

Representation Learning for Spoken Term Detection

Raghavendra Reddy Pappagari

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Technology



भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Electrical Engineering

June 2015

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

P. Raghavendra Reddy

(Signature)

Raghavendra Reddy Pappagari

(Raghavendra Reddy Pappagari)

EE12M1023

(Roll No.)


Approval Sheet

This Thesis entitled Representation Learning for Spoken Term Detection by Raghavendra Reddy Pappagari is approved for the degree of Master of Technology from IIT Hyderabad

(Dr. Balasubramaniam Jayaram) Examiner
Department of Mathematics
IITH



(Dr. Ketan P. Detroja) Examiner
Department of Electrical Engineering
IITH



(Dr. Sumohana Channappayya) Examiner
Department of Electrical Engineering
IIT



(Dr. K. Sri Rama Murthy) Adviser
Department of Electrical Engineering
IITH

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. K Sri Rama Murty for his valuable guidance and constructive criticism. It would have been impossible for me to complete this thesis without his insightful thinking into working of algorithms and results. It is his persistence for perfection that has provided me concrete base for research.

I am also thankful to my lab mates Kallola Rout, Shekhar Nayak and Karthika Vijayan for their valuable help in my research. My sincere thanks to Mettu Srinivas and Mohammed Rafi for their comments on my thoughts and thought provoking discussions. My special thanks to Mohammed Rafi and B. Naresh Reddy for their excellent efforts in creation of Telugu database.

Dedication

I would like to dedicate this thesis to my parents and all of my friends who were with me in difficult times.

Abstract

Spoken Term Detection (STD) is the task of searching a given spoken query word in large speech database. Applications of STD include speech data indexing, voice dialling, telephone monitoring and data mining. Performance of STD depends mainly on representation of speech signal and matching of represented signal.

This work investigates methods for robust representation of speech signal, which is invariant to speaker-variability, in the context of STD task. Here the representation is in the form of templates, a sequence of feature vectors. Typical representation in speech community Mel-Frequency Cepstral Coefficients (MFCC) carry both speech-specific and speaker-specific information, so the need for better representation. Searching is done by matching sequence of feature vectors of query and reference utterances by using Subsequence Dynamic Time Warping (DTW). The performance of the proposed representation is evaluated on Telugu broadcast news data.

In the absence of labelled data i.e., in unsupervised setting, we propose to capture joint density of acoustic space spanned by MFCCs using Gaussian Mixture Models (GMM) and Gaussian-Bernoulli Restricted Boltzmann Machines (GBRBM). Posterior features extracted from trained models are used to search the query word. It is noticed that 8% and 12% improvement in STD performance compared to MFCC by using GMM and GBRBM posterior features respectively. As transcribed data is not required, this approach is optimal solution to low-resource languages. But due to its intermediate performance, this method can not be immediate solution to high resource languages.

In the presence of labelled data i.e., in supervised setting, Hidden Markov Model (HMM)- Multi Layer Perceptron (MLP) hybrid model is employed to extract phonetic posterior features. Speech is segmented into phoneme labels by using HMM. Obtained phoneme segments are supplied to MLP training. It is observed that the information in phonetic posteriors contributed to 25% improvement is seen in STD performance compared to MFCCs. Effect of speaking mode of query words is also investigated in this approach. By careful selection of path weights in accumulated matrix calculation in DTW, great improvement in performance of STD system for query words recorded in isolated manner is observed.

It can be seen that speech information in MFCC is enhanced to extract better features for STD task. Minimal pair ABX (MP-ABX) tasks are used to analyse different features and develop insights into the nature of information in them. In this work, MFCC and features derived from analytic signal Frequency domain linear prediction (FDLP), instantaneous frequency (IF) coefficients are evaluated using MP-ABX tasks. FDLP features are derived from analytic magnitude and IF coefficients are derived from analytic phase of speech signals. The performance of the features derived from analytic representation are compared with performance of the MFCC. It is noticed that the magnitude based features- FDLP and MFCC delivered promising PaC, PaT and CaT scores in MP-ABX tasks, demonstrating their phoneme discrimination abilities. Combining FDLP features with MFCC had proven beneficial in phoneme discrimination tasks. The IF features performed well in TaP mode of MP-ABX tasks, emphasizing the existence of speaker specific information in them. The IF significantly outperformed FDLP, MFCC and their combination in speaker discrimination task.

Index Terms: Spoken Term Detection, GMM, GBRBM, HMM-MLP, Representation learning, Joint density estimation, Subsequence matching, DTW, MP-ABX.

Contents

Declaration	ii
Approval Sheet	iii
Acknowledgements	iv
Abstract	vi
Nomenclature	vii
1 Introduction	1
1.1 Introduction	1
2 Literature Survey	3
2.1 Review of Approaches for STD	3
2.1.1 Feature Representation for STD	3
2.1.2 Template Matching of Feature Representations	7
3 Unsupervised Approaches	8
3.1 Posterior extraction using GMM	8
3.1.1 STD using Gaussian posteriors	10
3.2 Posterior Extraction using GBRBM	11
3.2.1 STD using GBRBM Posteriors	15
3.3 Conclusion	16
4 Supervised Approaches	17
4.1 Posterior Extraction using Hybrid HMM-MLP	17
4.1.1 Speech segmentation using HMM	17
4.1.2 Extraction of Phonetic Posteriors using MLP	18
4.1.3 STD using Phonetic Posteriors	19
4.1.4 Effect of speaking mode	22
4.2 Conclusion	22
5 Analysis of Features from Analytic Representation of Speech using MP-ABX Measures	23
5.1 MP-ABX discrimination tasks	24
5.2 Analytic features of speech	24
5.2.1 Frequency domain linear prediction	25
5.2.2 Instantaneous frequency	25
5.3 Evaluation of analytic features	25
5.4 Discussions	28
5.5 Conclusions	29
6 Conclusions and Future Work	30
References	31

Chapter 1

Introduction

1.1 Introduction

It is difficult to manage and monitor increasing volumes of data on the internet. Resources can be efficiently managed, if only the required information can be retrieved from this data. In the case of speech data, we need to search and locate spoken query words in large volumes of continuous speech. This task is termed as Spoken Term Detection (STD). Some of the applications of STD include speech data indexing [1], data mining [2], voice dialling and telephone monitoring.

