in temporal characteristics in individual critical-bands. The same is true with unvoiced plosives (stop-consonants) such as /t/ and /p/.

2. The mean temporal patterns of phones cluster well into $6-7$ broad-phonetic categories such as plosives, nasals, fricatives, flaps, schwas, vowels, and silence categories of speech sounds. A simple correlation measure is used to cluster similar patterns into one group [64]. The correlation based measure is given by

$$D(x,y) = \frac{\sigma_{xy}^2}{\sigma^x \sigma^y}$$

where σ_{xy}^2, σ^x, and σ^y represent the cross-correlation between temporal patterns x and y, the auto-correlation of pattern x, and the auto-correlation of pattern y respectively.

The clusters of broad-phonetic categories show distinct temporal behavior in individual critical-bands. The cluster corresponding to vowels has an evolving peak in the center which characterizes formant like high energy regions in different frequency bands. The cluster corresponding to plosive-like sounds has a dip in energy off-center to left. This is due to the presence of a stop-closure of low energy which often precedes plosive sounds. The cluster corresponding to fricatives has a dip in the center, which depends on the low-energy frequency bands of individual fricative sounds. The energy increases around it due to the commonly surrounding vowels or high energy sounds. The clusters corresponding to 'flaps' and 'schwa' have rather complex temporal patterns. The mean temporal patterns of seven broad-categories are shown in Figure 3.5, 3.6, and 3.7.

3. The temporal patterns of some of the phones such as fricative /s/, nasal /m/, vowel /uw/, glides and approximants /w/,/l/, and /r/ differ considerably across critical bands. For example, the sounds /s/ and /sh/ show plosives-like temporal patterns in frequency bands up to 2 kHz. For higher frequencies, where these fricatives have high energy, they display vowel-like temporal pattern. The nasal sounds /m/ and /n/ show vowel-like temporal pattern up to $500-600$ Hz. In higher frequency bands they display fricative-like temporal patterns. The glides such as /w/, /y/ and, /l/ show vowel-like patterns in lower bands and fricative-like or plosive-like patterns in higher frequency bands. The mean temporal patterns of some of the phones such as the front-vowels /ae/, /iy/, schwa /axr/, nasal /m/, plosive /b/, fricative /sh/, and glide /l/ are shown in Appendix B.

Based on these observations we define seven broad-phonetic categories : nasals, vowels, fricatives, plosives, flaps, schwas, and silence. Note that here we ignored the fact that some sounds change cluster affiliation, e.g., fricative /s/ has vowel-like temporal characteristics in higher bands. We address this problem in the chapter 5 by defining new band-independent broad-phonetic categories.

Plosives	b,d,g,p,t,k,
Flaps	dx,nx
Fricatives	jh,ch,s,sh,z,zh,f,v,dh,th,hh
Nasals	m,n,ng,em,en,eng
Vowels	l,r,w,y,hv,el iy,ih,eh,ey ae,aa,ay,ah,er,axr ao,oy,ow,uh,uw
Schwas	ax,ix,ax-h
silence	sil,epi

Table 3.1: Seven broad-phonetic categories used for feature estimation from temporal patterns in individual critical-bands

3.3 Broad-phonetic classes from confusion matrix of phones

In this section, we analyze phone confusion matrix to find clusters of phones that are most often confused with each other during phone classification. Unlike the grouping of phones based on the similarity of their temporal characteristics in individual critical bands, this method of grouping is based on the similarity of the class(phones)-conditional feature distributions.

We train MLPs for estimating the phone posterior probabilities using DCT components of temporal patterns as the input feature vector in individual critical-bands. The phone confusion matrix obtained at the output of MLPs is far from being diagonal which suggests that narrow frequency-regions are not sufficient to detect phones as the speech sub-units. The confusion matrices for the first, fifth, and fifteenth bands are shown in Appendix D. The estimated local phone posterior probabilities are used as input to a Tandem system. The class-posterior estimator (MLP) in the Tandem system is trained to estimate phone posteriors for final classification. The training of all the MLPs is performed on the TIMIT dataset. Appendix D shows the phones confusion matrix at the output of the Tandem MLP. It can be seen that the clusters of highly confused phones are very similar to the groups of phones which have very similar temporal characteristics in individual critical-bands. It can also be observed that clusters of highly confused phones contain phones that have similar articulation properties.

3.4 Estimation of broad-phonetic category features

In this section, we describe the estimation process for broad-phonetic class features using temporal patterns. Figure 2.10 shows the block diagram of a temporal patterns based system for broad-phonetic feature estimation. The 50 DCT components computed from 101-point mean and variance normalized temporal patterns of log critical-band energies are used as input to class-posterior estimators (MLPs) in each critical-band. The phone labels are canonically mapped into the broad-phonetic categories. The output nodes of MLPs correspond to the broad-phonetic categories. There are two sets of broad-phonetic categories that are investigated in this work : 1) three categories based on voiced, unvoiced sounds , and background ; 2) seven broad-phonetic categories based on vowels, nasals, plosives, fricatives, flaps, schwa, and silence or background. OGI-Stories was used for training the band-specific class-posterior estimators (MLPs). The OGI-Numbers 29 phones were canonically mapped to one of the broad-categories. The category 'flap' was not present in the 29 phone set so it was ignored in the final broad-category set during feature estimation. The Tandem class-posterior estimator (MLP) was either trained on the target dataset, i.e., OGI-Numbers or an independent dataset, i.e., OGI-Stories.

3.5 Analyzing confusion matrix of broad-categories

Figures 3.1, 3.2, 3.3, and 3.4 show confusion matrix of seven broad-phonetic categories. We used TIMIT database for training MLPs and computing confusion matrices. The confusion matrix for these categories is more diagonal than the confusion matrix for phones. The confusion matrix of the broad-phonetic categories at the Tandem output is more diagonal than that obtained at the output of the band-specific MLPs. This suggests that information from all the critical-bands is needed to identify the broad-categories. The confusion among broad-categories in individual bands is due the fact that some sounds such as fricatives /s/ or front-vowel /iy/ change cluster affiliation in individual critical-bands. For example fricatives, /s/ or /sh/ in higher bands (above 2 kHz) have vowel-like high energy at the center, so they are confused with vowels. The mean temporal pattern of nasal /m/ shows vowel-like (peak at the center frame of temporal patterns) temporal characteristics

in lower frequency bands, while in higher bands it shows temporal characteristics similar to those of plosive-sounds (characterized by a narrow dip in energy off-center to the left). So it is likely for nasals to be confused with vowels in lower frequency bands and to be confused with plosives-like sounds in higher frequency bands. Table 3.2 shows the broad-phonetic category frame-level error rate at the output of the Tandem MLP measured (trained and tested) on the TIMIT dataset. The classification rates are higher for silence and vowel categories than other broad-phonetic categories.

3.6 Phone classification using broad-category features

To understand the extent to which the seven broad-phonetic features estimated from individual bands carry information about different phones, we use these features for phone classification. The estimated $15 * 6$ ('flap' was not present) broad-phonetic features from individual critical-bands are used as input features to a Tandem speech recognition system. The 29 phones are used as the target categories for classification at the output of the Tandem MLP. We analyzed the frame-level classification error rate for these phones. Table 3.3 shows the frame-level phone classification error of 26% at the output of the Tandem MLP.

3.7 Broad-phonetic category features in ASR

In this section, we describe performance evaluation of the broad-phonetic features in the Tandem speech recognition system. To evaluate the word recognition rate, the class-posteriors at the output of the Tandem MLP are Gaussianized, diagonalized (decorrelated), and used as input features to an HMM based recognizer (Section 1.4). The features are evaluated for a continuous digit recognition task on OGI-Digits database. The vocabulary is 11 words ($0 - 9$ and "oh").

Table 3.3 shows the performance of the broad-phonetic features in a Tandem speech recognition system. In this experiment, the broad-phonetic category $15 * 6$ features are used to estimate 29 phone category posteriors using the Tandem MLP (Section 1.4). Around 26 % frame-level phone classification error is obtained at the output of the Tandem

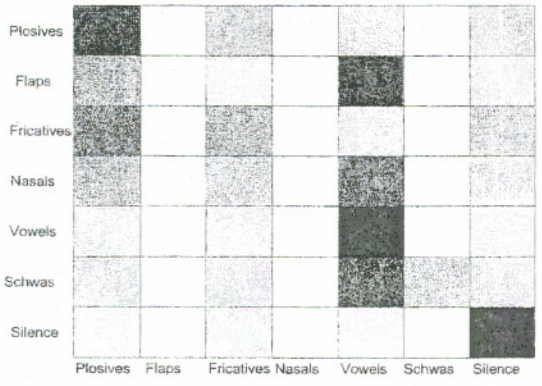

Figure 3.1: Confusion matrix in 1st band

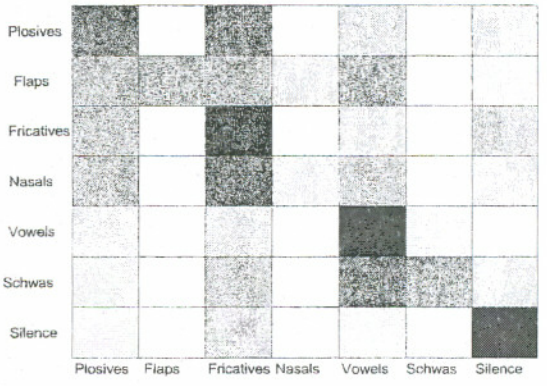

Figure 3.2: Confusion matrix in 5th band

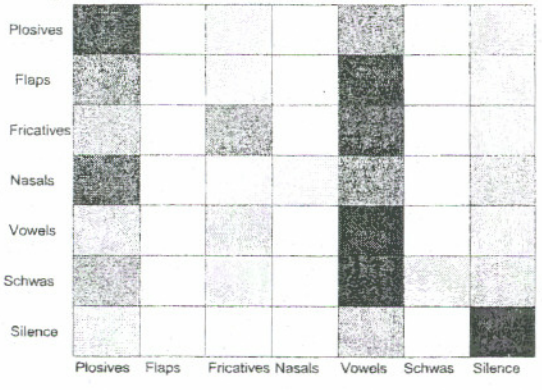

Figure 3.3: Confusion matrix in 15th band

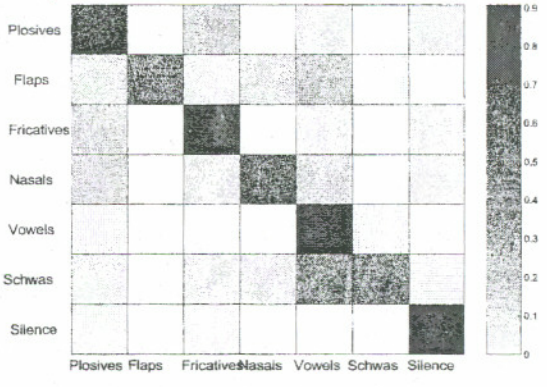

Figure 3.4: Confusion matrix at the Tandem MLP output

MLP. For word recognition performance evaluation, these 29 posteriors are gaussianized, decorrelated, and used as input features to the HMM based back-end recognizer (Section 1.4). In the HMM based back-end recognizer, each word is modeled as a sequence of context-independent monophones and each monophone is modeled using a five-state, three-mixture Gaussian hidden Markov model (HMM). These features give around 8.75 % word recognition error.

Table 3.4 shows the performance of broad-phonetic features when they are used in conjunction with Perceptual Linear Predictive (PLP) cepstral coefficients. The cepstral coefficients and broad-phonetic features are combined at the class-posterior level. In this experiment, the 29 phone posteriors , estimated from 9 frames of 24 PLP cepstral coefficients and their dynamic features, are concatenated with the final 6 broad-phonetic

class posteriors. The final 6 posteriors are obtained at the output of the Tandem MLP using $15 * 6$ broad-phonetic features as the input. The concatenated posteriors, $29 + 6$ in dimension , are gaussianized (by taking the log) and decorrelated using a whitening transform. The total dimensionality is reduced from 35 to 29 and finally these 29 features are used as input to the HMM based back-end recognizer (Section 1.4). Using the same framework, we also evaluate three broad-phonetic features based on voicing (i.e. voiced speech, unvoiced speech, and background) when used in conjunction with PLP features. The result obtained using these features is also shown in Table 3.4. We compare performance of phone-category features with broad-phonetic category features. To obtain the phone-category features band-specific posterior estimators (MLPs) estimate phone posteriors independently in each critical-band. The $29 * 15$ phone posteriors estimated from 15 critical bands are mapped to 29 phone posteriors using the Tandem MLP. These posteriors are referred to as phone-category features which were also proposed in earlier system [64]. Table 3.4 also shows recognition performance of these features when they are used in conjunction with PLP. The 29 phone posteriors at the output of the Tandem MLP are concatenated with 29 PLP-estimated phone posteriors. The concatenated posteriors are gaussianized, decorrelated, reduced in dimension from 58 to 29, and used as input features to the HMM based back-end recognizer. From the results, it can be observed that gain in the recognition performance is the same for six broad-phonetic category features as it is for 29 phone category features, both estimated from temporal patterns. This indicates that the amount of complementary information provided by 6 broad-phonetic features is the same as that carried by 29 phone-category features. This gain in the recognition performance is significant at 99 % confidence interval. We used Matched Pairs Sentence-Segment Word Error (MAPSSWE) test to perform significance test.

Table 3.5 compares the generalization properties of the broad-phonetic category and the phone-category features when parameters of the system are entirely trained on an independent dataset. For this experiment, we trained band-specific posterior estimators (MLPs) and the class-posterior estimator (MLP) used in the Tandem system on a task-independent dataset, OGI-stories. In this case, the broad-phonetic category features give better gain in word level recognition performance than phone category features.

Categories	priors	% error rate
plosives	17.0	32.8
flaps	0.6	52.9
fricatives	16.6	27.9
nasals	6.2	42.2
vowels	38.9	12.6
schwas	6.6	57.3
silence	13.8	9.3

Table 3.2: Frame-level broad-phonetic categories error rate (%) on TIMIT dataset.

	frame error	word error
broad-phonetic features	25.8	8.7

Table 3.3: Frame-level phone classification error (%) and WER on OGI connected digits task using $6 * 15$ broad-phonetic category features estimated from each band.

PLP features	with 3 broad-phonetic category features	with 6 broad-phonetic category features	with 29 phone category features
5.0	4.8	4.3	4.3

Table 3.4: Word Error Rates (%) on OGI-Digits, using cepstral features alone, or using cepstral augmented with temporal patterns estimated features. Band-specific posterior estimators (MLPs) were trained on OGI-Stories and the posterior estimator used in the Tandem system was trained on the OGI-Numbers. The dimensionality of final feature vector was kept at 29

PLP features	with 6 broad-phonetic category features	with 29 phone category features
7.4	6.9	7.3

Table 3.5: Word Error Rates (%) on OGI-Digits task, using cepstral features alone or cepstral augmented with temporal patterns estimated features. The band-specific posterior estimators (MLPs) and the posterior estimator used in Tandem system were trained on OGI-Stories, an independent dataset. The dimensionality of final feature vector was kept at 29

3.8 Conclusions

We showed that temporal patterns cluster well into a set of broad-phonetic categories in each critical band. The confusion matrix of phones obtained from phone classification also shows similar clusters of phones. The temporal patterns of log critical-band energies can be used to estimate broad-phonetic category features, independently from each frequency band.

The confusion matrix (obtained at the output of Tandem MLP) of broad-phonetic classes is more diagonal than that of phone classes. This indicates that broad-phonetic categories are better classified by temporal patterns than phones. By using just six broad-phonetic category features around 26 % frame-level phone error rate for 29 phones (Appendix C) is achieved. These features give around 8.75 % word error rate when used for word recognition.

A consistent and significant gain in the recognition performance is obtained by augmenting broad-phonetic features with short-term spectral features such as PLP cepstral coefficients. This indicates that the broad-phonetic features are complementary in nature with respect to short-term cepstral features. The gain in the recognition performance is significant at 99 % confidence interval. For confidence measure, we applied Matched Pairs Sentence-Segment Word Error (MAPSSWE) test [23].

The broad-phonetic TRAPS features give a similar gain in the recognition performance as that given by phone category TRAPS features. However, broad-phonetic category TRAPS features generalize better than phone-category TRAPS features when the system is entirely trained on an independent dataset. The reason for this is attributed to having more data to train each broad-phonetic categories as each category comprises several phones. The system uses 120 k $(15*(50*101+101*6)+15*6*200+200*6)$ parameters which is around 50 % less than the parameters $(15*(50*101+101*29)+15*29*300+300*29)$ used in the earlier proposed TRAPS system (Section 1.3).