Audio search can be broadly categorized into keyword spotting (KWS) and spoken-term detection (STD), depending on the domain (text or audio) of the query supplied. In KWS task, the query word is supplied as text [3], [4]. Since the query word (text) and reference data (audio) are in two different domains, one of them needs to be transformed into the other domain to perform the search. Hence, the knowledge of the language and pronunciation dictionary is required to implement the KWS system. In the case of STD task, the query word is also supplied in the audio format [5], [6]. Since the reference and query are in the same domain, the STD task does not require the knowledge of pronunciation dictionary. However, it suffers from channel mismatch, speaker variability and differences in speaking mode/rate as opposed to KWS task. One of the main issues in STD is to devise a robust and speaker-invariant representation for the speech signal, so that the query and reference utterances can be matched in the new representation domain. In this work, we compare and contrast the supervised and unsupervised approaches to learn robust speech-specific representation for the STD task.

In this work, we have used posterior representation of speech for developing the STD system. Posterior features are extracted in both unsupervised and supervised approaches. In the absence of labelled data, the posterior features are obtained using two unsupervised methods, namely, GMM and GBRBM. In supervised approach, HMM-MLP hybrid modelling is employed to extract phonetic posteriorgrams using labelled data. Experiments have been conducted on each of these techniques to choose optimal set of parameters. Subsequence DTW is applied on the posterior features to perform query search. Average Precision, $P@N$, is used as an evaluation metric, which indicates the number of correctly spotted instances of the query word, out of total occurrences N of that word.

Speech and speaker information contained different representations MFCC, FDLP and IFCC are analysed by using MP-ABX tasks: PaC, PaT, CaT and TaP. DTW is used to find whether A or B is similar to X. Average error rate is computed as the ratio of total errors to total number of triplets.

Rest of the thesis is organized as follows: Chapter 2 highlights the importance of feature representation for the STD task, and presents a survey of state-of-the-art approaches for arriving at a robust representation. Unsupervised learning of representations from speech signals is discussed in Chapter 3 using Gaussian mixture models (GMM) and Gaussian Bernoulli Restricted Boltzmann Machine (GBRBM). Building STD system using phonetic posteriors, a representation learned using supervised approach, is discussed in Chapter 4. Evaluation

of features derived from analytic representation using minimal-pair ABX task is presented in Chapter 5. Chapter 6 summarizes the important contributions of this study, and directions to the important issues to be addressed in future.

Chapter 2

Literature Survey

2.1 Review of Approaches for STD

The task of STD can be accomplished in two important stages: (i) extracting a robust representation from the speech signal, and (ii) matching the representations obtained from reference and query utterances for detecting the possible locations.

2.1.1 Feature Representation for STD

Performance of STD system depends critically on the representation of the speech signal. The acoustic waveforms, obtained by measuring sound pressure levels, of a word uttered by two different speakers look completely different, and yet carry the same linguistic information. For example, the waveforms of the word “səmaik^hjə” uttered by two different speakers is shown in Fig. 2.1(a). Though these two waveforms carry the same linguistic information, it is impossible to match them because of the natural variability (associated with fundamental frequency, rate of speech, context, etc.). Even though the similarity between them is better visible in the spectrogram representations, shown in Fig. 2.1(b), still it is difficult to match them using a machine.

The variability associated with the spectrograms, in Fig. 2.1(b), can be reduced by considering the average energies over a bands of frequencies. The mel-spaced filter bank energies, shown in Fig. 2.1(c), match better than the waveforms and their spectrograms. The cepstral coefficients derived from the mel-spaced filter bank energies, popularly called as mel-frequency cepstral coefficients (MFCCs), are the most widely used features in speech recognition systems [7]. Although they are well suited for the statistical pattern matching, like in HMMs for speech recognition, they may not be the best representation for template matching in the STD task. To illustrate this point, consider the task of searching the spoken query word “səmaik^hjə” in the reference utterance “ra:stɾa:ni səmaik^hjəŋga unca:ləntu si:mɑ:ndrəne:tələ.” The distance matrix ($\mathbf{D} = [d_{ij}]$) between the MFCC features of the reference and query utterances, for matched speaker condition, is shown in Fig. 4.2(a). The element d_{ij} in the distance matrix denotes the Euclidean distance between the MFCC features of i^{th} reference frame and j^{th} query frame. The distance matrices shown in Fig. 4.2(a) is a 3-dimensional representation, where x-axis denotes the sequence of reference frames, y-axis denotes the sequence of query frames, and intensity denotes the inverse of the pair-wise distances between reference and test frames. If the query word occurs in the reference utterance, then it is supposed to get reflected as an approximate diagonal path in the distance matrix. When the reference and query utterances are from the same speaker, there is a clear diagonal path in their distance matrix shown in Fig. 4.2(a). On the other hand, when the reference and query utterances are from different speakers, such a diagonal path can not be clearly spotted as in Fig. 4.2(b). This behaviour could be attributed to the speaker-specific nature of MFCC features, i.e., they do contain speaker-specific information. Notice that the MFCC features are used for speaker recognition also [8]. In the case of statistical pattern matching, the speaker-specific nature of MFCC features is normalized by pooling data from several different speakers. For STD also, we need to derive a robust speech-specific feature, from the speaker-independent