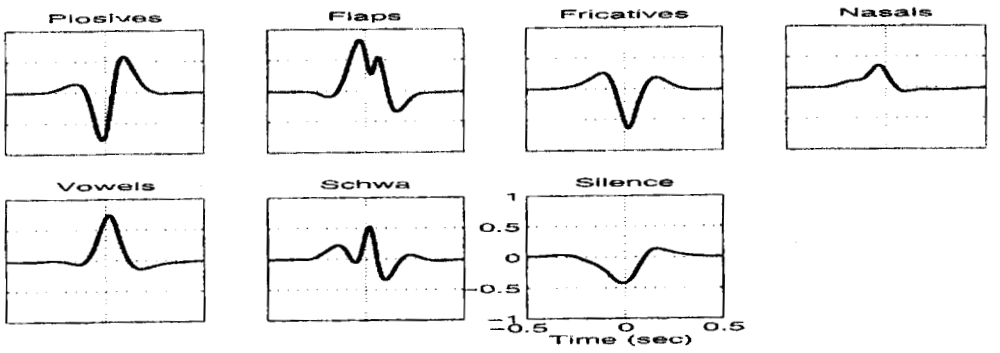

Figure 3.5: Mean Temporal Patterns of seven broad categories in 1st critical-band

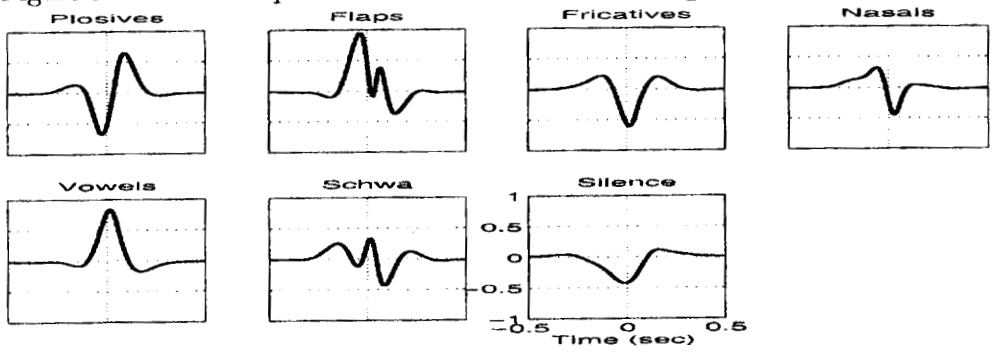

Figure 3.6: Mean Temporal Patterns of seven broad categories in 5th critical-band

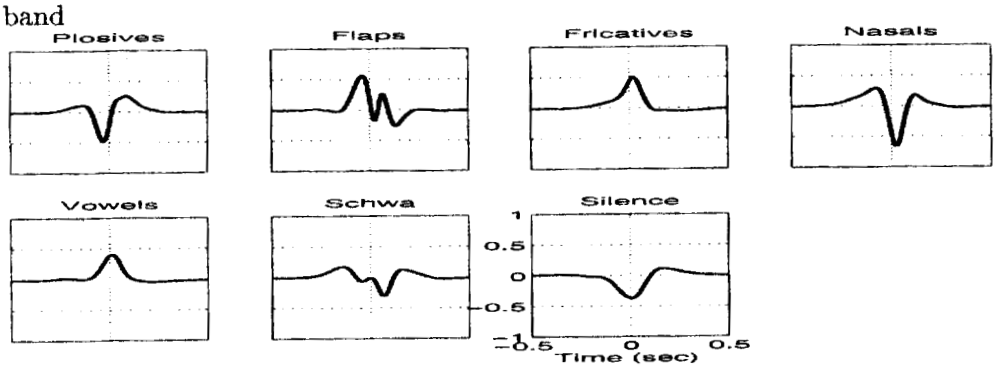

Figure 3.7: Mean Temporal Patterns of seven broad categories in 15th critical-band

Chapter 4

Broad-phonetic features : Applications

In the previous chapter, we investigated the estimation of broad-phonetic features from temporal patterns and showed benefits in recognition performance on incorporating these features in a conventional ASR system. In this chapter, we investigate further the use of broad-phonetic features in two practical applications : 1) robust Voice Activity Detection (VAD) and 2) Distributed Speech Recognition (DSR) in cellular telephony.

The broad-phonetic features that are investigated for these practical applications are the same features as described in Chapter 3 i.e. 1) based on Voicing the three categories : voiced sounds, unvoiced sounds, background; 2) based on seven broad-phonetic categories : vowels, nasals, fricatives, plosives, flaps, and background or silence.

The chapter is organized as follows. Section 4.1 describes proposed temporal patterns based system for voice activity detection (VAD). It also describes a GMM-based VAD scheme, the baseline system used for evaluating the proposed VAD. It presents some results and conclusions of this work in the Subsection 4.1.5. Section 4.2 describes the use of broad-phonetic features in DSR system. It gives a summary and conclusion of this work in Subsection 4.2.6.

4.1 Broad-phonetic features for Voice Activity Detection (VAD)

In previous section, we investigated TempoRAl PatternS (TRAPS) of log critical-band energies for the estimation of broad-phonetic features. We observed that temporal patterns are quite accurate for detecting non-speech or background portions of the speech

signal in individual bands. That leads us to investigate the use of temporal patterns for robust speech/non-speech detection or VAD . The results show the benefits of exploiting the multi-band framework of this system in band-limited noisy conditions. We compare the proposed VAD technique with another GMM-based VAD technique by using word recognition performance on the hypothesized speech segments as an evaluation metric.

4.1.1 Voice activity detection (VAD)

The separation of an audio signal into speech and non-speech segments, a process known as voice activity detection, is an important part of many speech processing systems. Efficient speech and non-speech segmentor has two benefits in speech recognition. In speech recognition systems, processing time is reduced because recognition is performed only on speech segments and recognition accuracy is improved by eliminating noisy non-speech segments that may be erroneously recognized as speech (causing insertion errors).

4.1.2 Frequency-localized temporal patterns for VAD

The temporal patterns of log critical-band energies are used to estimate the posterior probabilities of three classes : voiced speech, unvoiced speech, and non-speech, given a feature vector. The stream of non-speech posterior probabilities is then processed through a median filter or Viterbi search to produce a segmentation. The critical-band MLPs have 101 hidden units, a softmax output layer and are trained using error backpropagation algorithm with a cross-entropy error criterion. The final class posteriors are estimated by averaging the band-local log-posteriors and processing the vector of averaged log-posteriors through a softmax non-linearity. Two procedures for obtaining the final speech segments from the averaged log-posteriors were tested. In first case, the speech and non-speech segments are generated by passing the stream of non-speech posteriors through a 51-sample (0.51 s) median filter and marking frames with filtered silence posteriors less than 0.5 as speech. Hypothesized speech segments separated by less than 250 ms. of silence are merged , and the remaining speech segments are expanded by an additional 20 frames (10 at the beginning and 10 at the end of the segment). In second case, a Viterbi search is applied to the averaged log-posteriors to mark final speech segments.

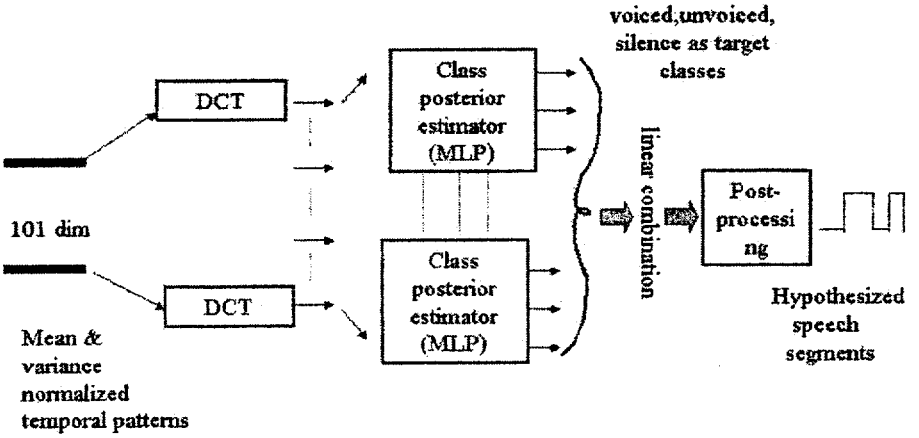

Figure 4.1: Frequency-localized temporal patterns for VAD

The speech segments produced by the Viterbi search are expanded by an additional 40 frames to keep enough surrounding context and low-energy unvoiced speech sounds around the hypothesized speech segments. The block diagram of this system is shown in Figure 4.1. We refer this system as the TRAPS-based VAD system.

4.1.3 GMM/HMM based VAD

In this section we briefly describe a baseline system for the VAD task. This system was developed at IBM [10, 11] recently. This system is different from the temporal patterns system (Section 4.1) in two aspects : 1) it uses Gaussian mixture models (GMM) for acoustic modeling of the speech and non-speech regions and 2) it uses log-energy and degree of voicing as features.

The HMM-based VAD system uses two five-state, left-to-right HMMs with no skip transitions to model speech and non-speech segments. The output distributions are mixtures of sixteen diagonal-covariance Gaussian densities. They are tied across all five states

in an HMM. The features used in the HMM system are derived from frame-level estimates of log-energy and degree of voicing. Both the log-energy and degree-of-voicing estimates are computed from 25 ms, mean-removed frames that have been weighted with a Hanning window. The log-energy features are normalized to have zero mean over an entire conversation side. The degree of voicing measure , v(t), is

$$v(t) = max \frac{r_i(t)}{r_0(t)}, t = 50, ..., 150 \tag{4.1}$$

where $r_i(t)$ is the biased autocorrelation at lag i (in sample, with a 16 kHz sampling rate) computed from frame t. The range of lags over which the maximization is performed corresponds to a pitch range of $106 - 320$ Hz. The degree of voicing measure is a very similar to the periodicity measure proposed by Thomson and Chengalvarayan [14]. The features for segmentation are computed by concatenating 17 frames of raw energy and voicing features, sorting the features of each type into increasing order and projecting the resulting $34-$ dimensional vector down to two dimensional using a discriminant, diagonalizing transform. Finally, segmentation is performed using a Viterbi search. A segment-insertion penalty controls the number and duration of hypothesized segments. The speech segments produced by the Viterbi search are expanded by 40 frames, 20 on each side of the segment, to capture any low-energy, unvoiced speech sounds at the segment boundaries and to provide sufficient acoustic context to the speech recognizer. This HMM VAD algorithm was used in IBM's entry in the SPINE-2 evaluation [10].

4.1.4 The Database and Task description

We used the SPINE-2 dataset for evaluation of the three VAD schemes. SPINE audio data is collected from a pair of speakers playing a collaborative war-game in which they locate and destroy targets on a game grid. The speakers sit in individual sound booths in which some background noise environment is reproduced. The noise environments represent a range of military environments, including a quiet office, the combat information center of an aircraft carrier, the cockpit of a helicopter and the interior of a Bradley fighting vehicle. The signal-to-noise ratio in the training and testing conditions ranges from $5-20$ dB. As of the SPINE-2 evaluation there were 17 hours of data available for training and development

testing. The SPINE-2 evaluation data comprises 128 unsegmented conversation sides with an average duration of 200 s. To measure the quality of the three VAD algorithms, the segmentations of the SPINE-2 evaluation data were generated using each algorithm. These hypothesized speech segments were recognized using an ASR system and the word error rates on the evaluation data are used as the quality metric. Earlier work on SPINE data has demonstrated that accurate speech/non-speech segmentation is crucial for obtaining accurate recognition performance [56]. The raw features used by the ASR system are 19 PLP features computed from the 16 kHz audio signal. Vocal tract length normalization is applied via linear scaling of the frequency axis up to $\pm 20\%$ prior to the Mel binning. The PLP features are normalized on a per-side basis to have zero mean and unit variance, except for c0, which is normalized to have a fixed maximum value in each segment. The feature vectors on which recognition is performed are produced by concatenating 9 frames of normalized PLP features (± 4 frames around the current frame) and projecting to a 39 dimensional feature space using a discriminant, diagonalizing transform. A Multi-pass decoding strategy is applied in recognizing the SPINE-2 evaluation data using a series of adaptation steps and two speaker- and environment-normalized acoustic models. The first acoustic model is a VTLN (Vocal Tract Length Normalization) model for which training and testing data are warped to match the characteristics of a canonical speaker. The second acoustic model is trained using speaker adaptive training (SAT). For the SAT system, both training and testing data are affine transformed into a canonical space. This process is identical to applying constrained maximum-likelihood linear regression (MLLR) to the acoustic models, but operates in the feature space. Three decoding passes are performed on each segmentation of the SPINE-2 evaluation data. The first pass uses the VTLN acoustic models and the VTLN-warped PLP features. The second decoding pass SAT uses the SAT acoustic model and single FMLLR (Full-covariance maximum-likelihood linear regression) transform per test speaker. The FMLLR transforms are trained on the one-best hypothesis from the VTLN decode. For the third pass of decoding (SAT-n), multiple regression class based FMLLR transforms are trained for each test speaker on the one-best hypotheses from the SAT-1 decoding, and recognition is performed using the SAT acoustic model. All passes of decoding are performed using IBM's rank-based

stack decoder. The language model is a class trigram trained only on SPINE training data. Words in the 5720−word recognition lexicon are modeled as sequences of decision-network clustered, context-dependent sub-phone units. Each sub-phone unit is modeled by a one-state HMM with a self-loop and a forward transition. Output distributions on the HMM transitions are modeled using mixtures of diagonal covariance Gaussian densities. The VTLN acoustic model comprises 1912 sub-phone units and 18842 mixtures, while the SAT model comprises 1537 sub-phone units and 15302 mixtures.

4.1.5 Results and Conclusions

Table 4.1 summarizes the recognition performance on decoding hypothesized speech segments from three different VAD schemes at various stages of feature and model adaptation. For all three decoding stages, recognition performance is 1 % absolute better for the TRAPS with Viterbi based segmentation than for the GMM based segmentation. By looking at the breakdown of recognition errors into substitutions, deletions, insertions and gap insertions (words hypothesized during non-speech segments), it is observed that the VAD systems differ mostly in deletions and gap insertion error rates. The GMM-based segments has low gap insertion and high deletion rates, whereas the TRAPS with median filtering VAD has high gap insertion and low deletion rates. The reason for this can be attributed to using locally mean and variance normalized temporal patterns of critical-band energies. By replacing median filtering by global Viterbi search as the final post-processing step to get hypothesized speech segments, the gap insertions in the TRAPS VAD can be reduced. Tables 4.2 and 4.3 show a breakdown of recognition performance by noise conditions for each VAD scheme. For narrow-band noises such as car, street, and office the TRAPS with Viterbi VAD works significantly better than the GMM-based system due to its multi-band framework. For carrier noise, which includes bursts of relatively high energy speech in time, the gap insertion rate for the TRAPS VAD is higher than for the GMM-based VAD scheme, which results in its poor performance in this particular condition. This indicates that the TRAPS VAD does not handle time-localized, non-stationary noisy conditions well.

VAD	WER	sub.	del.	ins.	gap ins.
VTLN					
HMM	41.5	22.2	13.5	4.8	0.9
TRAPS+Median	42.4	23.4	9.6	5.6	3.8
TRAPS+Viterbi	40.6	22.8	11.2	5.2	1.4
SAT-1					
HMM	34.3	17.0	12.8	3.6	0.9
TRAPS+Median	34.6	18.0	8.8	4.1	3.7
TRAPS+Viterbi	33.5	17.4	10.8	3.8	1.5
SAT-n					
HMM	33.7	16.6	12.8	3.3	0.9
TRAPS+Median	33.8	17.7	8.7	3.9	3.6
TRAPS+Viterbi	32.7	17.1	10.5	3.6	1.5

Table 4.1: % Recognition performance for each segmentation of the SPINE-2 evaluation data and each pass of decoding. The word error rate is broken down in terms of substitutions, deletions, insertions and gap insertions. Gap insertions are caused by segmentation errors: the recognizer hypothesizes words during non-speech segments. For all three decoding passes, the best accuracy is obtained with the TRAPS+Viterbi segmentation.

4.2 Broad-phonetic features in cellular telephony for DSR

In this section, we evaluate broad-phonetic features, estimated from the temporal patterns, for DSR. We propose the estimation of broad-phonetic features at the server-end. The broad-phonetic features are used as an additional feature-stream in the recognizer.

4.2.1 Distributed Speech Recognition (DSR) for cellular telephony speech

This section describes briefly different processing blocks of a distributed speech recognition (DSR) system, which was submitted to the AURORA-2 evaluation [1]. The AURORA-2 task comprises English (TIDIGITS) and European languages (Spanish, Italian, and Finnish) connected digits. There were three different testing conditions which were based on the degree of mismatch between training and testing environments. The recognizer was fixed in configuration by an ETSI committee [31, 17] where each digit was modeled using a whole-word model. The models were based on 16−state, 3−mixture whole word HMMs. The silence model had 3-state and 6-mixture per state. Also, a one-state short

Channel A speakers						
VAD	noise	WER	sub.	del.	ins.	gap ins.
HMM	Bradley	41.9	20.4	14.2	5.9	1.4
	office	32.6	14.3	15.4	2.2	0.8
	carrier	24.3	12.9	6.6	3.5	1.4
	car	32.3	14.0	15.5	2.3	0.4
TRAPS	Bradley	42.0	20.5	14.8	5.6	1.1
	office	28.6	15.1	10.2	2.4	0.9
	carrier	27.6	13.1	6.5	3.6	4.3
	car	26.8	14.3	9.1	2.6	0.8

Table 4.2: SAT-n recognition performance for SPINE-2 evaluation data, by noise condition and segmentation. The table is divided into two parts corresponding two speaker groups in the data. Because there are only sixteen speakers per group, meaningful comparisons can only be made within a speaker group.

pause model is used and is tied with the middle state of the silence model. The processing of features for noise compensation was divided into two parts.

4.2.2 Handset-side processing

In the proposed DSR system [1], the following techniques were incorporated for compensating the various mismatched testing conditions. The feature processing at the handset-side is as follows:

1. Wiener filtering was applied in the power spectral density domain as a first step for additive noise compensation.

2. RASTA like filtering was applied on 23 Mel critical-band energies, which were computed on the Wiener filtered power spectral densities, as a channel compensation technique. The filter-bank computation is described in detail in Appendix H.

3. Finally 15 cepstral coefficients were computed on RASTA filtered Mel critical-band energies.

Channel B speakers						
VAD	noise	WER	sub.	del.	ins.	gap ins.
HMM	helicopter	40.9	21.6	14.0	3.5	1.8
	F-16	34.8	20.8	9.4	3.8	0.9
	quiet	34.1	16.1	13.9	3.3	0.9
	street	32.7	16.9	12.3	3.1	0.3
TRAPS	helicopter	41.8	22.8	13.1	4.6	1.2
	F-16	35.3	20.6	9.7	4.1	0.9
	quiet	34.2	17.1	12.3	3.4	1.4
	street	32.0	17.5	9.7	3.7	1.0

Table 4.3: SAT-n recognition performance for SPINE-2 evaluation data, by noise condition and segmentation. The table is divided into two parts corresponding two speaker groups in the data. Because there are only sixteen speakers per group, meaningful comparisons can only be made within a speaker group.

4.2.3 Server-side Processing

The server-side feature processing further compensates for mismatched noise and channel testing conditions. The following techniques were proposed for feature processing at the server-side:

1. Adaptive mean and variance normalization was applied on the received 15 cepstral features. The estimates for local mean and variance were updated only for speech-detected frames using information from a voice activity detector.