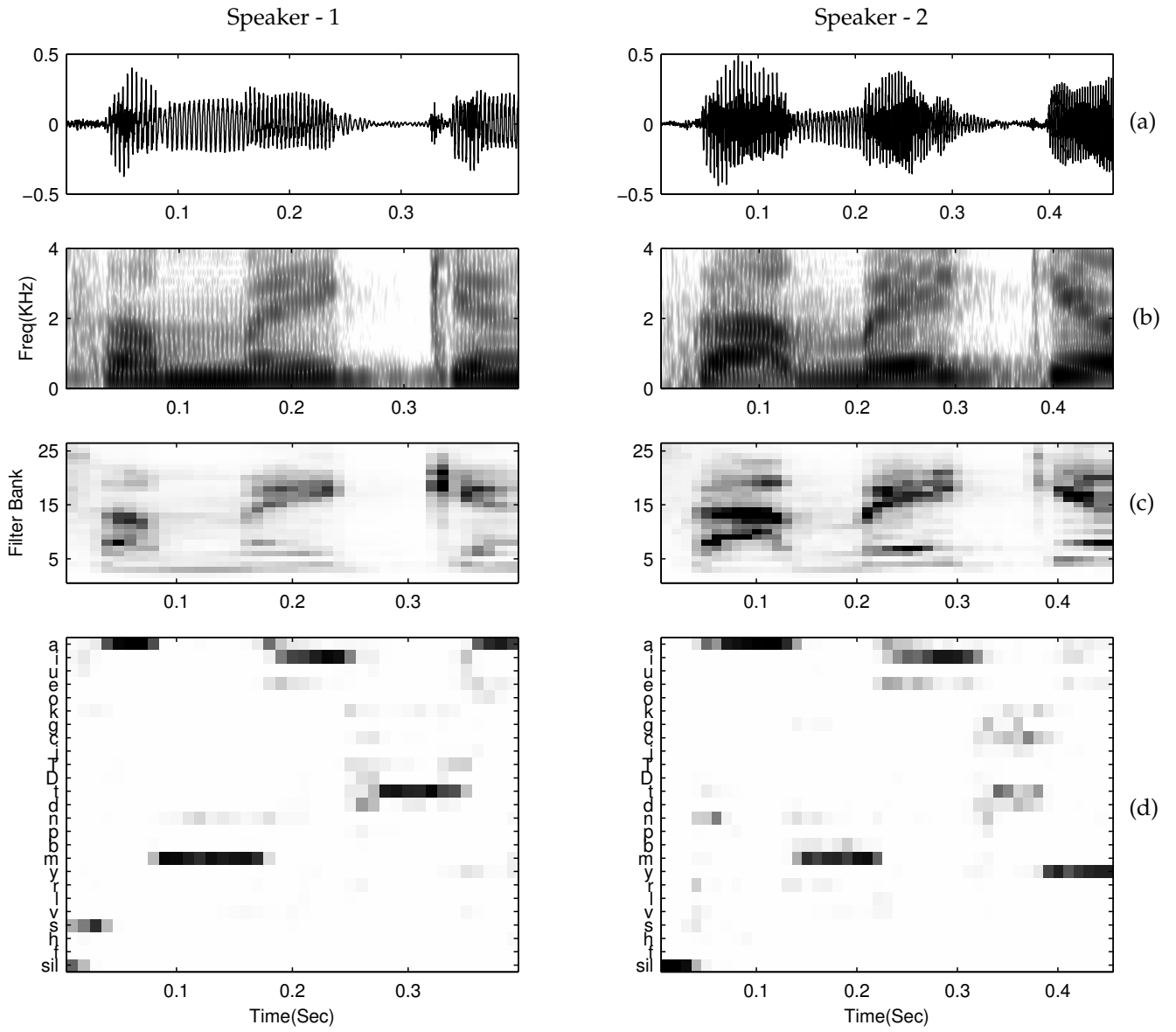


Figure 2.1: Illustration of suitability of phonetic posteriors for template matching of the word “səmaik^hjə” spoken by two different speakers. (a) Wave form (b) Spectrogram (c) Mel filter bank energies (d) MLP posteriors

Table 2.1: Summary of representation learning approaches for STD

	Discrete	Continuous
Supervised	Sequence of phoneme-like labels Eg: LVCSR [9], Sub-word modelling [10] Pros: Good performance, faster matching Cons: Requires transcribed data, language knowledge and pronunciation dictionary	Sequence of posterior probability vectors of phonemes Eg: HMM [11], MLP [12], Hybrid HMM-MLP [5] Pros: Good performance, phoneme decoding is not required Cons: Requires transcribed data, slower matching
Unsupervised	Sequence of cluster indices obtained from acoustic similarity Eg: Vector Quantization (VQ) [13] Pros: Does not require transcribed data, simple to implement and faster matching Cons: Poor performance	Sequence of posterior vectors obtained from a trained statistical/neural network models Eg: GMM [14], ASM [15], GBRBM [16] Pros: Does not require transcribed data, optimal solution for low resource languages Cons: Intermediate performance, slower matching

component of MFCCs, for template matching.

A good feature representation for STD should be speech-specific, and at the same time, it should be robust to speaker and channel variability. Several attempts have been made to learn suitable representation from the data (MFCCs of any other feature) using statistical models and neural network models. Depending on the learning paradigm (supervised or unsupervised) and nature of the resultant representation (discrete or continuous), the representations for the STD are broadly classified into four categories. Table 2.1 presents the summary of different approaches to learn speech-specific representations for the STD task.

Learning discrete representations using supervised approaches

In this class of approaches, discrete representation is obtained using supervised approaches where labelled information is utilized. In [17], [9], [18] a well trained large Vocabulary Continuous Speech Recognizer (LVCSR) is used to convert the speech into text. Generally, it performs Viterbi search on word lattice to find the most probable sequence of words based on likelihoods of the trained acoustic models and the language model. It was observed that N-best hypothesis Viterbi search performs better than 1-best hypothesis, particularly when Word Error Rate (WER) of the system is high [19]. Variants of word lattices, namely, Word Confusion Network (WCN) [20], [21] and Position Specific Posterior Lattice (PSPL) [22] were proposed to achieve good performance for IN-Vocabulary (INV) words. For spotting the locations of the query word in the reference utterance, text based searching methods were employed on the machine generated text transcripts [23]. This method is suitable mainly for high resource languages, as it requires large amount of transcribed speech data for training the language specific LVCSR system. Since these are word based recognizers, despite the high recognition accuracies for INV words, it suffers from Out-of-Vocabulary (OOV) problem. That is, it can not handle words which are not in the predefined dictionary. In real-time implementation of query based search, encountering OOV words, like nouns, is very common. Hence, LVCSR based method can not be a practical solution even for high resource languages.

In order to address this issue, subword LVCSR based recognizers [10], [24], [17] were proposed, where recognition is done based on subwords instead of words. Since any word can be represented with basic subword (like phonemes) set, it is possible to recognize any word using subword recognizers. But the performance of subword-based recognizers deteriorates for INV in comparison to LVCSR based recognizers. The evidences from both phonetic and word lattices are combined to improve the performance [19], [25]. Generally, in subword based techniques, words are represented using phonetic lattices. In the matching stage, query search is accomplished by matching the phonetic lattices of the words. Despite of its good performance and faster

matching, it is not practical solution for low resource languages which do not have enough labelled data.

Discrete representation using unsupervised approaches

This class of approaches do not use labelled information to obtain discrete representation. Vector Quantization (VQ) can be employed to represent speech as a sequence of discrete symbols, in an unsupervised manner [13]. In this method clustering is performed on feature vectors extracted from speech signal, and the set of mean vectors is stored as a codebook for representing speech. Feature vectors, derived from speech signal, can be represented as a sequence of codebook indices depending on their proximity to the cluster centres. In the matching stage, the sequence of codebook indices, obtained from the reference and query utterances, can be matched using approximate substring matching techniques [26]. Although this method does not require transcribed data and can be implemented with less complexity, its performance is significantly lower than the supervised approaches.

Continuous representation using supervised approaches

Conversion of speech signal into discrete sequence of phonemes, like in LVCSR, is less accurate because of the high degree of confusion among certain phonemes, especially among stop consonants. For example, the phoneme /k/ gets confused with /g/, and phoneme /d/ often gets confused with /t/ and /D/ and so on. This can lead poor transcription of reference and query utterances, resulting in lower STD performance. However, unlike speech recognition, STD task does not require the conversion of speech signal into sequence of phonemes. It was shown that phonetic posteriors, i.e., probability of the phoneme given the observed feature vector, are better suited for the STD task [27]. The posterior vector \mathbf{y} for an observed feature vector \mathbf{x} is defined as

$$\mathbf{y} = [P(c_1/\mathbf{x}) P(c_2/\mathbf{x}) P(c_3/\mathbf{x}) \cdots P(c_M/\mathbf{x})]^T \quad (2.1)$$

where $P(c_i/\mathbf{x})$ denotes probability of frame \mathbf{x} belonging to the i^{th} phoneme class c_i , and M denotes number of classes. The sequence of posterior vectors, commonly referred to as posteriorgram, forms a template representation of speech segment.

Phonetic posteriorgrams are extracted using well trained phoneme recognizers, built using hidden Markov models (HMMs) or Multilayer perceptrons (MLP) [28], [12]. It was shown that the phonetic posteriorgrams extracted using deep Boltzmann machines (DBM) are useful when limited amount of transcribed data is available [29]. In the next stage, dynamic time warping (DTW) is used to match the posteriorgrams extracted from the reference and query utterances.

Representing under resourced languages using well built multiple phone recognizers, each trained with one high resource language, was also explored by many researchers, and they were briefly summarized in [5], [6]. The features extracted from each recognizer for a given query, which can be from low resource language, were used to represent that query. However, all these approaches require labelled data in at least one language, which may not be feasible always.

Continuous representation using unsupervised approaches

Even though the supervised approaches are successful, they cannot be applied on a new language or under-resourced language, where manual transcriptions are not readily available. In such cases, unsupervised methods are preferred for posterior extraction [14], [15]. Most of the unsupervised posterior extraction methods rely on estimating the joint probability density of the feature vectors of the speech signal. Gaussian mixture model (GMM) has been used for estimating the density function of the speech data and, there by, posterior extraction [14]. One drawback with this method is that GMM cannot model the temporal sequence in which the feature vectors have evolved. In order to address this issue, Acoustic Segment Models (ASM), which take temporal structure of speech into account, were proposed [15]. In this approach, each pseudo phoneme class is modelled as a HMM, where the class labels were obtained from pre-trained GMM. The classes represent

pseudo phoneme like units, as they were obtained based on their acoustic similarity rather than their linguistic identity. It was shown that ASM outperforms well trained phoneme recognizer in language mismatched environment, and also GMM modelling. In the matching stage, a variant of DTW was used on GMM/ASM posteriors for searching the query word in the reference utterance. It was observed that the performance improves significantly by applying speaker normalization techniques, like Constrained Maximum Likelihood Linear Regression (CMLLR) and Vocal Tract Length Normalization (VTLN).

2.1.2 Template Matching of Feature Representations

Statistical behaviour of vocal tract system makes it impossible to generate two exactly similar speech segments in natural speech. So, no two speech utterances have equal durations of phonemes, and they are distributed non-linearly. In all the two stage approaches, matching stage plays crucial role in bringing out the excerpts of queries from continuous speech. Searching time and memory requirement are two main constraints in matching. Different matching methods are followed based on the representation of speech. In discrete representation, query words are represented as sequence of symbols or lattices. In [30], three methods for matching lattices were presented: Direct index matching, edit distance alignment and full forward probability. In the case of continuous representation, speech is represented as sequence of frames called as template. Matching templates was attempted in [31] using DTW. Degree of similarity of two given utterances can be approximated through cost of optimal path of DTW. Limitation of DTW is that durations of two utterances used for matching should be comparable, otherwise the computed similarity score denotes the closeness of long utterance with small utterance, like spoken word which are not comparable. Segmental DTW [32], subsequence DTW [33], unbounded DTW [34] and information retrieval DTW (IR-DTW)[35] are few of the variants of DTW, which are tuned to search queries in continuous speech. Improvements to IR-DTW using hierarchical K-means clustering was proposed in [36].

In [32], segmental DTW was proposed for unsupervised pattern discovery. In this technique, starting point in one utterance was fixed, and DTW was applied with a constraint on the degree of warping path. Starting point was slid through the chosen utterance. In the next step, the obtained paths at each point were refined using length-constrained minimum average (LCMA) algorithm [37] to get minimum distortion segments. In [14], modified segmental DTW, where LCMA is not applied on the paths, was used to accomplish STD goal. As this method requires DTW at periodic intervals of time in either of the two utterances, it requires high computation and memory resources, making it practically impossible on large archives.

In [33], subsequence DTW was proposed for STD task where calculation of accumulated matrix is different from conventional DTW. In the conventional DTW, dissimilarity values along the reference utterance were accumulated forcing the backtracked path to reach first frame. Using the fact that query can start from any point in the reference utterance, accumulation of values along first row of the reference utterance was not done, which makes the back-traced path to end at any point in reference utterance. Path was backtracked from minimum accumulated distortion.

In [35], IR-DTW was proposed for the STD task, in which memory requirement was reduced by avoiding calculation of full distance matrix. Starting indices in both query and reference utterances were chosen, based on exhaustive similarity computation. By using non-linear subsequence matching algorithm, ending indices were found, and the best matching segments were filtered. This algorithm can be scaled up for large amounts of data with reduced memory requirements. When query term contains multiple words, or if query term is not present exactly in the reference utterance, but parts of query term are jumbled in the reference utterance (like in QUESST task in MediaEval 2015), then this method is useful. This method was further improved using K-means clustering algorithm instead of exhaustive distance computation [36]. By imposing constraints on similarity score on matching paths, searching can be made faster as in [38], [39], [40].