2. Dynamic features (15 delta and 15 double-delta) were computed on the normalized cepstral features.

The dimension of final feature vector was 45 after including dynamic features. We refer these features as robust MFCC features. More details of the proposed DSR system can be found in [1].

4.2.4 Aurora Dataset

Aurora-2 is the TIDIGITS dataset artificially distorted by additive noise and simulated channel distortion. Two kinds of training are used: clean speech training (denoted as

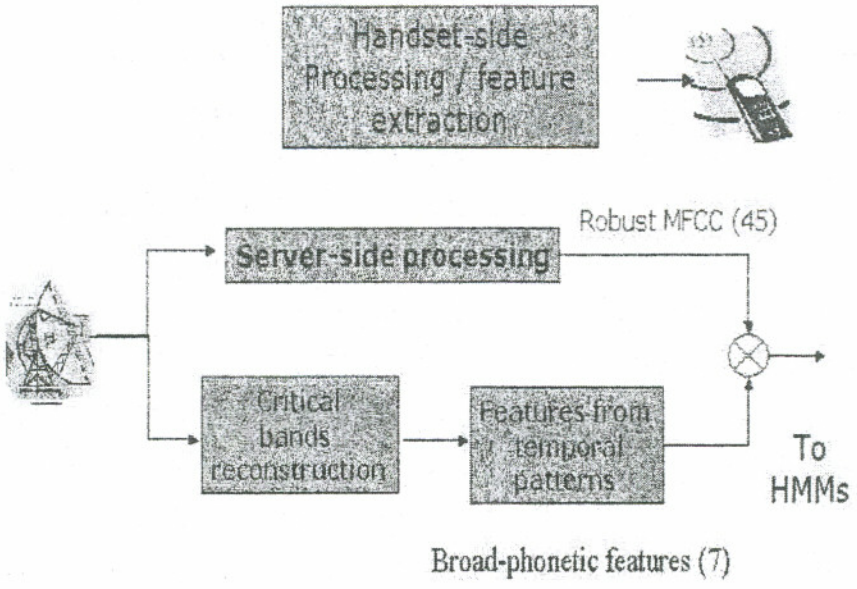

Figure 4.2: Distributed Speech Recognition (DSR) system

Clean), and training by using both clean and noisy speech (denoted as Multi). For each training set, three tests are realized : (A) - matched training and testing noises, (B) - mismatched training and testing additive noises, and (C) - mismatched training and testing additive noises and channel (convolutive) distortions. Aurora-3 is a set of multi-language SpeechDat-Car databases recorded in-car under different driving conditions with close-talking and hands-free microphones. Three recognition experiments are defined for Aurora-3 with different levels of training and testing mismatch : Well-matched, Medium -mismatched, and Highly-mismatched.

For Italian, the training and testing conditions cover stop-motor-running, town-traffic, low-speed-rough-road, and high-speed-good-road driving conditions. The data is recorded using close-talking (CT) and hand-free microphones (HF) . The Well-matched training and testing datasets have CT and HF recorded data from all the driving conditions. The Medium-mismatched training set has HF recorded data covering three driving conditions

: stop-motor-running, town-traffic, and low-speed-rough-road. Whereas the Medium-mismatched testing condition has HF recorded high-speed-good-road condition. The Highly-mismatched training set has CT recorded all the driving conditions whereas testing condition has HF recorded data in the town-traffic, low-speed-rough-road, and high-speed-good-road conditions [68].

For Finnish, the data is recorded in 0 kmh and engine on (low-noise), $40 - 60$ kmh car-speed (low-noise), $40 - 60$ kmh and window open (low-noise), $100 - 120$ kmh (high-noise), and $100 - 120$ kmh with music on (high-noise) driving conditions. The data is recorded with CT and HF microphones. The Well-matched training and testing datasets cover all the driving conditions with both HF and CT microphone recorded data. The Medium-mismatched training dataset covers 70% of quiet and low noise (with HF microphone) conditions whereas Medium-mismatched testing dataset covers 30% of high-noise driving conditions. The Highly-mismatched training set comprises 70% of all the driving conditions with CT microphone whereas Highly-mismatched testing dataset has 30% of low-noise and high-noise conditions, recorded with HF microphone [48].

For Spannish, the data is recorded in quiet (stop-motor-running), low-noise (town-traffic+low-speed-rough-road), and high-noise (high-speed-good-road) driving conditions again with HF and CT microphones. The Well-matched training and testing datasets cover all the driving conditions with data recorded with both HF and CT microphones. The Medium-mismatched training set has HF recorded in quiet and low-noise conditions whereas the testing condition has HF recorded high-noise conditions. The Highly-mismatched training set has CT recorded all the driving conditions whereas testing set has HF recorded low- and high-noisy conditions [69].

4.2.5 Broad phonetic features estimated from temporal patterns in DSR

In the DSR framework, we propose estimation of broad-phonetic features from temporal patterns of log critical-band energies of 15 bands, which are reconstructed from the received 15 cepstral features (prior to adaptive mean and variance normalization and dynamic feature computation), at the server-side. 50 DCT components of temporal trajectories of log critical-band energies are used to estimate posteriors of seven broad phonetic classes :

vowels, plosives, fricatives, nasals, flaps, schwa and silence (Table 3.1). A set of multilayer perceptrons are trained using the backpropagation algorithm with cross-entropy criterion and these seven categories as target classes in individual critical-bands. The estimated posteriors from the multilayer perceptron (MLP) in each critical-band are then used as input features to a Tandem MLP for final posterior estimation. The seven outputs from this MLP are taken out without applying the softmax non-linearity and concatenated with the robust MFCC features. The components of the resulting 52-dimensional feature vector are decorrelated and used in an HMM based back-end recognizer. Table 4.5 and Table 4.6 show the results of speech recognition experiments. The results show that by incorporating the seven additional features with cepstral features a $1 - 3\%$ absolute reduction in % word error rate is obtained in various testing conditions of the TIDIGITS task. Whereas for the SpeechDat car condition a 0.3 to 0.7 % absolute reduction in word error rate is achieved but this improvement does not hold good on the highly mismatched condition. The improvements in the most of the widely varying testing conditions in TIDIGITS and SpeechDat car tasks show that seven features largely generalize well across different environments. Note that the training of the posterior estimators (MLPs) are entirely performed on an independent dataset i.e. the noisy TIMIT dataset.

Table 4.4 shows the effect of using shorter temporal context (190 ms) on the recognition performance. It can be observed that the gain in the recognition performance is consistent with reduced system delay (90 ms) in a real-time applications however longer context (1 s) further improves the recognition performance but introduces more delay (0.5 s) in a real-time applications.

	multi-style		clean-style	
	robust MFCC	with TRAPS	robust MFCC	with TRAPS
Well matched	9.0	8.0	14.4	13.0
Medium mismatched	9.4	7.9	14.8	14.2
Highly mismatched	9.9	9.0	14.5	13.6
overall	9.4	8.3	14.6	13.6

Table 4.4: Word error rates (%) on Aurora-2, TIDIGITS data, by training data (multi-style vs. clean) and by feature set (noise-robust MFCCs alone,or robust MFCCs augmented with TRAPS-estimated broad-phonetic features, context window = 190 ms).

	robust MFCC	with TRAPS
Well matched	3.3	3.0
Medium mismatched	8.5	7.8
Highly mismatched	13.0	14.2

Table 4.5: Word error rates (%) on Aurora-3, SpeechDat-car data, by testing conditions by feature set (noise-robust MFCCs alone,or robust MFCCs augmented with TRAPS-estimated broad-phonetic features, context window = 1 s).

	multi-style		clean-style	
	robust MFCC	with TRAPS	robust MFCC	with TRAPS
Well matched	9.0	7.2	14.4	11.1
Medium mismatched	9.4	7.1	14.8	11.0
Highly mismatched	9.9	8.3	14.5	12.1
overall	9.3	7.4	14.6	11.3

Table 4.6: Word error rates (%) on Aurora-2, TIDIGITS data, by training data (multi-style vs. clean) and by feature set (noise-robust MFCCs alone,or robust MFCCs augmented with TRAPS-estimated broad-phonetic features, context window = 1 s).

4.2.6 Conclusions

We obtain consistent gains in recognition performance on combining broad-phonetic features and conventional cepstral features. We achieve 0.3 to 3.0 % absolute reduction in word error rates on most of the conditions of AURORA datasets. The number of parameters used in the temporal pattern system is around 98040. The temporal patterns system has an algorithmic latency of 190-500 ms. We showed that it is feasible to reconstruct critical-band energies from the transmitted cepstra for a temporal patterns system at the server-end without disturbing the terminal-side feature processing. Thus an additional set of features can be extracted by utilizing the information from frequency-localized temporal patterns of the speech signal and improved recognition performance can be achieved.

Chapter 5

Universal TempoRAl PatternS (UTRAPS) based system

In the previous chapter, band-specific broad-phonetic categories were investigated for feature estimation. The broad-phonetic categories were obtained by clustering mean temporal patterns of phones in individual critical-bands (Section 3.4). By looking at the confusion matrices of broad-phonetic categories, it can be seen that the nasals and vowel categories are most often confused with each other in lower frequency bands, while the nasals and plosive categories are confused with each other in higher frequency bands. Likewise fricatives and plosives are confused with each other in lower bands while in higher bands fricatives and vowels are confused with each other (Figures 3.1, 3.2, 3.3, and 3.4) This is due to the fact that some of the phones change cluster affiliation in different frequency bands. This depends often on the type of speech activity present in that particular band. For example fricative /sh/ has a high concentration of energy in higher frequency regions which makes temporal patterns of /sh/ in these regions have a peak at the center , much like vowels have (Appendix B).

These observations lead us to define clusters (new categories) based on distinct speech activities in different frequency bands. Instead of defining broad-phonetic categories by clustering temporal patterns independently in individual critical-bands, we propose new band-independent broad-phonetic categories which can be obtained from clustering temporal patterns from all the critical-bands. These clusters of temporal patterns represent distinct speech activities or events manifested in log critical-band energies of the speech signal. In our work described in the previous chapters, we have used independent posterior

estimators in each frequency band. In this chapter, We propose a new system which uses single, universal class-posterior estimator for estimating posterior probabilities of speech-events categories. These posteriors are used as features in the Tandem speech recognizer for final recognition.

This approach is a step towards providing more sharable acoustic information by defining consistent and distinct acoustic regions in the time-frequency representation of the speech signal. We show that the new features are robust across various noisy environments. They are also complementary in nature with respect to short-term spectral features such as cepstral coefficients.

The chapter is organized as follows. In Section 5.1, we describe the clustering of temporal patterns of phones into speech-event categories. Section 5.2 describes a Universal TempoRAl PatternS (UTRAPS) system for feature estimation. Section 5.4 shows the performance of the UTRAPS features in ASR. Section 5.5 gives the summary and conclusion of this work.

5.1 Clustering the temporal patterns to define speech events

This section describes how we obtain band-independent speech-event categories for feature estimation. The mean temporal patterns are computed using 101-sample (1 s in temporal context) mean subtracted, variance normalized, and hamming windowed temporal patterns of log critical-band energies for each phone on a labeled dataset (TIMIT) [60]. The details of computation of mean temporal patterns is described in Appendix B. An agglomerative hierarchical clustering technique [55] is used to obtain new categories. A correlation measure is used as a similarity measure for clustering the temporal patterns. It is given by,

$$S(x,y) = \frac{\sigma_{xy}^2}{\sigma^x \sigma^y}$$

For number of TIMIT phones = 56 and number of bands = 15, the agglomerative clustering procedure starts with $56 * 15$ mean temporal patterns as the singleton clusters. At each iteration the two closest together clusters are merged with each other. This is performed iteratively until the number of clusters reach finally to 9 clusters. The matlab code used

for generating these cluster is given in the Appendix F.

This stopping point in clustering is based on some heuristics - had we continued further in the clustering, final clusters would not be able to capture the distinct 'flap' or 'schwa' mean temporal patterns, and we have chosen to keep these distinct patterns in our inventory of speech events. The final nine clusters are shown in the Figure 5.1. They represent distinct frequency-localized temporal patterns of the speech signal.

1. SILENCE-like - e.g. mean temporal patterns of the silent speech regions for all the critical-bands.

2. PLOSIVE-like - e.g. mean temporal characteristics of most of the plosives which shows a dip in energy off-center to the left as plosives are usually preceded by a stop-closure. This pattern is also found in the 8-9 bands of the glide /w/ and in the 5-6 bands of the glide /y/.

3. NASAL-like - e.g. mean temporal characteristics of a nasal /em/ in the lower 1-7 bands.

4. GLIDE-like - a peak in energy off-center to the right which represents mean temporal characteristics of glides /r/, /w/ in 1-6 bands , and of /y/ in the lower 1-4 bands and the higher 10-15 bands.

5. LOW VOCALIC ENERGY - a small burst in the energy off-center to the left followed by a small dip in energy off-center to the right. This pattern is often seen in the middle bands of vowel /iy/ and higher bands of vowel /axr/ etc.

6. SCHWA-like - mean temporal characteristics of the schwa sounds such as /ix/, /ax/ in most of the bands.

7. FLAP-like - mean temporal characteristics of the flap-like sounds such as /nx/, /dx/ in most of the bands.

8. HIGH VOCALIC ENERGY - mean temporal characteristics of most of the vowels such as /aa/, /ae/ in 1-15 critical-bands, of a vowel /iy/ in the lower 1-5 and upper

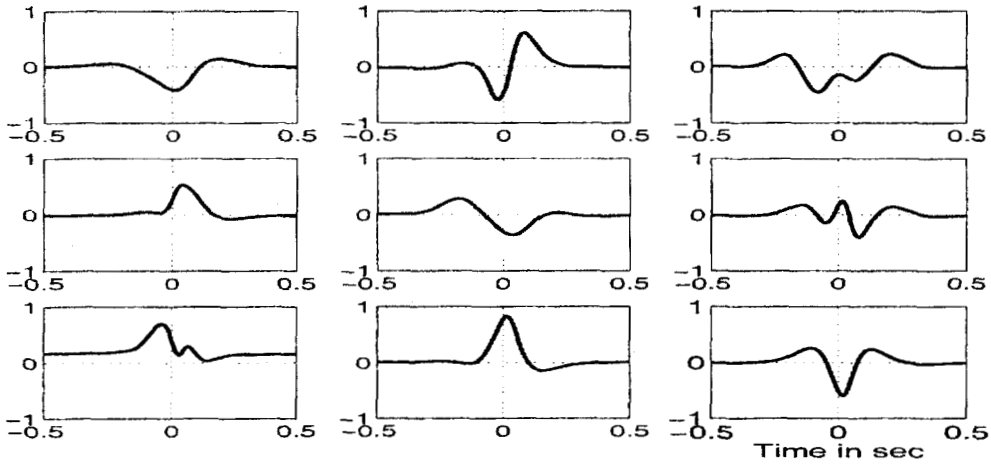

Figure 5.1: Mean temporal patterns of nine speech-events

10-15 bands, and of a dipthong /oy/ in the lower 1-6 bands. This pattern is also found in the higher 12-15 bands of fricatives such as /zh/, /sh/ etc.

9. FRICATIVE-like - The nineth cluster represents mean temporal patterns of most of the fricatives such as /s/, /sh/, /f/ and of affricatives such as /ch/, /jh/ in the lower 1-9 bands. It also represents mean temporal characteristics of nasals /m/, /n/ in the higher bands

The grouping of the phones based on these nine clusters in individual bands is shown in the Appendix E.

5.2 A Universal TemPoRAl PatternS (UTRAPS) system

A single posterior estimator (MLP) is trained with 9 clusters as the target categories. The 50 lowest DCT components computed from 101-point, mean and variance normalized temporal patterns of log critical-band energies are used as input to the MLP. The MLP has 101 hidden units. The new class-labels are obtained by mapping mean temporal patterns of phones from each of the critical-bands to one of the 9 speech-event clusters using a similarity (correlation coefficient) measure. The phone-labels of the training dataset are obtained by forced-alignment or manual-transcription. The MLP is trained with backpropagation

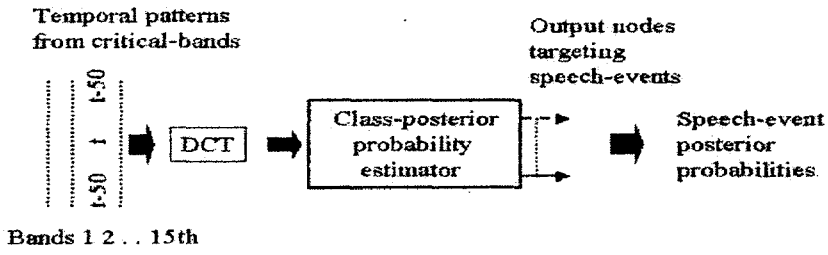

Figure 5.2: Block diagram of the UTRAPS system

algorithm with cross- entropy as the error criterion. At every frame, class-posteriors are obtained by doing a forward pass for each temporal pattern from the 15 frequency bands through the MLP. The $9 * 15$ class-posteriors are used as input features to a Tandem system for final recognition. The block diagram of the UTRAPS system is shown in Figure 5.2.

5.3 Speech-event for phone and broad-phonetic category classification

In this section we compare phone classification rates for 56 TIMIT phones based on speech-event features and based on the previously described broad-phonetic features. We used temporal patterns of 15 Mel filterbank, which are described in Appendix H, for this experiment. The $9 * 15$ features are used as input to the Tandem MLP which targets 56 phones for classification. The UTRAPS MLP is trained with 101 hidden units and the Tandem MLP is trained with 200 hidden units. Similarly the $7 * 15$ broad-phonetic features are used as input to the Tandem MLP for phone classification rate. In this system, the band-specific MLPs are trained with 101 hidden units and the Tandem MLP is trained with 200 hidden units. All the MLPs are trained on the TIMIT dataset using the backpropagation algorithm with a cross-entropy error criterion.

Table 5.1 shows that similar or lower phone frame-level error rates are obtained for most of the phones using speech-event UTRAPS features. Table 5.2 shows frame-level classification error rates for the seven broad-phonetic categories obtained using speech-event UTRAPS features and broad-phonetic TRAPS features. The speech-event features give similar classification error rates for vowels, nasals, fricatives, and plosives categories than broad-phonetic features that are estimated independently from each frequency band. For flap, schwa, and silence categories they perform worse than broad-phonetic features.

5.4 Speech-event features in ASR

The speech-event features are evaluated on the small vocabulary continuous digits tasks : OGI-Digits and AURORA tasks. Description of the datasets can be found in Section 2.3 and Section 4.2.4.

For OGI-Digits task, we used 50 DCT components of temporal patterns of 15 logBark filterbanks energies. The used filterbank is described in Appendix H. The UTRAPS MLP is trained with 101 hidden units and 9 target classes on an independent dataset i.e. OGI-stories. The $15 * 9$ speech-event posteriors from UTRAPS MLP are taken before soft-max non-linearity and used as the input features to a Tandem MLP that estimates posteriors for the seven broad-phonetic categories posteriors. For this task, since 'flap' was not present we used six broad-phonetic categories for posterior estimation at the output of the Tandem MLP. The Tandem MLP is trained with 200 hidden units. These six posteriors output from the Tandem MLP are used as features and evaluated in conjunction with PLP cepstral features. The posteriors are combined at the class-posterior level as described earlier in section 3.7. For estimating class-posteriors from cepstral features, 9 frames of 24 PLP cepstra along with dynamic features (delta+double delta) are used as input to a MLP for estimating 29 phone posteriors. These posteriors are concatenated with 6 broad-phonetic posteriors. The 35 concatenated posteriors are gaussianized, decorrelated, reduced in dimension to 29 (using the PCA transform), and finally used as input to the HMM based back-end (Section 3.7). We had two cases based on the dataset used for

Phone	TRAPS	UTRAPS	Phone	TRAPS	UTRAPS
b	82.3	**80.1**	m	56.4	**55.4**
d	89.8	**87.6**	em	99.9	99.9
g	86.1	**80.9**	n	56.7	**57.3**
p	**71.7**	72.1	nx	**80.6**	82.1
t	61.6	**60.5**	ng	86.9	**85.0**
k	59.2	**55.7**	en	83.4	83.6
dx	**44.7**	49.8	l	50.8	**50.2**
bcl	77.2	**77.0**	el	69.6	**65.4**
dcl	**77.0**	77.9	r	51.9	**51.6**
gcl	**91.4**	93.1	w	**54.8**	55.8
pcl	79.8	**77.0**	y	79.5	**77.2**
tcl	**63.6**	64.9	hh	88.9	**87.5**
kcl	62.5	**62.4**	hv	69.8	**69.0**
jh	75.0	**72.9**	iy	33.1	**31.4**
ch	75.9	**72.9**	ih	74.5	**72.8**
s	32.8	**31.4**	eh	69.4	**68.2**
sh	47.6	**45.6**	ey	54.0	**51.5**
z	69.3	**66.9**	ae	48.8	**48.6**
zh	91.4	**90.0**	aa	**58.6**	58.9
f	69.1	**68.7**	aw	80.1	**77.6**
th	99.6	**99.4**	ay	50.2	**47.3**
v	83.5	83.5	ah	76.4	**74.9**
dh	83.6	**83.0**	ao	59.5	**58.3**
oy	75.6	**71.5**	ow	74.2	**69.6**
uh	97.0	**96.0**	uw	71.3	**66.3**
er	63.7	**59.7**	axr	**65.9**	66.4
ax	68.2	**68.0**	ix	**56.4**	59.2
si	**7.1**	7.9	q	73.5	**73.2**

Table 5.1: % Frame-level phone classification error rates obtained at the output of the Tandem MLP, using the UTRAPS features and using broad-phonetic TRAPS features estimated independently from individual critical-bands. Results are obtained on the TIMIT dataset. The minimum error is shown in bold type

training the Tandem MLP. In the first case, we trained it on the target dataset i.e. OGI-Numbers (contains OGI-Digits) and in the second case, we trained it on an independent dataset i.e. OGI-stories.

We compare the result with six broad-phonetic features that are obtained from the previously described TRAPS system (Section 3.4). In the TRAPS system, six broad-phonetic posteriors are obtained from each frequency band independently at the output of band-specific posterior estimators. The estimated $15 * 6$ posteriors are mapped to 6 final broad-phonetic posterior using a Tandem MLP. These 6 posteriors are evaluated in conjunction with PLP cepstral features as the speech-event features are. We also compare this result with phone-based TRAPS features. The phone-based TRAPS features are obtained from each frequency band independently at the output of band-specific posterior estimators using the 29 phones as the target categories. Table 5.4 shows that we obtain similar gain in recognition performance using the UTRAPS speech-event features or the both phonetic and broad-phonetic TRAPS features in conjunction with PLP cepstral features. In this experiment, the Tandem MLP is trained on the target dataset, OGI-Numbers, and the UTRAPS MLP and band-specific TRAPS MLPs are trained on OGI-Stories.

Table 5.5 shows the performance of the features when the Tandem MLP, the UTRAPS MLP, and the band-specific MLPs, all are trained on OGI-Stories. In this case the UTRAPS features outperform the broad-phonetic category features. The UTRAPS features are significantly better (at 98 % confidence interval) than both phone-based and broad-phonetic TRAPS features. For confidence measure, we applied Matched Pairs Sentence-Segment Word Error (MAPSSWE) test [23].

Table 5.3 shows that speech-event features performance when they are used alone in the Tandem recognizer. In this experiment, $15*9$ features are used as input to the Tandem MLP. The Tandem MLP estimates 29 phone posterior probabilities and these posteriors are gaussianized, decorrelated, and used as input to the HMM back-end recognizer. The result is compared with that obtained from $15 * 6$ broad-phonetic TRAPS features. The $15 * 6$ broad-phonetic TRAPS features are used as input to the Tandem MLP for estimating again 29 phone posterior probabilities. These phone posteriors are gaussianized,

decorrelated, and used as input to the HMM back-end recognizer for WER computation.

For the AURORA task (Section 4.2), $15 * 9$ speech-event features are evaluated in conjunction with the robust MFCC features in the DSR framework. In this experiment, speech-event features are estimated from the reconstructed 15 Mel critical-band energies at the server-end. The computation of reconstructed Mel critical-band energies is described in Appendix H. The 50 DCT components of 101-sample temporal patterns are used as input to the UTRAPS MLP. The MLP is trained with 101 hidden units and 9 target categories at the output layer. Every 10 ms, $15 * 9$ class-posteriors are estimated using temporal patterns from all the frequency bands. They are mapped to 7 broad-phonetic class posteriors using the Tandem MLP. The Tandem MLP is trained with 200 hidden units and 7 target categories at its output layer. These 7 posteriors are taken from the Tandem MLP without applying the softmax non-linearity and then concatenated with the robust MFCC features at the server-end. The description of the robust MFCC features can be found in Section 4.2.1. The final 52 dimensional feature vector is decorrelated (using a whitening transform) and is used as input to the HMM based back-end recognizer. The description of this back-end recognizer can be found in Section 4.2.1. The speech-event UTRAPS features are compared with previously described broad-phonetic TRAPS features used in DSR (Section 4.2). Table 5.6 and Table 5.7 show that the speech-event UTRAPS features give very similar word recognition performance on the Aurora-2 and Aurora-3 datasets as the broad-phonetic TRAPS features give.

Note that the UTRAPS system has only 5959 $(50 * 101 + 101 * 9)$ parameters as opposed to the TRAPS system which has 86355 (N*$(50 * 101 + 101 * 7)$ where N= 15 bands) parameters.

5.5 Summary and Conclusion

We proposed band-independent categories for robust feature estimation. These categories characterize distinct speech events in the frequency-localized temporal domain of the speech signal. These categories exploit the redundancy of events across different frequency bands as well as across different phones. Each category has more data for training

Broad-phonetic categories	UTRAPS %error	TRAPS % error
Plosives	31.6	32.8
Flaps	58.5	52.9
Fricatives	27.5	27.9
Nasals	42.3	42.2
Vowels	11.0	12.6
Schwas	60.9	57.3
Silence	10.7	9.3

Table 5.2: % Frame-level broad-phonetic classification error rates obtained at the output of the Tandem MLP, using the UTRAPS features and using broad-phonetic features estimated independently from individual critical-bands. Results are obtained on the TIMIT dataset

PLP cepstral features	UTRAPS features	broad-phonetic TRAPS features
5.0	8.4	8.3

Table 5.3: Word Error Rates (%) on OGI-Digits : PLP cepstral features, the UTRAPS features, or the broad-phonetic TRAPS features.

PLP cepstral features	UTRAPS features	broad-phonetic TRAPS features	phone-based TRAPS features
5.0	4.3	4.3	4.3

Table 5.4: Word Error Rates (%) on OGI-Digits : PLP cepstral features alone, PLP cepstra augmented with UTRAPS features, or cepstra augmented with various TRAPS features.

PLP-cepstral features	with UTRAPS features	with broad-phonetic TRAPS features	with phone-based TRAPS features
7.4	6.1	6.9	7.3

Table 5.5: Word Error Rates (%) on OGI-Digits : PLP cepstral features alone, PLP cepstra augmented with the UTRAPS features, or PLP cepstra augmented with various TRAPS features. In this case, MLPs training was entirely done on an independent dataset, OGI-Stories.

	robust MFCC	with broad-phonetic TRAPS features	with UTRAPS features
Well matched	8.9	7.2	7.4
Medium mismatched	9.4	7.1	7.2
Highly mismatched	9.9	8.3	8.2

Table 5.6: Word error rates (%) on the Aurora-2, TIDIGITs data and by feature set (noise-robust MFCCs alone, robust MFCCs augmented with broad-phonetic TRAPS features or the UTRAPS features), with context window = 1 s.

	robust MFCC	with broad-phonetic TRAPS features	with UTRAPS features
Well matched	3.3	3.0	2.9
Medium mismatched	8.5	7.8	7.8
Highly mismatched	13.0	14.2	11.2

Table 5.7: Word error rates (%) on the Aurora-3, by testing conditions and by feature set (noise-robust MFCCs alone, robust MFCCs augmented with broad-phonetic TRAPS features or robust MFCCs augmented with the UTRAPS features), with context window = 1 s.

as it uses temporal patterns from all the critical-bands as opposed to using band-specific patterns for training (as used in the broad-category features). These categories are defined by the clusters obtained from an agglomerative clustering technique which is applied to mean temporal patterns of context-independent phones. The used clustering techniques is only one of many alternatives which may be employed for deriving the speech events from speech data so this particular technique for deriving speech events is very much an open issue.

We proposed a universal class posterior estimator for speech-events feature estimation. We showed that similar or better phone classification rates can be obtained using speech-event features versus using the broad-phonetic features. A consistent and significant gain (at 98 % confidence interval) in recognition performance is achieved by augmenting these features with cepstral features. This indicates features are complementary to short-term cepstral features. We compared speech-event UTRAPS features with broad-phonetic and phone-based TRAPS features. We showed UTRAPS features give similar gain as TRAPS features, when used in conjunction with PLP cepstral features. We showed UTRAPS features generalize better than both sets of TRAPS features. A significant reduction in the number of parameters is achieved on using a UTRAPS system over earlier proposed TRAPS system.

Chapter 6

Incorporating broader-frequency context

In the previous chapters, we investigated various categories such as context-independent phones, band-dependent broad-phonetic classes , and band-independent speech-events categories for noise-robust feature estimation using the temporal patterns of the speech signal. The temporal patterns of individual critical-bands were treated as independent feature streams for estimating posterior probabilities of classes (Section 3).

To know the nature and the spread of signal interaction across the critical-bands, we investigate the effect of broadening the frequency context on the estimated features and ASR. The frequency context is gradually increased from a single critical-band to several critical-bands by using temporal patterns jointly from adjacent critical-bands as input to class-posterior estimators.

Section 6.1 analyzes two-dimensional principal components (PCAs) from the time-frequency feature space. It shows the effect of broadening the frequency context by incorporating temporal patterns from several adjacent critical-bands on the estimated features. Section 6.2 describes the experimental setup. It shows the effect of broadening the frequency context on the recognition performance using the features in ASR. Section 6.3 describes a joint transform using a DCT basis as a replacement for joint PCAs. It also describes the results obtained from using the joint DCT transform and compares it with that obtained from joint PCAs. The chapter ends with summary of this work in Section 6.4.

6.1 On incorporating broader frequency context

In contrast to estimating features from the entire frequency spectrum in a single step, previously proposed TRAPS systems estimate features (class-posterior probabilities) from 1 s (or 101−samples) temporal trajectory of log critical-band energies from a single critical-band (Sections 1.3, 3.4) in its first stage of feature processing. In such a system, only within-channel (critical-band) cues are used and across-channel cues are ignored for estimating class (phone) posteriors in its initial stage of processing. To study the effect of incorporating the local across-channel cues on ASR, we investigate combining temporal dynamics of patterns from several adjacent critical-bands in a TRAPS based system. The results also show the effect of broadening the frequency-context on ASR.

6.1.1 Evidence of across-channel processing

S. A. Shamma et. al. work on Spectro-Temporal Modulation Transfer function (STMF) [35, 65] indicates spectral-temporal modulations play significant role in the perception of speech signal. They measured the sensitivity of human subjects to a range of spectro-temporal modulations. Their results indicate that STMF exhibits low-pass characteristics in both the dimensions, spectral and temporal. They also investigated STRF (spectro-temporal response field) which represents the spectro-temporal patterns that best excites the auditory cells. They found that the temporal selectivity of these cells ranges from rapid (over 16 Hz) to very slow (under 2 Hz) whereas spectral selectivity ranges from 0.5 to 2 octaves. This indicates that receptive fields gather information not only from longer temporal-span (order of 250 ms) of the signal but also from a range of spectral frequencies. These studies indicate that auditory system utilizes not only within-clannel cues but also across-channel cues for identifying a particular sound. The other evidence for across-channel processing is comodulation masking release (CMR), known as a psycoacoustic phenomenon. The work on CMR by Hall et. al. [16] indicates that initial stage of auditory filtering is followed by across-channel processing. They found that for masker bandwidths larger than the critical-band signal-thresholds in the modulated-noise condition go down than in the reference condition where unmodulated bandpass noise was used as the masker.

This difference between the modulated and the unmodulated condition diminishes when the masker bandwidth is smaller than the critical band-width. Based on these findings Hall et. al. suggested [30] that the energy-envelops fluctuations between different auditory channels are compared and thereby across-channel cues are used in detecting the signal during auditory processing [33].

6.1.2 Computation of joint PCAs

To incorporate broader frequency-context, joint principal components (PCAs) are estimated from a two dimensional time-frequency tile of log critical-band energies. The temporal context is kept the same as before : 1 s (101−samples from 8 kHz speech). The frequency context is increased from one critical-band to five critical-bands in steps of one critical-band. The joint PCAs are used as input features to successive phone posterior estimators (MLPs).

For joint principal components computation, the 101-sample long temporal patterns are concatenated from the adjacent N critical-bands where N is the used frequency context. These temporal patterns are mean and variance normalized prior to estimating joint PCAs. For the $NX101$ sample long vector total covariance is computed using a dataset. We used an independent TIMIT dataset for this computation. The joint PCAs computation can be described mathematically as follows,

$$Y_N^n = \frac{(X_N^n - m^n)}{\rho^n};$$

Where vector Y_N^n represents a normalized 101-sample temporal pattern of Nth critical-band at time n, X_N^n represents 101-sample temporal vector of log critical-band energies of the same critical-band formed by concatenating 50 samples in the past and in the future of the current frame at time n. At every 10ms, the temporal pattern X_N^n is mean subtracted and variance normalized. The local mean m_N^n and the variance ρ_N^n of temporal pattern X_N^n in Nth critical-band are computed as follows,

$$m_N^n = \frac{1}{101} \sum_{k=1}^{k=101} X_N^n(k)$$

$$\rho_N^n = \frac{1}{101} \sum_{k=1}^{k=101} (X_N^n(k) - m_N^n)^2$$

The vector Y^n is formed by concatenating the temporal patterns from N adjacent critical-bands. C is the total covariance matix computed over vector Y. Singular value decomposition of matrix C gives eigenvalue matrix S and eigenvector matrix U. The columns of the matrix U represent joint principal components (PCAs) of the concatenated temporal patterns. This computation can be described mathematically as follows,

$$Y^n = [Y_1^n, Y_2^n, \cdots, Y_N^n]$$
$$C = cov[Y^n]$$
$$C = U * S * U'$$

For N= 3 critical-bands, the joint PCAs are shown in Figures 6.5 and 6.6. It can be seen that the joint PCAs can be approximated by combinations of cosine functions of different modulation frequencies. For example in Figure 6.5 each principal component can be viewed as a concatenation of three segments of each 101-samples which can be approximated by a cosine function of appropriate frequency. All three cosine components are inphase with each other. Whereas in Figure 6.6, each principal component can be viewed as a concatenation of two out of phase 101-sample cosine segments. These joint PCA basis vectors in effect do spectral subtraction (Figure 6.6) along the frequency axis, which essentially incorporates information about local spectral-slope into the projected feature components.

6.2 Experiment and Results

We have used two databases for evaluation : OGI-Stories and OGI-Numbers (described in Sections 2.3 and 2.3.1). OGI-Stories database is used for training the band-specific posterior estimators (MLPs), and the Tandem MLP is trained on OGI-Numbers. Testing is done on continuous OGI-Digits.

The mean subtracted, variance normalized, and hamming windowed 101-sample temporal patterns of log critical-band energies are concatenated from the given number of critical-bands. The concatenated patterns are projected to 75 components using the joint Principal components (PCAs) and used as input to class-posterior estimators (MLPs). The class-posterior estimators are trained with 29 phones as the target categories in each frequency channel. Here, a frequency channel is a group of N adjacent critical-bands. The channels have an overlap of $N-1$ critical-bands where N is the number of critical-bands that are combined and used as input to the class-posterior estimators. The joint PCAs are computed a priori on an independent dataset (TIMIT) using $N*101$-samples from a $2-$D time-frequency tile of log critical-band energies. The estimated local phone posteriors are used as input to a Tandem MLP for the final phone posterior estimation. All MLPs have 300 hidden units and 29 phones as the target categories for the class-posterior estimation. The 29 phone posteriors at the output of the Tandem MLP are gaussianized, decorrelated, and used as the input features to the HMM based back-end recognizer (Section 2.3.1).