Chapter 3

Unsupervised Approaches

Main objective of modelling techniques is to build a system which can discriminate different classes of inputs. In unsupervised approaches, this goal is accomplished without using labels. In the case of low resource languages, where less or no labelled data is available, unsupervised approaches can be a promising solution. In this study, generative models, namely GMM and GBRBM, are studied for unsupervised posterior feature extraction.

3.1 Posterior extraction using GMM

Mixture models capture the underlying statistical properties of data. In particular, GMM models the probability distribution of the data as a linear weighted combination of Gaussian densities. That is, given a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, the probability of data X drawn from GMM is

$$p(X) = \sum_{i=1}^M w_i \mathcal{N}(X/\mu_i, \Sigma_i) \quad (3.1)$$

where $\mathcal{N}(\cdot)$ is Gaussian distribution, M is number of mixtures, w_i is the weight of the i^{th} Gaussian component, μ_i is its mean vector and Σ_i is its covariance matrix. The parameters of the GMM $\theta_i = \{w_i, \mu_i, \Sigma_i\}$ for $i = 1, 2, \dots, M$, can be estimated using Expectation Maximization (EM) algorithm [41].

Fig. 3.1 illustrates the joint density capturing capabilities of GMM, using 2-dimensional data uniformly disturbed along a circular ring. The red ellipses, superimposed on the data (blue) points, correspond to the locations and shapes of the estimated Gaussian mixtures. In the case of 4-mixture GMM, with diagonal covariance matrices, the density was poorly estimated at odd multiples of 45° , as shown in Fig. 3.1(a). As the number of mixtures increases, the density is better captured as shown in Fig. 3.1(b). Since the diagonal matrices cannot capture correlations between dimensions, the curvature of the circular ring is not captured well. In the case of diagonal covariance matrices, the ellipses are aligned with the xy-axes as shown in Fig. 3.1(a) and Fig. 3.1(b). The density estimation can be improved using full covariance matrices, as shown in Fig. 3.1(c). However, this improvement comes at the expense of increased number of parameters and computation. We need to estimate $M(2D + 1)$ parameters for an M-mixture GMM, with diagonal covariances, where D is the dimension of the data. For a GMM with full covariance matrices, we need to estimate $M(0.5D^2 + 1.5D + 1)$ parameters, which in turn requires large amount of data.

Given a trained GMM and a data point \mathbf{x} , the posterior probability that it is generated by the i^{th} Gaussian component c_i can be computed using the Bayes' rule as follows:

$$P(c_i/\mathbf{x}) = \frac{w_i \mathcal{N}(\mathbf{x}/\mu_i, \Sigma_i)}{p(\mathbf{x})} \quad (3.2)$$

The vector of posterior probabilities for $i = 1, 2, \dots, M$ is called Gaussian posterior vector. Gaussian posterior

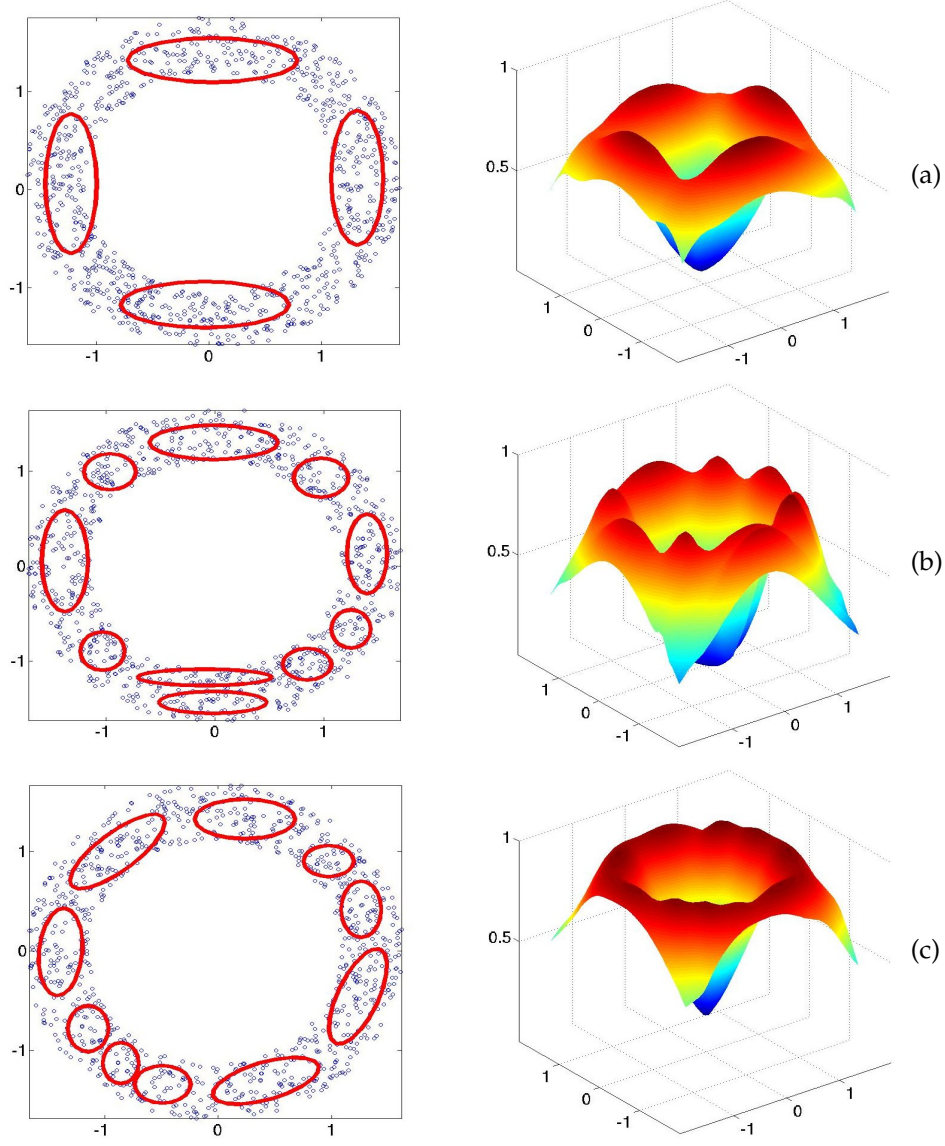


Figure 3.1: Illustration of distribution capturing capability of GMM. GMM trained with diagonal covariance matrices (a) 4-mixtures (b) 10-mixtures and (c) 10-mixture GMM trained with full covariance matrices

Table 3.1: Results of STD system using MFCC and GMM posteriorgrams with 64-mixtures GMM

Metric	MFCC	GMM-64		
	Euclidean Distance	Euclidean Distance	Dot Product	KL Divergence
$P@N$	45.68%	42.89%	51.89%	53.10%
$P@2N$	54.91%	50.68%	63.08%	63.69%
$P@3N$	60.81%	55.67%	67.77%	67.47%
$P@4N$	63.23%	58.54%	71.55%	70.34%
$P@5N$	64.29%	60.96%	73.67%	73.67%

representation was found be better suited for STD than the MFCC coefficients [14], [42].

3.1.1 STD using Gaussian posteriors

In this study, a GMM is built by pooling MFCC feature vectors extracted from 5 hours of speech data collected from different Telugu news channels. Our STD system is evaluated on 1 hour of news data with 30 query words listed in Table.4.4 which are written in IPA script. Using this model, the reference and query utterances are represented as a sequence of GMM posterior features. Speaker invariant nature of GMM posteriors is illustrated, in Fig. 4.2(c) and 4.2(d) using the task of searching for a query word "səmaik^hjə" in the reference utterance "ra:stra:ni səmaik^hjənga unca:ləntu si:ma:ndrone:tələ." We have considered two cases: the query word is from the same speaker as the reference utterance and the query word is from a different speaker. Euclidean distance and symmetric Kullback-Leibler Divergence (KL divergence) are used to find the distance between two MFCC and two posterior features, respectively. The distance matrices computed from MFCC features of reference and query utterances are shown in Fig. 4.2(a) and 4.2(b). The distance matrices computed from GMM posterior features are shown in Fig. 4.2(c) and 4.2(d). In the case of matched speakers, there is a DTW path at correct location, indicating the presence of query word in the reference utterance, in distance matrices computed from both MFCC and posterior features. When the speakers do not match, the desired location of query word is successfully found in the distance matrix computed from the GMM posterior features, but not from the MFCC features. This case study depicts the speaker-invariant nature of GMM posterior features, and thereby their effectiveness in STD.

Subsequence DTW is used to match the GMM posteriorgrams of reference and query utterances, to perform the STD task. Each vector in posteriorgram can be interpreted as a probability distribution. If any element in posterior vector \mathbf{y} is zero then the KL divergence with any other vector is infinity, which is not desired. To avoid this, smoothing method is suggested in [12]. So, the new posterior vector \mathbf{y}_{new} is

$$\mathbf{y}_{new} = (1 - \lambda)\mathbf{y} + \lambda\mathbf{u}$$

where \mathbf{u} is a sample vector drawn from uniform distribution \mathcal{U} , and λ is smoothing constant. For all experiments in this study, λ is empirically chosen as 0.01. Performance of STD system is evaluated in terms of average precision ($P@N$), where N is number of occurrences of the query in the reference utterance. The evaluation metric $P@N$ is calculated as proportion of query words located correctly in top N hits from reference utterance. Detected location is chosen as hit, if it overlaps more than 50% with reference location.

Importance of local distance measure used in Subsequence DTW for different input features is tabulated in Table 3.1. For this experiment 64-mixtures GMM is used. It is observed that for euclidean distance measure, MFCC features are better than GMM posteriors. However, when negative logarithm of dot product and KL divergence are used as local distance measures combined with GMM posteriorgrams, our STD system performance is improved by 9% and 10.21% respectively compared to euclidean distance measure. This shows that euclidean distance is not suitable to find the distance between two probability distributions. For all the remaining experiments KL divergence is used as local distance measure. Table 3.1 also shows results for

Table 3.2: Effect of number of mixtures in GMM on STD performance

Metric	MFCC	GMM-16	GMM-32	GMM-64	GMM-128
$P@N$	45.68%	46.20%	48.90%	53.10%	52.80%

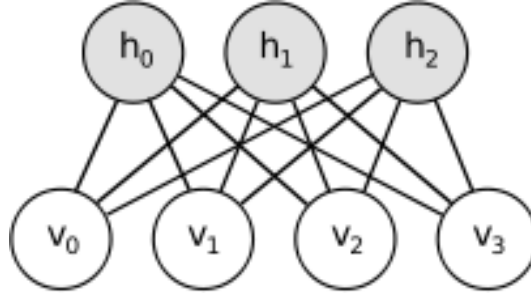


Figure 3.2: Network architecture of a Restricted Boltzmann Machine

different values of k . As k increases, more number of query words can be retrieved. The number of query words that can be recovered in top $2N$ hits using MFCCs can be located in top N hits using GMM posteriorgrams. We can recover all instances of a query using GMM posteriorgrams for smaller value of k .

The performance of the STD system, with varying number of Gaussian components, is given in Table 3.2 with symmetric KL divergence as distance measure. The performance of the system is improved as the number of mixtures increase. It can be seen that for GMM trained with less number of mixtures the STD performance is low, which could be due to clustering of multiple phonemes under same mixture. The lower performance of 128-mixture GMM may be attributed to association of each phoneme class with more than one mixture. It can be clearly seen that GMM posteriors are more robust across speakers, giving better STD performance over MFCC. However, STD performance can be improved further by exploiting dependencies between dimensions. In the next section, extraction of posterior features using GBRBM is explained in which these correlations are be captured with less number of parameters to be estimated.

3.2 Posterior Extraction using GBRBM

A Restricted Boltzmann machine (RBM) is an undirected bipartite graphical model with visible and hidden layers. In contrast to a Boltzmann machine, intra-layer connections do not exist in RBM, and hence the word *restricted*. Example architecture of RBM is shown in Fig. 3.2. In an RBM, the output of a visible unit is conditionally Bernoulli given the state of hidden units. Hence the RBM can model only binary valued data. On the other hand in a GBRBM, the output of a visible unit is conditionally Gaussian given the state of hidden units, and hence it can model real valued data. Both in RBM and GBRBM, the output of a hidden unit is conditionally Bernoulli, given the state of visible units, and hence can assume only binary hidden states. Since the same binary hidden state is used to sample all the dimensions of the visible layer, GBRBM are capable of modelling correlated data.