Table 6.1 shows WER in clean condition for different number of combined critical-bands. From the result it can be seen that on increasing frequency-context from 1 band to 3 bands recognition performance significantly (at 98 % confidence interval) increases. On increasing further the context it decreases. Figures 6.1, 6.2, 6.3, and 6.4 show ASR performance for different number of combined critical-bands in noisy conditions. The results show that word error rate reduce significantly (at 98 % confidence interval) when the frequency context increases from 1 critical-band to 3 critical-bands in all the $10-20$ dB SNR, noisy testing conditions. For these conditions, beyond three critical-bands of frequency context, the recognition error rate increases or remains the same. However, for low SNR ($0-5$ dB), in Pink and Babble noisy conditions, the WER significantly reduces on increasing frequency-context from 1 band to 3 bands but it increases on increasing context from 3 to 4 bands and then it again decreases on increasing further the context from 4 to 5 bands. We used a Matched Pairs Sentence-Segment Word Error (MAPSSWE) test to perform significance test on the results [23]. The effects of increasing frequency-context on frame-level phone accuracies on clean testing condition are shown in Table 6.3. For

most of the vowels, plosives, and fricatives, accuracy increases with increasing frequency-context from one critical-band to three critical-bands. For more than 3 critical-bands of frequency context, the phone accuracy reduces or remains the same.

Table 6.1: % Word Error Rate (Recognition Performance) on using joint PCAs on time-frequency plane of speech representation for feature estimation at clean condition

Frequency-context (number of critical-bands)	1-band (WER)	2-band (WER)	3-band (WER)	4-band (WER)	5-band (WER)
WER	6.6	5.3	5.0	5.9	5.8

6.3 Replacing Joint PCAs by DCT bases

The previous section, we noticed that the joint PCAs can be represented by a combination of DCT bases. In this experiment, for the 3-band case, two 101-sample vectors are computed. The first 101-sample is obtained by averaging 101-sample temporal patterns from three adjacent critical-bands, while the second is obtained by taking the difference between patterns in the first and third bands. These two 101-sample vectors are projected independently onto a 38-component DCT basis. This gives 76 projected components spanning $1 - 20$ Hz of modulation frequency. These components are used as input to the class-posterior estimators for estimating phone posterior probabilities in each frequency channel. Table 6.2 shows the recognition performance obtained using 3−band joint DCT

Table 6.2: % Word Error Rate on using 76 joint DCTs on the temporal patterns of three adjacent critical-bands, 75 joint 3-bands PCAs, 50 1-band DCT in clean, and in noisy condition

Frequency-context (number of critical-bands)	clean (WER)	babble (WER)	pink (WER)	white (WER)
3-band DCT	5.7	8.6	5.8	5.8
3-band PCAs	5.0	8.7	5.8	5.3
1-band DCT	6.0	10.3	7.5	7.1

components, 3−band joint PCAs components, and 1−band DCT components on clean as

well as 10 dB SNR noisy condition. With joint DCT components, we obtain similar or worse results than those obtained from 3−band joint PCAs, But the joint DCT results are still significantly (at 98 % confidence interval) better than those obtained from 1−band DCT components.

Table 6.3: % Frame-level phone error rates obtained using $1-5$ critical-bands of frequency context. The minimum error is shown in bold type.

phones	1-band	2-band	3-band	4-band	5-band
t	20.7	18.7	**17.5**	17.7	18.5
k	41.9	39.8	**38.2**	39.6	39.6
tcl	28.8	27.3	**26.9**	27.0	28.5
kcl	12.9	11.2	10.8	**10.7**	11.5
s	16.9	17.6	**16.0**	16.7	18.3
z	25.2	23.1	**21.2**	21.3	23.5
f	21.0	21.9	**21.4**	21.5	21.1
th	29.7	26.4	**26.1**	26.9	26.5
v	34.2	32.0	30.5	**29.8**	30.2
n	17.2	16.4	**15.6**	**15.6**	16.2
r	21.7	**17.6**	18.3	19.1	19.5
w	19.1	18.2	**17.4**	17.6	18.9
iy	21.4	20.7	**19.7**	20.0	21.0
ih	21.2	22.4	**21.6**	22.4	23.7
eh	28.5	25.1	**24.2**	24.2	24.0
ey	24.5	23.5	**20.7**	22.5	22.1
ay	12.0	9.3	**9.2**	9.8	9.7
ah	34.1	30.6	**28.8**	29.0	28.7
ow	23.2	18.6	18.5	**18.4**	17.8
uw	17.2	15.5	14.3	**14.2**	15.3
sil	13.9	**12.3**	12.5	13.0	12.7

6.4 Summary

In this work, we investigated the effect of broadening the frequency-context on ASR. The joint PCA components are computed and used as input for phone posterior estimation. From the structure of the joint PCAs it can be seen that joint PCA bases capture local spectral-slope information in the projected temporal patterns. We showed that broader than one critical-band frequency context is required for higher recognition performance.

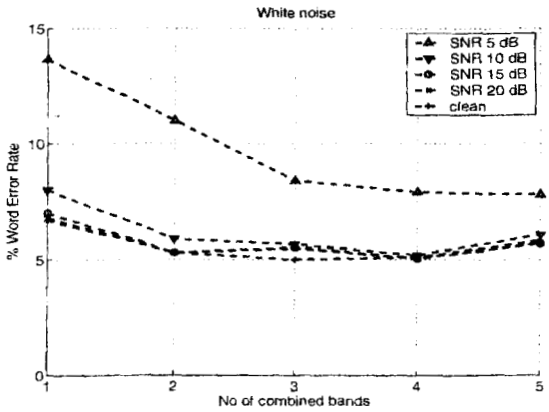

Figure 6.1: Effect of broadening frequency context on ASR using TRAPS features

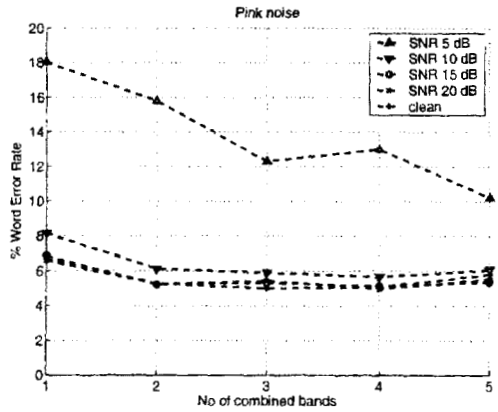

Figure 6.2: Effect of broadening frequency context on ASR using TRAPS features

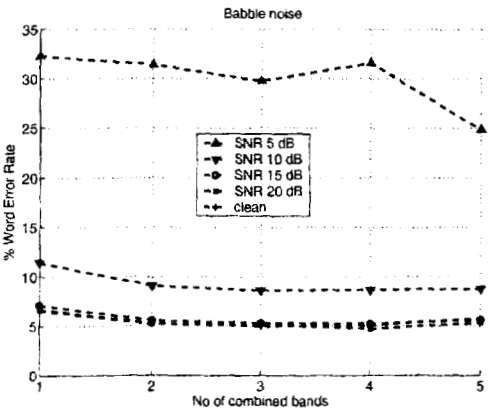

Figure 6.3: % Effect of broadening frequency context on ASR using TRAPS features

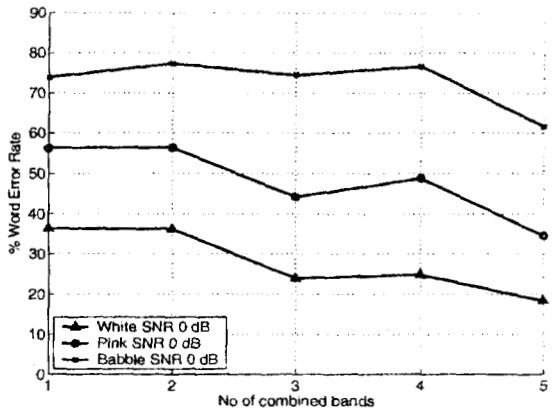

Figure 6.4: Effect of broadening frequency context on ASR using TRAPS features, At 0 dB SNR conditions

The robustness of the estimated features doesn't suffer on incorporating more than 1 critical-bands of frequency context during feature estimation in all the testing conditions. This indicates that local spectral-slope information is required for achieving higher phones accuracies. We showed that joint PCA bases can be approximated by a combination of several DCT bases. The trend of improved recognition performance on increasing frequency-context still holds on using joint DCT components. The 3-band DCT components outperform (significant at 98 % confidence interval) 1-band DCT components. In conclusion, we showed that modeling the interaction among adjacent critical-bands is required for obtaining higher recognition performance.

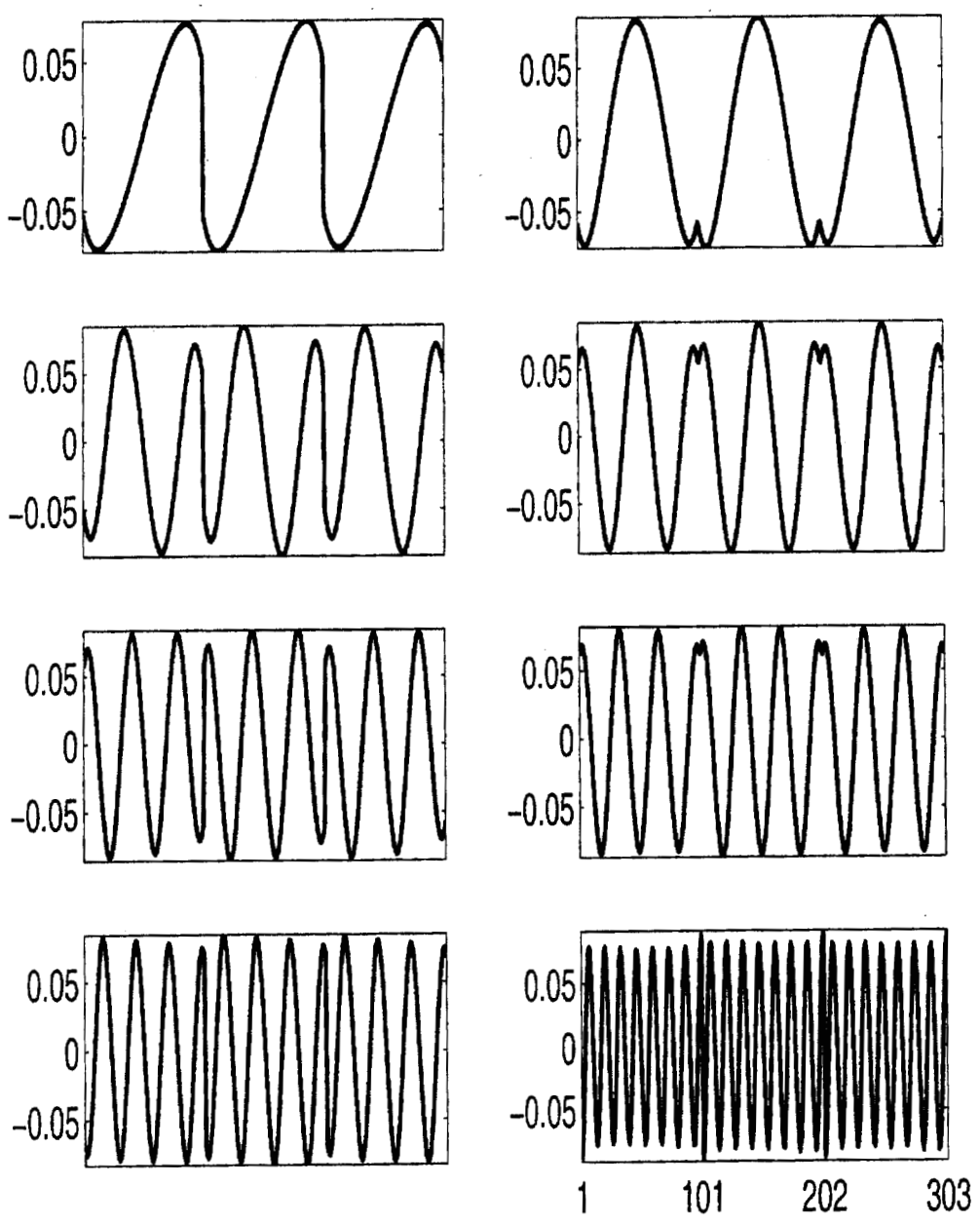

Figure 6.5: First 8 PCAs over 3 critical-bands and 1 s (101-samples) temporal context

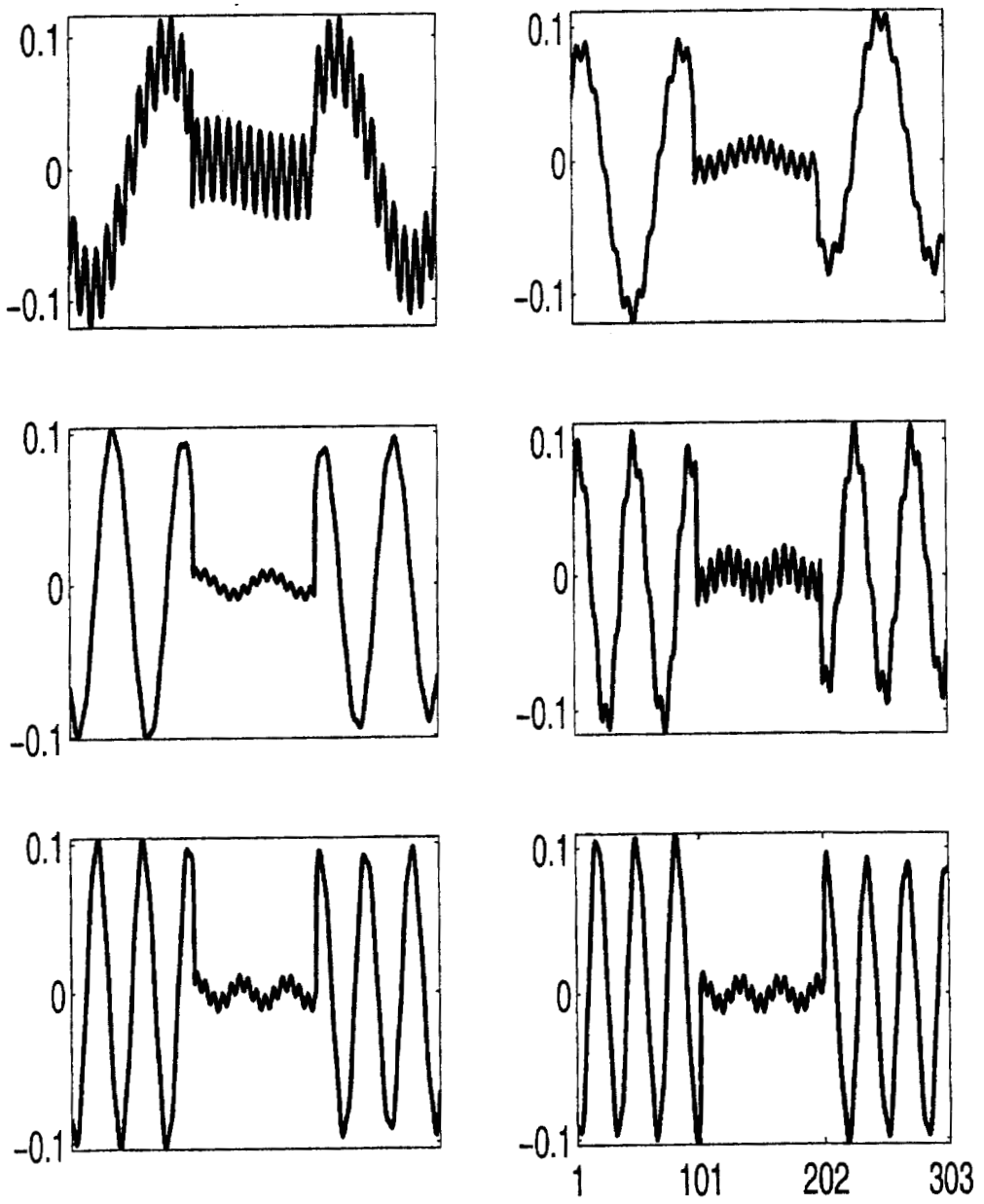

Figure 6.6: Some of the next 6 PCAs over 3 critical-bands and 1 s (101-samples) temporal context

Chapter 7

Summary

In this chapter, we summarize the thesis work in the section 7.1. We present some of the future directions in the section 7.2 and the final comments in the section 7.3.

7.1 Summary

The thesis was organized in five parts. Chapter 2 described various ways of parameterizing temporal patterns. The work in this chapter showed that linear components of temporal patterns can be used to parameterize patterns for obtaining better recognition performance in clean as well as in noisy environments. The temporal patterns closely follow the properties of a first-order Markov sequence, which results in their principal components (PCAs) being close to a DCT basis. Around 50 PCA (DCT) components cover $1 - 25$ Hz of modulation spectrum and 98 % of the total variability of 101-sample temporal patterns. Using DCT or DFT components of temporal patterns give significantly better recognition performance than using temporal patterns directly for recognition. DCT and DFT components performance very similarly in most of the testing conditions. However, LDA components of temporal patterns perform significantly worse than DFT or DCT components in low SNR, noisy environments. This work also showed the effect of low-pass filtering the modulation spectrum of temporal patterns on ASR. The recognition performance drastically goes down on cutting off modulation spectrum below 8 kHz of modulation frequency components. The recognition performance doesn't drop significantly as long as upto around 16 Hz of modulation spectrum of temporal patterns is retained for the successive feature estimation.

Chapter 3 investigated broad-phonetic category features. These features are estimated from temporal patterns independently in each frequency band. The features are evaluated for phone classification and continuous word recognition tasks. We showed that broad-phonetic features give significant gain in the recognition performance when they are combined with short-term features (PLP cepstra) in ASR. This indicates their complementary nature to short-term cepstral features. On using broad-phonetic features, the gain in recognition performance is the same as that carried by phone category features which are estimated from temporal patterns. By using broad-phonetic features, the number of parameters of the system is drastically reduced with improved generalization properties of these parameters.

Chapter 4 showed two practical applications of broad-phonetic features estimated from temporal patterns. The first one is robust Voice Activity Detection (VAD). The proposed VAD scheme is compared with a state-of-art GMM VAD technique. Speech recognition performance on the hypothesized speech segments was used as the performance metric for evaluating different VAD techniques. The proposed temporal patterns VAD system outperformed the GMM based VAD scheme in various noisy conditions. The second application is DSR in cellular telephony. We showed that broad-phonetic features can be estimated from the temporal patterns obtained from the reconstructed critical-band energies at the server-end without disturbing the terminal-end feature processing modules. On augmenting the broad-phonetic features with the short-term cepstral features we achieved significant gain in the recognition performance in various noisy conditions.

Chapter 5 proposed a new UTRAPS system. This system uses a single, universal, class-posterior estimator for estimating the speech-event category posteriors. The band-independent categories are obtained by using an agglomerative hierarchical clustering technique applied to the mean temporal patterns of the context-independent phones collected from all the critical-bands. A simple correlation measure was used as a distance measure to obtain the final clusters of temporal patterns. These patterns indicate distinct temporal activities manifested in the frequency localized temporal domain of the speech signal. We showed that UTRAPS features give consistent gains, similar to TRAPS broad-phonetic features, in recognition performance when they are combined with conventional

ASR features. These features are complementary to short-term cepstral features. They have better generalization properties than previously proposed broad-phonetic features, as data is shared not only across different phones but also across different frequency bands during system training. The UTRAPS system used just 5959 parameters as opposed to 86355 parameters used in the earlier TRAPS system.

Chapter 6 studies the effect of broadening the frequency-context on the estimated phone category features. We show that broader than one critical-band frequency context is required for higher recognition performance. The robustness of the estimated features doesn't suffer on incorporating more than 1 critical-band of frequency context during feature estimation in most of the testing conditions. We show that joint PCA bases can be approximated by a combination of several DCT bases. The trend of improved recognition performance on increasing frequency-context still holds on using joint DCT components. The 3-band DCT components outperform 1-band DCT components. In conclusion, we showed that modeling the interaction among adjacent critical-bands is required for obtaining higher recognition performance.

7.2 Future Directions

This work can be further extended by exploring band-independent speech-event categories obtained from 2-D dimensional time-frequency regions of the speech signal. A universal posterior estimator can be used to classify these categories and the estimated class-posteriors can be used as the robust features for ASR. The categories should target consistent and trainable the acoustic regions of the speech signal. The frequency-context can be around 3 critical-bands wide and the temporal-context can be around $200 - 500$ ms long to identify different speech-events of the speech recognition.

For merging the information from individual frequency-bands, we used the Tandem system. In this system, narrowband features (broad-phonetic or speech-events) are concatenated and mapped to the final class-posteriors using the Tandem MLP. The other methods for features combination need to be investigated. The reliability measure needs to be incorporated for selecting narrowband features on their reliability during combining

them.

7.3 Final Comments

In this work, we defined broad-phonetic categories and speech-events categories on the similarity in temporal characteristics of context-independent phones in different critical-bands. We assumed that phone like entities represents different sounds and thus can be used as the basis to find different events in time-frequency plane of the speech signal. The speech-event like categories were also obtained by clustering mean temporal patterns of the context-independent phones. Choosing the context-independent phones as the basis for defining speech-events categories may not be an optimal way for robust feature estimation. The more investigation is needed for defining these categories for feature estimation.

Appendix A

First-order Markov sequences

A random sequence $u(n)$ is called Markov-p or pth-order Markov if the conditional probability of $u(n)$ given the entire past is equal to the conditional probability of $u(n)$ given only $u(n-1), ..., u(n-p)$,

$$P(u(n)|u(n-1), u(n-2), ...) = P(u(n)|u(n-1), ..., u(n-p)), \forall n.$$

Another interpretation of a pth-order Markov sequence is that if the present, i.e. $u(j), n-p \leq j \leq n-1$, is known, then the past i.e., $u(j), j \leq n-p$, and the future, i.e. $u(j), j \geq n$, are independent. A Markov-p scalar sequence can also be expressed as a $(px1)$ Markov-1 vector sequence. The covariance function of a first-order stationary Markov sequence $u(n)$ is given as

$$r(n) = \rho^{|n|}, |\rho| \leq 1, \forall n$$

where $|\rho|$ is close to 1. The NXN discrete cosine transform (DCT) is very close to the Karhunen Loeve (KL) transform of a first-order stationary Markov vector sequence of length N whose covariance matrix is given by $r(n)$, when the correlation parameter ρ is close to 1. The reason is that R^{-1}, the inverse of the covariance matrix, is a symmetric tridiagonal matrix [8].

The covariance matrix of a first order N-point long Markov sequence has a following

structure.

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho^3 & \cdots & \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{pmatrix}$$

It is a Toeplitz matrix. The inverse matrix of $\mathbf{R}$, for the value of $\beta^2 = \frac{(1-\rho^2)}{(1+\rho^2)}$ and $\alpha = \frac{\rho}{(1+\rho^2)}$ satisfies the following relation.

$$\beta^2\mathbf{R}^{-1} = \begin{pmatrix} 1-\rho\alpha & -\alpha & \cdots & 0 \\ & & \mathbf{0} & \\ -\alpha & 1 & \cdots & \\ & & \ddots & \\ & \mathbf{0} & & \\ 0 & & -\alpha & 1-\rho\alpha \end{pmatrix}$$

For ρ 1, the $\mathbf{R}^{-1}$ matrix can be approximated by the following symmetric tridiagonal matrix Q_c

$$\mathbf{Q_c} = \begin{pmatrix} 1-\alpha & -\alpha & \cdots & 0 \\ & & \mathbf{0} & \\ -\alpha & 1 & \cdots & \\ & & \ddots & \\ & \mathbf{0} & & \\ 0 & & -\alpha & 1-\alpha \end{pmatrix}$$

The basis vectors of the cosine transform are the eigen-vectors of the symmetric tridiagonal matrix Q_c.

Appendix B

Mean Temporal patterns

The computation of 101-sample mean temporal patterns of different phones is as follows,

$$M_{phone,B} = \frac{1}{N_{phone}} \sum_{n=1}^{n=N_{phone}} \frac{(X^n_{phone,B} - m^n)}{\rho^n};$$

Where $M_{phone,B}$ is the mean temporal pattern computed over N_{phone} observations of a phone, *phone* in the critical-band B. Where $X^n_{phone,B}$ represents 101-sample temporal vector of log critical-band energies formed by concatenating 50 samples in the past and in the future of the current frame at time n. At every 10ms, the temporal pattern X is mean subtracted and variance normalized. The local mean m^n and the variance ρ^n of X^n are computed as follows,

$$m^n = \frac{1}{101} \sum_{k=1}^{k=101} X^n_{phone,B}(k)$$

$$\rho^n = \frac{1}{101} \sum_{k=1}^{k=101} (X^n_{phone,B}(k) - m^n)^2$$

We used 15 Mel critical-band energies for computing mean temporal patterns. The computation of 15 Mel critical-bands are described in detail in the Appendix H. We used TIMIT dataset for this computation.

The mean temporal patterns of some of the phones are shown in the following figures.

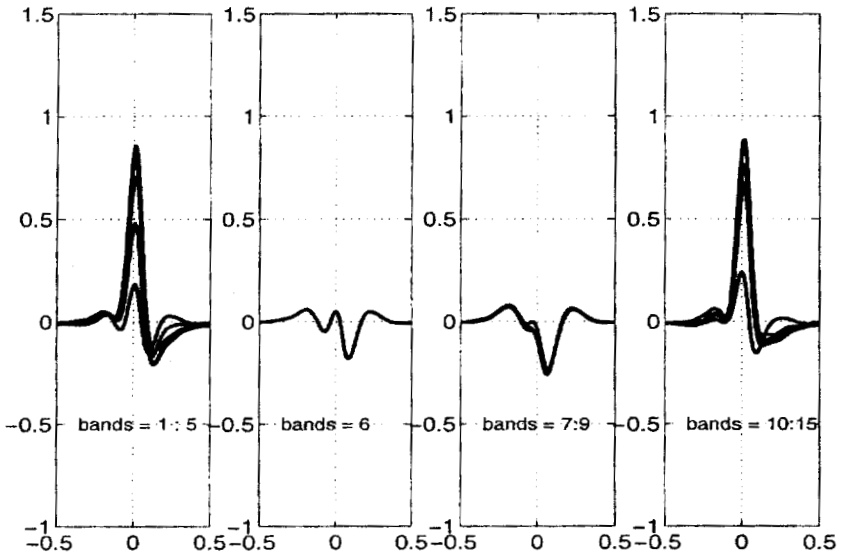

Figure B.1: Mean Temporal Patterns for a front vowel /iy/

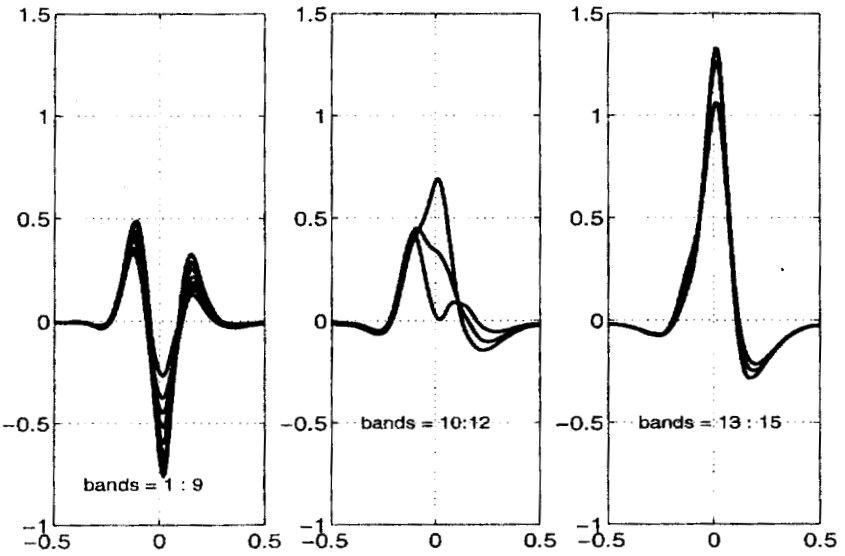

Figure B.2: Mean Temporal Patterns for a fricative /sh/

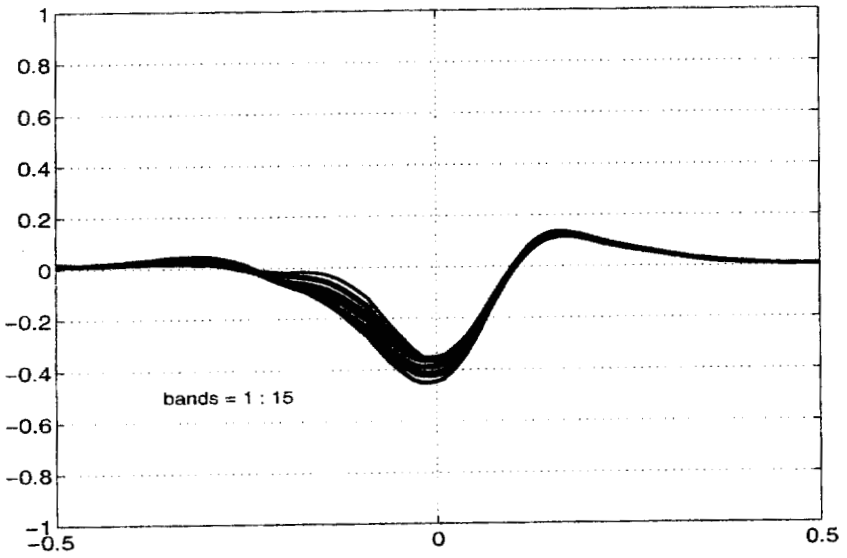

Figure B.3: Mean Temporal Patterns for silence /sil/

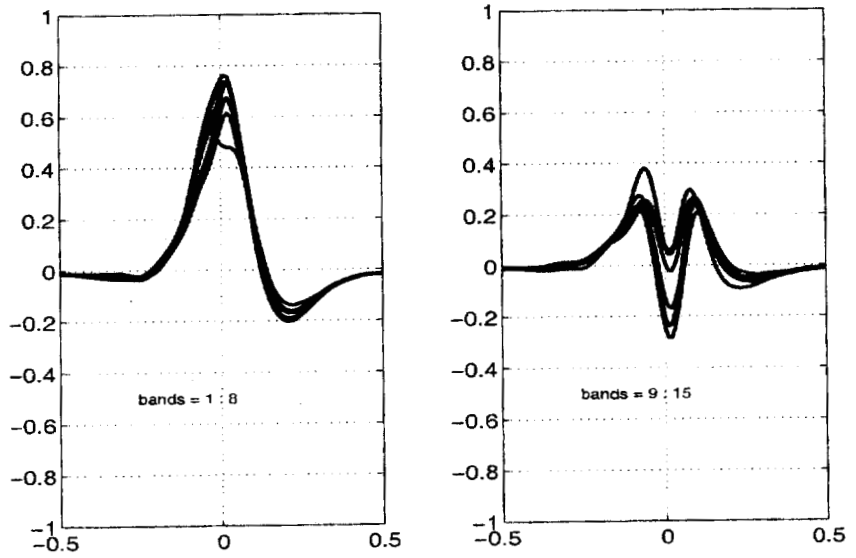

Figure B.4: Mean Temporal Patterns for a glide /l/

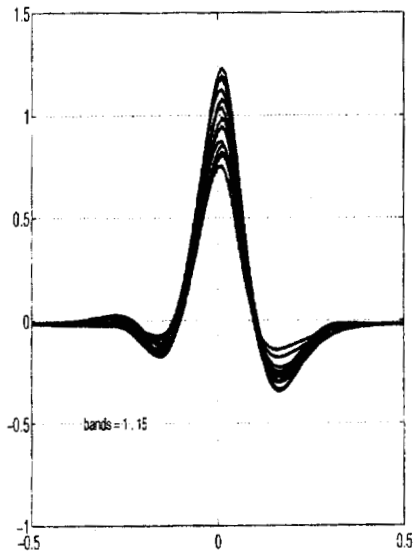

Figure B.5: Mean Temporal Patterns for a vowel /ae/

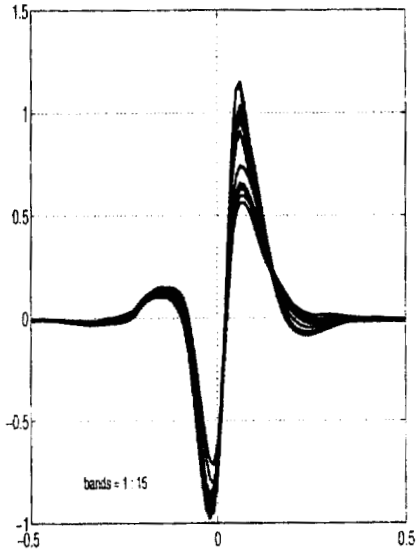

Figure B.6: Mean Temporal Patterns for a plosive /b/

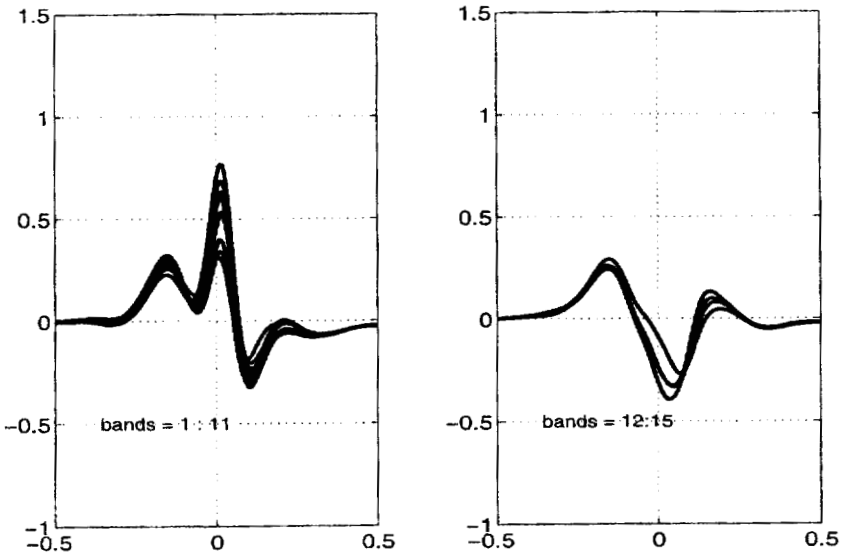

Figure B.7: Mean Temporal Patterns for a schwa /axr/

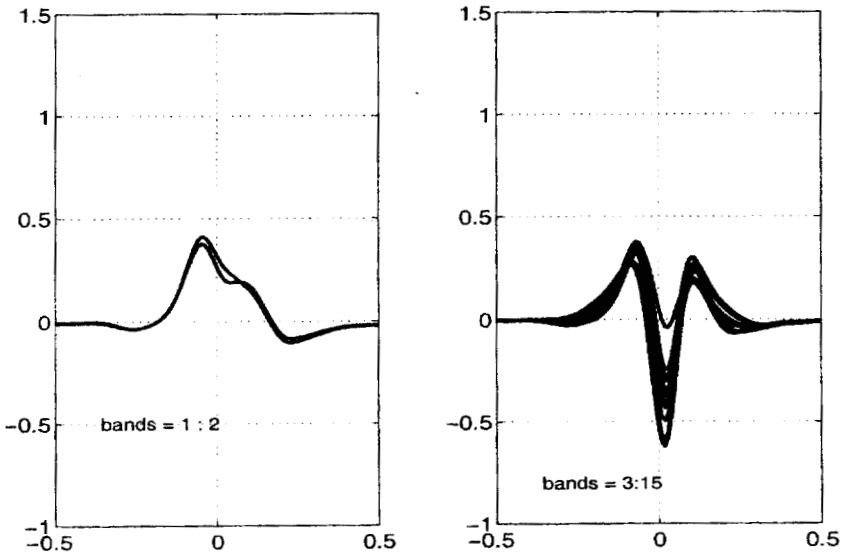

Figure B.8: Mean Temporal Patterns for a glide /m/

Appendix C

Phone Set

Table C.1: The set of 56 phones in TIMIT.