A GBRBM can be completely characterized by its parameters, i.e., weights, hidden biases, visual biases and variances of the visible units. Since the hidden layer activations are stochastic, any initialization works but badly initialized models require large number of iterations to get converged. For example, initializing all weights to zeros or very large values make large percentage of hidden activations to 0 or 1 which are not representatives of input data point. Usually, weight matrix is initialized to small random values sampled from zero-mean Gaussian distribution. The GBRBM associates an energy for every configuration of visible and hidden states. The parameters of the GBRBM are estimated such that the overall energy of GBRBM, over the ensemble of training data, reaches a minima on the energy landscape. The energy function for GBRBM, for a

particular configuration of real-valued visible state vector \mathbf{v} and binary hidden state vector \mathbf{h} , is defined as [43]

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^V \frac{(v_i - b_i^v)^2}{2\sigma_i^2} - \sum_{j=1}^H b_j^h h_j - \sum_{i=1}^V \sum_{j=1}^H \frac{v_i}{\sigma_i} h_j w_{ij}, \quad (3.3)$$

where V and H are total number of visible and hidden units, v_i is the state of i^{th} visible unit, h_j is the state of j^{th} hidden unit, w_{ij} is the weight connecting the i^{th} visible unit to the j^{th} hidden unit, b_i^v is the bias on the i^{th} visible unit, b_j^h is the bias on the j^{th} hidden unit, σ_i is the variance of the i^{th} visible unit.

The joint density of the visible and hidden unit states is related to the energy of the network as

$$p(\mathbf{v}, \mathbf{h}) \propto e^{-E(\mathbf{v}, \mathbf{h})} \quad (3.4)$$

The parameters of the GBRBM are estimated by maximizing the likelihood of the data. Because of the issues in tractability of true gradient of the likelihood, Markov Chain Monte Carlo (MCMC) approximation methods were used to train RBMs. Contrastive Divergence (CD) [44] is one such technique which is proven to work well in practice. Energy of the system which is directly related to likelihood, is minimized in CD algorithm. In one cycle of CD algorithm, CD_1 , the probability of firing of a hidden unit, j , is activation of sigmoid function for an input of weighted sum of previous layer (visible layer) activations as shown in the following equation. Since the hidden units are stochastic, output is forced to be 1 if output of activation function is greater than a random number sampled from uniform distribution [0, 1]. Binary activations are sent back to visible layer for reconstruction. Visible units are assumed to be Gaussian in GBRBM. Visible activations are sampled from Gaussian distribution with mean equal to weighted sum of inputs from hidden layer and learnt variance. For the second cycle of CD, these reconstructions are fed back as input to visible units. The updates for the parameters can be estimated using Contrastive Divergence (CD) algorithm, as follows:

$$\begin{aligned} \Delta w_{ij} &\propto \left\langle \frac{v_i h_j}{\sigma_i} \right\rangle_{data} - \left\langle \frac{v_i h_j}{\sigma_i} \right\rangle_{recall} \\ \Delta b_i^v &\propto \left\langle \frac{v_i}{\sigma_i^2} \right\rangle_{data} - \left\langle \frac{v_i}{\sigma_i^2} \right\rangle_{recall} \\ \Delta b_j^h &\propto \langle h_j \rangle_{data} - \langle h_j \rangle_{recall} \\ \Delta \sigma_i &\propto \langle \gamma \rangle_{data} - \langle \gamma \rangle_{recall} \end{aligned}$$

where

$$\gamma = \frac{(v_i - b_i^v)^2}{\sigma_i^3} - \sum_{j=1}^H \frac{h_j w_{ij} v_i}{\sigma_i^2}$$

and $\langle \cdot \rangle_{data}$ denotes expectation over the input data, and $\langle \cdot \rangle_{recall}$ denotes expectation over its reconstruction.

During each cycle of CD, the energy associated with the joint configuration of visible and hidden states is supposed to decrease, although there is no theoretical guarantee. After a large number of iterations, the expectation of the energy does not change any more, indicating thermal equilibrium of the network. At thermal equilibrium, the GBRBM models the joint density of the training data. A well trained GBRBM model is capable of generating the data points which resemble the training data.

In the problem of learning underlying probability distribution, probability of a test data point drawn from current model gives an idea of usability of model. But it is difficult to compute probability in the case of GBRBM as calculation of partition function is computationally intensive. However, comparison of free energies of training data and validation data is enough as probability is directly related to free energy. Large difference between free energies of validation data and training data denotes model overfitting. Generally, overfitting occurs when number of examples used for training are not sufficient to estimate model parameters. Generalizability is required for a model to be usable for a test data point unseen in the training data. For a test data point supplied to overfitted model, the outcome is erroneous which can not be expected before hand.