Phone	Example	Phone	Example	Phone	Example	Phone	Example
b	bee	d	day	g	gay	p	pea
t	tea	k	key	dx	muddy	bcl	BCL B iy
dcl	DCL D ey	gcl	GCL G ey	pcl	PCL P iy	tcl	TCL T iy
kcl	KCL K iy	jh	joke	ch	choke	s	sea
sh	she	z	zone	zh	azure	f	fin
th	thin	v	van	dh	then	m	mom
em	bottom	n	noon	nx	winner	ng	sing
en	button	l	lay	el	bottle	r	ray
w	way	y	yacht	hh	hay	hv	ahead
iy	beet	ih	bit	eh	bet	ey	bait
ae	bat	aa	bott	aw	bout	ay	bite
ah	but	ao	bought	oy	boy	ow	boat
uh	book	uw	boot	er	bird	axr	butter
ax	about	ix	debit	sil	pause	q	glottal stop

Table C.2: The set of 29 phones in OGI-Numbers.

Phone	Example	Phone	Example	Phone	Example	Phone	Example
d	day	t	tea	k	key	dcl	DCL D ey
tcl	TCL T iy	kcl	KCL K iy	s	sea	z	zone
f	fin	th	thin	v	van	m	mom
n	noon	l	lay	r	ray	w	way
iy	beet	ih	bit	eh	bet	ey	bait
ae	bat	ay	bite	ah	but	ao	bought
ow	boat	uw	boot	er	bird	ax	about
sil	pause						

Appendix D

Confusion matrices for phones

In a TRAPS based system (Section 1.3, the confusion matrices are computed at the output of the band-specific class posterior estimators (MLPs) as well as at the output of the Tandem MLP. For that, the MLPs are trained with 56 phones as the target classes. For every 10ms, a decision is made about the winning phone whose posterior probability was found greater than the other phone classes. Using the correct phone labels of each feature frame, confusion matrix was estimated. We used TIMIT dataset for this estimation.

The confusion matrices at the output of band-specific MLPs show more confusion among those phone classes which have similar temporal patterns of log critical-band energies. The Tandem MLP learns the specific patterns of the confusion matrices from individual critical-bands and try to combine that knowledge to estimate final phone class-posteriors. It plays the role of a merging net in a TRAPS based system.

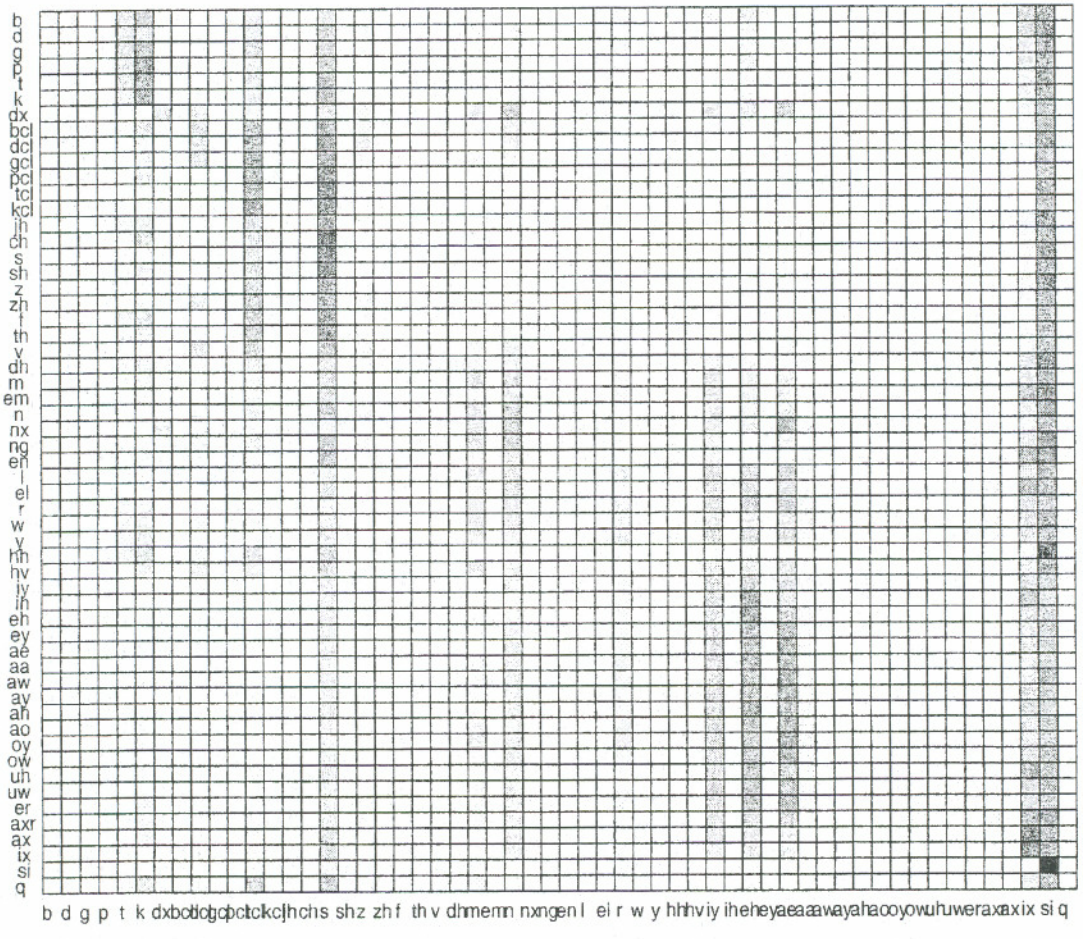

Figure D.1: For 56 TIMIT phones in the 1st critical-band, 115 – 265 Hz

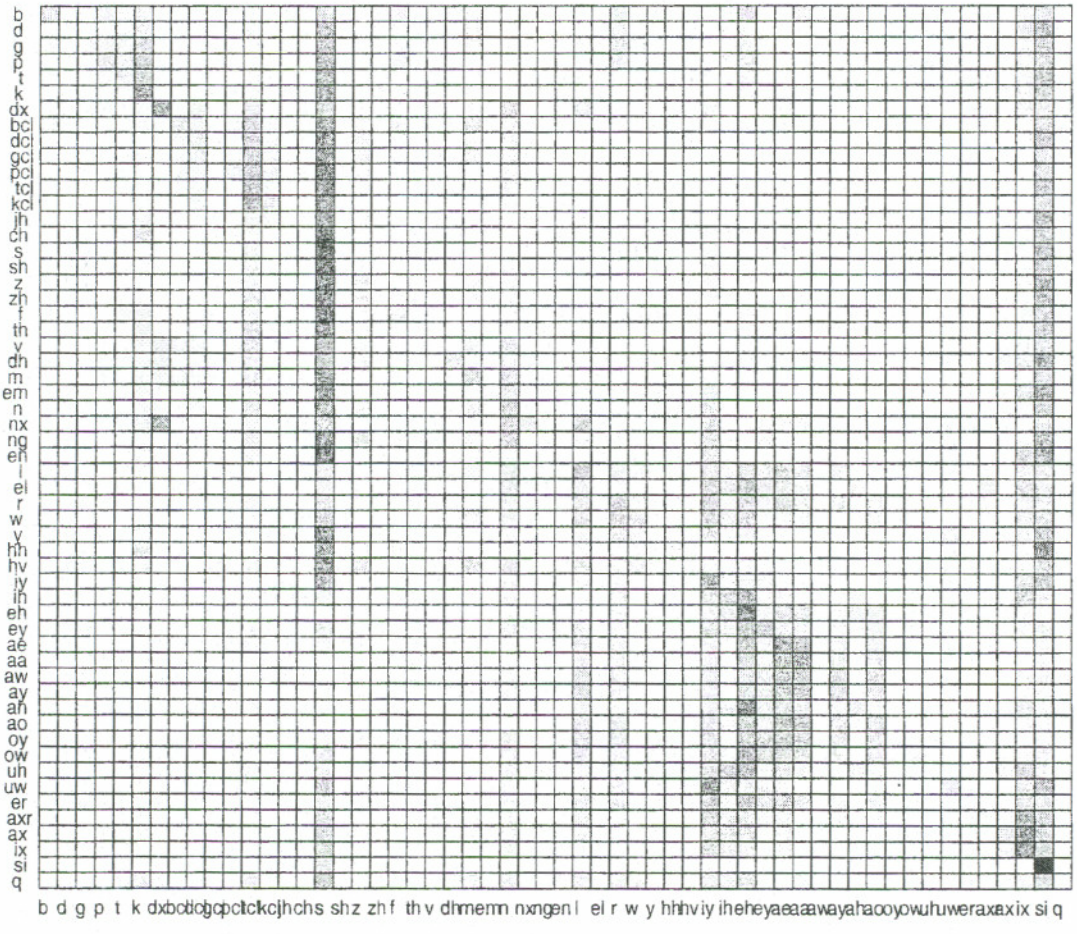

Figure D.2: For 56 TIMIT phones in the 5th critical-band, 4.538 – 629 Hz

Figure D.3: For 56 TIMIT phones in the 15th critical-band, 2962 – 3769 Hz

Figure D.4: For 56 TIMIT phones at the Tandem outputs

Appendix E

Clusters of speech events

The table shows which phone classes fall under which speech-event after applying clustering in a given critical-band. Each phone is quantified by one of the 9 speech-events in each critical-band. Phones change cluster affiliations based on distinct speech-activities present in each critical-band. Here N represent the critical-band index and $C1 - C9$ represent nine speech-event clusters.

N	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	si	b g p k dh t hh	em	l r w y hv ao oy	en z	ax	dx v ng m n	nx el iy ih eh ey ae aa aw ay ah oy ow uh uw er axr ix	q th f zh sh s ch jh kcl tcl pcl gcl dcl bcl d
2	si	b g p k dh hh	em	r w y oy	en z	ax	dx v ng m n	nx el iy ih eh ey ao ae aa aw ay ah oy ow uh uw er axr ix l	q th f zh sh s ch jh kcl tcl pcl gcl dcl bcl d t hv
3	si	b g p k dh t hh	em	r w y oy	en	ax ix	dx v ng nx n m	el iy ih eh ey ao ae aa aw ay ah oy ow uh uw er axr l	q th f z zh sh s ch jh kcl tcl pcl gcl dcl bcl d hv

N	C1	C2	C3	C4	C5	C6	C7	C8	C9
4	si	b g p k dh t hh	em	r w y oy	en	ax ix	dx n ng nx	el iy ih eh ey ao ae aa aw ay ah oy ow uh uw er axr l	q th f v z zh sh s m ch jh kcl tcl pcl gcl dcl bcl d hv
5	si	b g p k dh y q t hh	em	r w oy	en z	ax ix	dx n ng nx l	el iy ih eh ey ao ae aa aw ay ah oy ow uh uw er axr	th f v zh sh s m ch jh kcl tcl pcl gcl dcl bcl d hv
6	si	b g p k dh y q t hh	em	r w oy	en ng z	ax ix iy	dx n nx	el ih eh ey ao ae aa aw ay ah oy ow uh uw er axr l	th f v zh sh s m ch jh kcl tcl pcl gcl dcl bcl d hv
7	si	b g p k dh q t hh	em	w oy	en ng z iy	ax ix	dx n nx	el r ih eh ey ao ae aa aw ay ah oy ow uh uw er axr l	th f v zh sh s m ch jh kcl tcl pcl gcl dcl bcl d hv y
8	si	b g p k dh w q t hh	em	 oy ao	en ng z iy	ax ix	dx v nx m l	el r ih eh ey ae aa aw ay ah oy ow uh uw er axr	q th f n zh sh s ch jh kcl tcl pcl gcl dcl bcl d hv y
9	si em	b g p k dh w y t hh		oy ao oy	en el z iy	ax ix	dx nx l	er axr r ih eh ey ae aa aw ay ah ow uh uw	q th f n v zh sh s m ng ch jh kcl tcl pcl gcl dcl bcl d hv

N	C1	C2	C3	C4	C5	C6	C7	C8	C9
10	si em	b g p k dh t hh		hv r oy y ao oy	en el z	ax ix	dx nx sh n sh	iy ay ah ih eh ey ae aa aw ow uh uw er axr	q th f v s m ng bcl d ch jh kcl l tcl pcl gcl dcl bcl d w
11	si em	b g p k dh q hh		hv r t ch y ao oy ay	en el z s	ax ix	dx sh nx ng aw zh n axr	iy ih eh ey ae aa ah ow uh uw er	th f v m jh kcl l w tcl d pcl gcl dcl bcl
12	si em	b g p k dh q hh		hv r t ch oy y ay jh	en el z s axr	ax ix	dx nx sh ng n zh er	iy ih eh ey ae aa ao ah aw w uh uw	th f v l w m d kcl tcl pcl gcl dcl bcl
13	si em	b g p k dh r hh		hv t ch l oy y jh d	en el z axr er	ax ix	dx zh nx ng s uw	iy ih eh ey sh ae aa ao ah aw ay uh w	th f v n w m q gcl kcl tcl pcl dcl bcl
14	si em	b g p k dh r hh		hv t y oy jh d	en el ax er	ax ix	dx nx ng l uw	iy s z ih eh ey ch ae aa ah aw ay w uh zh sh ao	th f v n m q w gcl kcl tcl pcl dcl bcl
15	si em	b g p k dh hh		oy t y l d	el axr er	ax ix	dx nx uw	iy s z th ih eh ey ch ae aa ah aw ay w uh zh jh f sh ao	en hv v n m q r kcl w tcl pcl gcl dcl bcl ng

Table E.1: Clusters in individual critical-bands

Appendix F

Code for generating Speech events

The following code is used to generate nine speech events.

```
clear Data Data
Data=[];
%% Number of critical-bands
Bands=15
%% Number of phones
Classes=56;

        % Reading the mean temporal patterns of 56 phones
        % (TIMIT dataset) from individual critical-bands (15)
        for i =  1:15,
                str=strcat('Mean_TRAPS_phones',int2str(i-1));
                MeanTRAPS{i}=load(str);
                Data=[MeanTRAPS{i}; Data];
        end

        % Randomizing the data
        NN=randperm(length(Data));
        Data=Data(NN,:);

        % Calculating the correlation matrix
        Cv=(Data*Data');
        dd=sqrt(diag(Cv));
```

```
        pp=dd*dd';
        Corr=Cv./pp;
        DD=diag(diag(Corr));
        DD=-1*DD;
        Corr=Corr+DD;

        TMP=Data;
for iter = 1 : 831,
% position of the maxima in the correlation matrix
        [my x]=max(Corr);
        [mx y]=max(my);
        searchVec=1:size(TMP,1);
        X=x(y);
        Y=y;

% Merging the two nodes of maximum correlation
          tmp=mean([TMP(X,:);TMP(Y,:)]);
          searchVec=setdiff(searchVec,X);
          searchVec=setdiff(searchVec,Y);
          TMP=TMP(searchVec,:);
          TMP=[TMP;tmp];

% Updating the correlation matrix after merging the nodes
         Cv=(TMP*TMP');
         dd=sqrt(diag(Cv));
         pp=dd*dd';
         Corr=Cv./pp;
         DD=diag(diag(Corr));
         DD=-1*DD;
         Corr=Corr+DD;
end
```

Appendix G

Linear Transforms of temporal patterns

G.1 Linear transform from Principal Component Analysis (PCA)

For PCA transform computation [20] of the temporal patterns, first total covariance matrix of the normalized temporal patterns is estimated using TIMIT or OGI-Stories dataset.

$$X^n = [X_{n,0}, \cdots, X_{n,101}];$$

$$Y^n = \frac{(X^n - m^n)}{\rho^n};$$

Where X^n represents 101-sample temporal vector of log critical-band energies formed by concatenating 50 samples in the past and in the future of the current frame at time n. At every 10ms, the temporal pattern X^n is mean subtracted and variance normalized. Y^n represents a mean and normalized temporal pattern. For normalization, local mean m^n and the variance ρ^n of X^n are computed as follows,

$$m^n = \frac{1}{101} \sum_{k=1}^{k=101} X^n(k)$$

$$\rho^n = \frac{1}{101} \sum_{k=1}^{k=101} (X^n(k) - m^n)^2$$

Next total covariance matrix is estimated as follows,

$$Cov = \frac{1}{N} \sum_{n=0}^{n=N} [(Y^n - m_y)^T * (Y^n - m_y)];$$

Where m_y is given by,

$$m_y = \frac{1}{N} \sum_{n=0}^{n=N} Y^n;$$

Where Cov represents $101x101$ covariance matrix of the vector Y, m_y represents mean of normalized temporal vector Y, and N is the total number of frames in the dataset.

After computing total covariance matrix, singular decomposition is applied to compute eigenvector matrix V and eigenvalue matrix S.

$$Cov = U * S * V';$$

The columns of the eigenvector matrix V represent principal component bases of the temporal vector X. We used first 50 bases for estimating class-posterior probabilities. These 50 bases cover 98 % of the total covariance of dataspace Y.

G.2 Linear transform from Linear Discriminant Analysis (LDA)

For LDA transform computation [55] of the temporal patterns, first within-class covariance and across-class covariance matrices are computated as follows,

$$m_i = \frac{1}{N_i} \sum_{n \in C_i} Y^n,$$

$$m = \sum_{n \in C_i} \frac{N_i}{N} m_i,$$

$$S_B = \sum_{i=0}^{i=C} \frac{N_i}{N} \sum_{n \in C_i} (m_i - m)^T * (m_i - m),$$

$$S_W = \sum_{i=0}^{i=C} \frac{N_i}{N} \sum_{n \in C_i} (Y^n - m_i)^T * (Y^n - m_i),$$

Where m_i is the mean of the ith class, N_i is the number of sample-points for ith class, m is the global mean of the normalized temporal pattern, Y^n, S_B and S_W are the between-class and within-class covariances respectively for the C number of classes.

$$D = S_W^{-1} * S_B;$$

$$D = U * S * V';$$

On singular value decomposition of matrix D the eigenvector matrix V and eigenvalue matrix S are obtained. The eigenvectors are referred to as linear discriminants. These discriminants are sorted based on their eigenvalues. The leading discriminants are used as the projection bases.

We used 29 phones as the target classes for deriving 29 Linear Discriminant basis vectors (LDAs). The LDA bases are derived on the clean, OGI-Stories corpus, a dataset described in section 2.3.1. The first 15 LDA components of the temporal patterns are used as input features to sub-word class (phones) posterior estimators.

We also derive the first 50 LDAs basis vectors from 554 Gaussian mixtures (from 3-state monophone HMM models), used as the target categories, for our second experiment.

Appendix H

Computation of PLP critical-bands and Mel critical-bands

We used 1-Bark spaced PLP critical-bands for the continuous numbers task on OGI-Numbers. The cut-off frequencies are listed in the following table.

Table H.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

For Aurora experiments, we used 23 MEL filterbank energies for feature extraction. The frequency range (64 – 4000Hz) was used for computing MEL-warped spectrum for feature extraction. The cut-off frequencies of 23 critical-bands are listed in the following table. The 23 MEL filterbanks are generated as follows,

Table H.2: Cut-off frequencies of the used Mel critical band filters for 8 kHz sampling frequency.

Critical band number	Lower cut-off frequency (Hz)	Upper cut-off frequency (Hz)
1	64	186
2	155	248
3	217	341
4	279	403
5	372	496
6	434	589
7	527	682
8	620	806
9	713	930
10	837	1054
11	961	1178
12	1085	1333
13	1209	1488
14	1364	1674
15	1519	1860
16	1705	2047
17	1891	2264
18	2078	2512
19	2295	2760
20	2543	3008
21	2791	3318
22	3039	3628
23	3349	3969

$$Mel\{x\} = 2595 * \log\left\{1 + \frac{x}{700}\right\}$$

$$f_{c_i} = Mel^{-1}\left\{Mel(64) + i\frac{Mel(4000) - Mel(64)}{23+1}\right\}, i = 1, \cdots, 23$$

$$cbin_i = round\left\{\frac{f_{c_i}}{f_s} * 256\right\}$$

Where $floor(.)$ stands for rounding downwards the nearest integer. The output of the Mel filter is the weighted sum of the 256-point FFT power spectrum values (bin_i) in each band. f_s (8000Hz) is the sampling frequency. Triangular half-overlapped windowing is used as follows.

$$fbank_i = \sum_{k=cbin_{i-1}}^{k=cbin_i} \frac{k - cbin_{i-1}}{cbin_i - cbin_{i-1}} S_k + \sum_{k=cbin_i+1}^{k=cbin_{i+1}} \frac{cbin_{i+1} - k}{cbin_{i+1} - cbin_i} S_k$$

Where $k = 1, \cdots, 23$, $cbin_0$, and $cbin_{24}$ denote the FFT bin indices corresponding to the starting frequency of 64Hz and final frequency of 4000Hz respectively. S_k is the short-term power spectral density of the speech signal.

$$cbin_0 = round(64 * 256/8000)$$

$$cbin_{24} = 256/2 = 128$$

The 23 critical-band energies are projected to 15 DCT components and 15 cepstral based features (referred to as 15 robust MFCC) are computed at the handset side and transmitted to the network. At the server these 15 features are projected back to critical-band energies using a 15 point $IDCT$ transform and that gives 15 reconstructed critical-band energies. These reconstructed energies are used in the TRAPS and UTRAPS based system for class-posteriors estimation for the DSR experiments.

Bibliography

[1] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan and S. Sivadas. QUALCOMM-ICSI-OGI Features for ASR. In *Proc. of ICSLP* (Colorado, USA, 2002), vol. 1, pp. 21–24.

[2] Acero, A., and Stern, R. Environmental robustness in automatic speech recognition. In *Proc. of ICASSP'90* (1990), vol. 1, IEEE, pp. 849–852.

[3] Allen, J. Short term spectral analysis, synthesis and modification by discrete fourier transform. *IEEE - ASSP 25* (1977), 235–238.

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

[5] Allen, J. Articulation and intelligibility. *DIMACS workshop on complexity and inference 1* (2003), 200–212.

[6] Allen, J., and Rabiner, L. Unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE 65* (Nov. 1977), 1558–1564.

[7] Allen, J. B. Applications of the short-time fourier transform to speech processing, spectral analysis. *Proc. ASSP Workshop on Spectral Estimation* (1981), 6.3.1–6.3.5.

[8] Anil K. Jain. *Fundamental of Digital Image Processing.* PRANTICE HALL, New Jersey, 1989.

[9] B. H. Juang. Speech recognition in adverse environments. *Computer speech and language 1* (1991), 275–294.

[10] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya. Robust speech recognition in noisy environments: The 2001 ibm spine system. In *Proc. of ICASSP* (Orlando, USA, 2002), vol. 1, pp. 53–56.

[11] B. Kingsbury, P. Jain, and A. Adami. A hybrid hmm/traps system for robust voice activity detection. In *Proc. of ICSLP* (Denver, USA, 2002), pp. 1073–1076.

[12] BOURLARD, H., AND DUPONT, S. A new ASR approach based on independent processing and re-combination of partial frequency bands. In *Proc. of ICSLP* (Philadephia,PA, 1996), vol. 1, pp. 426–429.

[13] D. ELLIS, R. SINGH AND S. SIVADAS. Tandem acoustic modeling in large-vocabulary recognition. In *Proc. of ICASSP* (Salt Lake City, Utah, 2001), vol. 2, pp. 12–18.

[14] D.L. THOMSON AND R. CHENGALVARAYAN. Use of periodicity and jitter as speech recognition features. In *Proc. of ICASSP* (Seattle, USA, 1998), vol. 1, pp. 21–23.

[15] E. ZWICKER AND E. TERHARDT. Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *The Journal of The Acoustical Society of America 68* (1980), 1523–1525.

[16] EMILY BUSS, JOSEPH W. HALL. The role of auditory filters in comodulation masking release (cmr). *Journal of the Acoustical Society of America 103* (1998), 3561–3566.

[17] ETSI STANDARD DOC. Speech prcessing, transmission, and quality aspects ; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm. *ETSI ES 202 050 v0.1.0 (2002-04)* (2002).

[18] F. JELINEK. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, 1997.

[19] FLETCHER, H. *Speech and Hearing in Communication.* Krieger, New York, 1953.

[20] FUKUNAGA, K. *Introduction to Statistical Pattern Recognition.* Academic Press, Indiana, 1990.

[21] FURUI, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE - ASSP 34* (1986), 52–59.

[22] FURUI, S. Recent advances in robust speech recognition. *Tutorial and Research Workshop on Robust Speech Recognition for unknown communication channels* (Apr. 1997).

[23] GILLICK, L., AND COX, S. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. of ICASSP* (1989), vol. 1, IEEE, pp. 532–535.

[24] H. BOURLARD AND N. MORGAN. *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publishers, Boston, 1994.

[25] H. HERMANSKY. Perceptual linear predictive (PLP) analysis for speech. *The Journal of The Acoustical Society of America 87* (Apr. 1990), 1738–1752.

[26] H. HERMANSKY AND S. SHARMA. Traps classifiers of temporal patterns. In *Proc. of ICSLP'98* (Sydney, Australia, 1998), vol. 2, pp. 615–618.

[27] H. HERMANSKY, D. ELLIS, AND S. SHARMA. Tandem connectionist feature extraction stream extraction for conventional hmm systems. In *Proc. of ICASSP* (Istanbul, 2000), pp. 1635–1638.

[28] H. HERMANSKY, N. MORGAN, A. BAYYA, AND P. KOHN. Compensation for the effect of the communication channel in auditory-like analysis of speech RASTA–PLP. In *Proc. of EUROSPEECH '91* (Genova, Switzerland, 1991), ESCA, pp. 1367–1370.

[29] H. YANG, S. VAN VUUREN AND H. HERMANSKY. Relevancy of Time-Frequency Features for Phonetic Classification Measured by Mutual Information. In *Proc. of ICASSP* (Phoenix, USA, 1999), pp. 225–228.

[30] HALL J. W., HAGGARA, M. P., AND FERNANDEZ, M. A. Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America 76* (1984), 50–56.

[31] HIRSCH, H.-G., PEARCE, D. The aurora experimentral framework for the performance evaluations of speech recognition systems under noisy conditions. *ISCA, ITRW, ASR 2000* (2000).

[32] J. R. BELLEGARDA, P. V. DE SOUZA, A. J. NADAS, D. NAHAMOO, M. A. PICHENY, AND L. BAHL. Robust speaker adaptation using a piecewise linear acoustic mapping. In *Proc. of ICASSP* (, USA, 1992), pp. 445–448.

[33] J. W. HALL, III, JOHN H. GROSE. Comodulation masking release and auditory grouping. *Journal of the Acoustical Society of America 88* (1990), 119–125.

[34] J.R. LEHMAN, W. JESTEAD, S. B. Forward masking as a function of frequency, masker level and signal delay. *The Journal of The Acoustical Society of America 71* (Apr. 1982), 950–962.

[35] K. WANG, S. SHAMMA. Representation of acoustic signals in the primary auditory cortex. *IEEE Trans. Audio and Speech Processing 3* (1995), 382–395.

[36] KIRCHHOFF, K. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *ICSLP* (Sydney, Australia, 1998), pp. 891–894.

[37] L. DENG, D. SUN. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of Acoustic Society of America* (1994), 2702–2719.

[38] L. R. BAHL, S. BALAKRISHNAN-AIYER, J. BELLEGARDA, M. FRANZ, P. GOPALAKRISHNAN, D. NAHAMOO, M. NOVAK, M. PADMANABHAN, M. PICHENY AND S. ROUKOS. Performance of the IBM large vocabulary continuous speech recognition system on the ARPA wall street journal task. In *Proc. of ICASSP* (Detroit, USA, 1995), pp. 41–44.

[39] LAWRENCE K. SAUL. MAZIN G. RAHIM, AND J. B. ALLEN. A statistical model for robust integration of narrowband cues in speech. *Computer Speech and Language 15* (2001), 175–194.

[40] LAWRENCE RABINER, BIING-HWANG JUANG, BILNG-HWANG JUANG. *Fundamentals of speech recognition.* Prentice Hall Signal Processing Series, 1998.

[41] LIU, S. Landmark detection of distinctive feature-based speech recognition. In *JASA* (1996), pp. 3417–3430.

[42] L.R. RABINER. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77* (Feb. 1989), 257–285.

[43] M. J. F. GALES AND S. J. YOUNG. An improved approach to the hidden markov model decomposition of speech and noise. In *Proc. of ICASSP* (, USA, 1992), pp. 181–184.

[44] MILLER G., NICELY P. An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am. 27* (1955), 338–352.

[45] N. KANEDERA, T. ARAI, H. H., AND PAVEL, M. On the importance of various modulation frequencies for speech recognition. In *Proc. of EUROSPEECH'97* (Rhodes, Greece, 1997), ESCA, pp. 1079 - 1082.

[46] N. KANEDERA, H. HERMANSKY, T. A. Desired characteristics of modulation spectrum for robust automatic speech recognition. In *Proc. of ICASSP'98* (Seattle WA, USA, 1998), vol. 2, pp. 613–616.

[47] NIYOGI, P. R. Distinctive feature detection using support vector machines. In *Proc. of ICASSP* (Phonix, USA, 1999), pp. 425–428.

[48] NOKIA. Availability of finnish speechdat-car database for etsi stq wi008 front-end standardisation. *STQ Aurora DSR working group, au21799* (1999).

[49] O. GHITZA. Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. *Journal of Phonetics 16* (1988), 109–124.

[50] PAVEL, M. *Homogeneity in complete and partial masking.* PhD thesis, New York University, New York, 1980.

[51] PAVEL, M., AND HERMANSKY, H. Temporal masking in automatic speech recognition. *The Journal of The Acoustical Society of America 95* (May 1994), 2876–2880.

[52] R. COLE, M. NOEL, T. L., AND DURHAM, T. New telephone speech corpora at cslu. In *The Fourth European Conference on Speech Communication and Technology* (1995), pp. 821–824.

[53] R. COLE, M. NOEL AND T. LANDER. Telephone speech corpus development at CSLU. In *Proc. of ICSLP* (Yokohama, Japan, 1994), pp. 1815–1818.

[54] R. DRULLMAN, J. F., AND PLOMP, R. Effect of temporal envelope smearing on speech reception. *The Journal of The Acoustical Society of America 95* (1994), 1053–1064.

[55] R. DUDA, P. HART, D. S. *Pattern Classification.* Wiley Interscienc, 2001.

[56] R. SINGH, MICHAEL L. SELTZER, B. R., AND STERN, R. M. Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination. In *Proc. of ICASSP* (Salt Lake City, Utah, 2001), pp. 7–11.

[57] ROBERT V. SHANNON, FAN-GANG ZENG, VIVEK KAMATH, JOHN WYGONSKI, MICHAEL EKELID. Speech recognition with primarily temporal cues. *Science 270* (1995), 303–304.

[58] RONALD E. CROCHIERE, LAWRENCE R. RABINER. *Multirate Digital Signal Processing*, fifth ed. PRANTICE HALL Signal Processing series, New Jersey, 1983.

[59] S. KAJAREKAR AND H. HERMANSKY. Analysis of information in speech and its application in speech recognition. In *Proc. of TSD* (Brno, Czech Republic, 2000), pp. 283–288.

[60] S. SHARMA, P. V., AND HERMANSKY, H. Combining information from multiple classifiers for speaker verification. In *Speaker Recogniton and its Commercial and Forestic Applications* (France, 1998), pp. 20–24.

[61] S. SHARMA, D. ELLIS, S. KAJAREKAR, P. JAIN AND H. HERMANSKY. Feature extraction using non-linear transformation for robust speech recognition on the AURORA data-base. In *Proc. of ICASSP* (Istanbul, Turkey, 2000), pp. 1117–1120.

[62] S. VAN VUUREN AND H. HERMANSKY. Data-driven design of rasta-like filters. In *Proc. of EUROSPEECH* (Rhodes, Greece, 1997), pp. 409–412.

[63] S.B. DAVIS AND P. MERMELSTEIN. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing 28* (Aug. 1980), 357–366.

[64] SHARMA, S. *Multi-Stream Approach To Robust Speech Recognition.* PhD thesis, OGI, Portland, USA, Apr. 1999.

[65] SHIHAB A. SHAMMA, TAISHIH CHI, YUJIE GAO, MATTHEW C GUYTON. Spectro-temporal modulation transfer functions and speech intelligibility. *Journal of the Acoustical Society of America 106* (1999), 2719–2732.

[66] S.S. STEVENS. On the psychophysical law. *Psychol. Rev. 64* (1957), 153–181.

[67] TIBREWALA, S., AND HERMANSKY, H. Multi-band and adaptation approaches to robust speech recognition. In *Proc. of EUROSPEECH'97* (Rhodes, Greece, 1997), vol. 5, ESCA, pp. 2619–2622.

[68] ULF KNOBLICH, ALCATEL. Description and baseline results for the subset of the speechdat-car italian database used for etsi stq aurora wi008 advanced dsr front-end evaluation. *STQ Aurora DSR working group, au23700* (2000).

[69] ULF KNOBLICH, ALCATEL. Spanish sdc-aurora database for etsi stq aurora wi008 advanced dsr front-end evaluation : Description and baseline results. *STQ Aurora DSR working group, auxxx00* (2000).

[70] X. D. HUANG AND K. F. LEE. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on speech and Audio processing 1* (1993), 150–157.

Biographical Note

Pratibha was born in Jabalpur, Madhyapradesh, India, on the 22nd of October 1973. She completed Bachelor of Electrical Engineering degree in 1995 from Goverment College of Engineering, Jabalpur. In August 1995, she got admission in Indian Institute of Technology (IIT), Kanpur, one of the prestegious school of India, for pursuing her master-level studies. She specialized in Digital Signal Processing (DSP). After finishing masters in Feb 1997, she moved to Indian Institute of Science (IISc), Bangalore for pursuing research in the area of image restoration techniques. In 1998 she joined Oregon Graduate Institute (OGI) as a Ph.D student. She worked in the area of automatic speech recognition (ASR) in Anthropic Signal Processing Group, OGI under the guidance of Prof. Hynek Hermansky. She worked on several projects which include recognition of speech in noisy environments (SPINE), Distributed Speech Recognition (AURORA). The following year in 2001, she was a summer intern at the T. J. Watson, IBM lab where she worked on techniques for improving robustness of speech recognition in noisy environment and techniques for robust Voice Activity Detection (VAD) under the supervision of Dr. Mukund Padmanabhan and Dr. Brian Kingsbury. During her Ph.D, she presented her work in many internation conferences. She authored and co-authored several international conference papers. The fields of her interest include machine learning, human-machine interaction systems, and information retrieval from multimedia signals.

Selected Publications

1. Hynek Hermansky and Pratibha Jain, Band-independent speech-event categories for TRAP based ASR, *EUROSPEECH*, Geneva, Switzerland, Sep. 2003.

2. Pratibha Jain and Hynek Hermansky, Beyond a single critical-band in TRAP based ASR, *EUROSPEECH* , Geneva, Switzerland, Sep. 2003.

3. Pratibha Jain, Hynek Hermansky, and Brian Kingsbury, Distributed speech recognition using noise-robust MFCC AND TRAPS-estimated manner features, *ICSLP*, Vol 1, pages 473-476, Denver, USA, Sep. 2002.

4. Brian Kingsbury, Pratibha Jain, and Hynek Hermansky, A Hybrid HMM/TRAPS model for robust voice activity detection, *ICSLP*, Vol 2, pages 1073-1076, Denver, USA, Sep. 2002.

5. Pratibha Jain and Hynek Hermansky, Down-sampling speech represent -ation in ASR, *European Conference on Speech Communication and Technology (EUROSPEECH)*, Vol 1, pages 73-76, Budapest, Hungary, 1999.