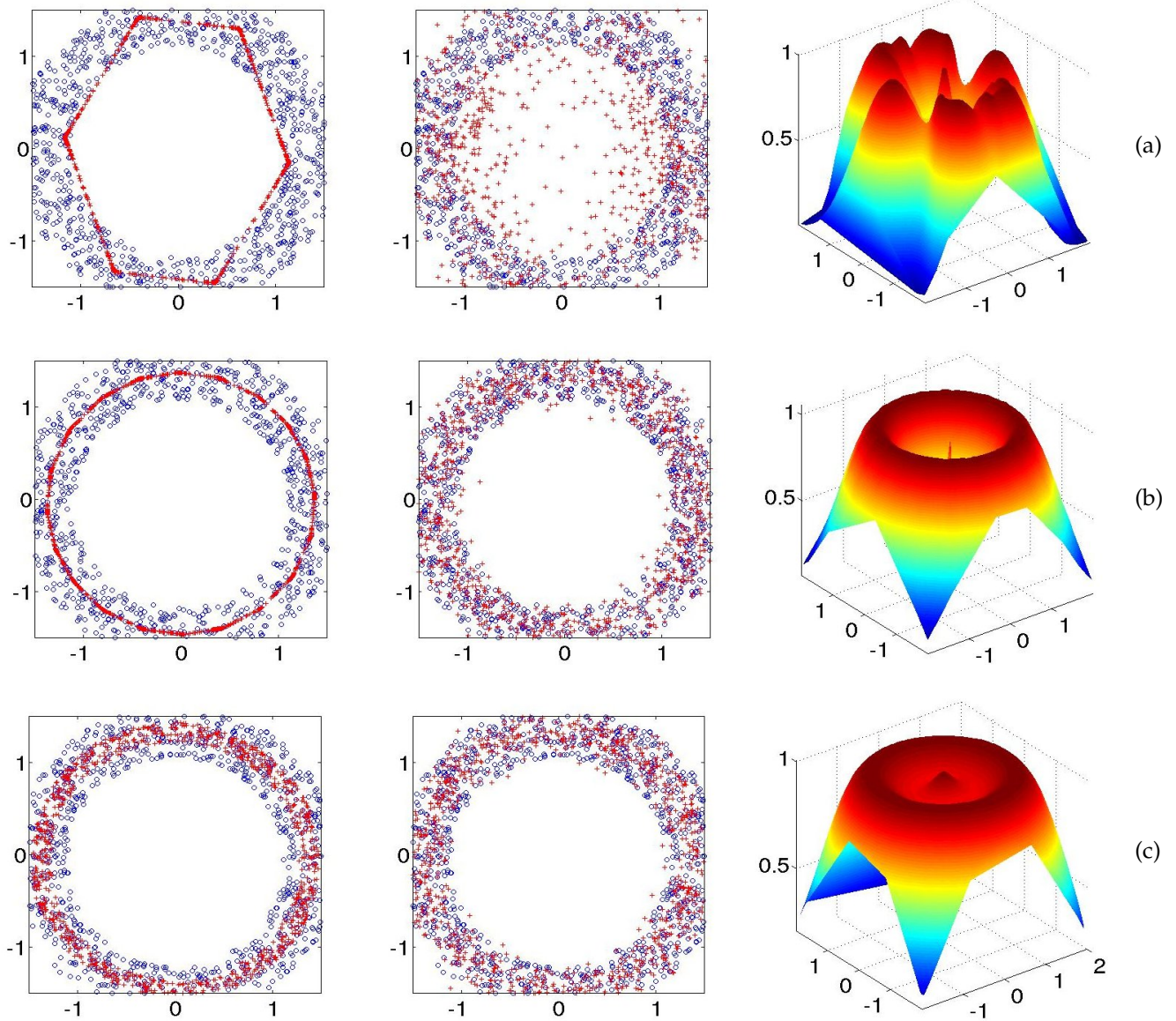


Figure 3.3: Illustration of distribution capturing capability of GBRBM. Left: Original training data (blue), and mean of the unbiased samples generated by trained GBRBM (red), Middle: Original training data (blue), and unbiased samples generated by GBRBM (red), Right: captured density plot. GBRBM plots trained with (a) 3 (b) 10 (c) 200 hidden layer neurons

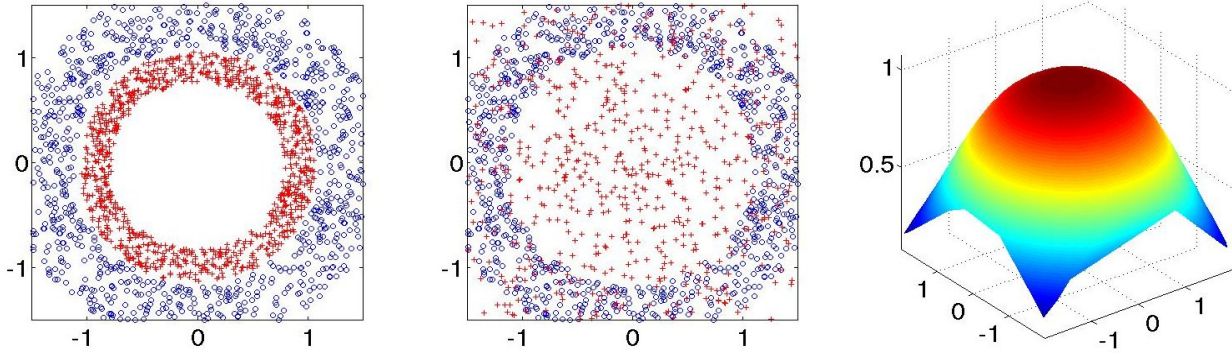


Figure 3.4: Importance of variance learning: Left: Original training data (blue), and mean of the unbiased samples generated by trained GBRBM (red), Middle: Original training data (blue), and unbiased samples generated by GBRBM (red), Right: captured density plot. Plots are generated from GBRBM trained with 10 hidden layer neurons

Overfitting can be avoided by using several techniques: Cross-validation, Regularization, early stopping. In this work, regularization is used to avoid overfitting. Sparsity and weight-decay terms are added in update equations for regularization. Reconstruction error can be used to monitor progress of learning but can not be relied entirely as it does not correlate with objective function, energy equation. It is the difference between input data point and reconstructed visible activations.

The distribution capturing capability of GBRBM is illustrated, in Fig. 3.3, with 2-dimensional data uniformly distributed along a circular ring. First column shows the mean of the unbiased samples generated by the model, second column shows the reconstructed data points, and third column shows the estimated density function. Input data points are plotted as blue 'o' and reconstructions are plotted as red '+'. A GBRBM with different number of hidden units is trained, to capture the joint density of this data, for 200 cycles using CD. The number of hidden units plays an important role in capturing the distribution of input data. For a GBRBM with H hidden units, the hidden state vector can take at most 2^H binary configurations. However, only a few of those 2^H hidden states sustain at the thermal equilibrium of the network. Hence, the reconstructed visible state can take a maximum of 2^H different mean vectors. The mean visible layer activations, for a GBRBM with 3 hidden units is shown in Fig. 3.3(a). In this case the mean of the circular ring is approximated by a hexagon, i.e., with 6 (out of $< 2^3$ possible) stable states at thermal equilibrium. As the number of hidden units increase, the number of possible stable states also increase, leading to a better approximation of the mean of the input data. The mean of the unbiased samples generated by a GBRBM with 10 hidden units, in Fig. 3.3(b), faithfully estimated the mean of the circular ring. When the number of hidden units is further increased to 200, the mean activations are spread over input data leading to an overfit. The data distributions captured by GBRBMs, with different hidden units, are shown in the third column of Fig. 3.3. It is clear that the GBRBM with 10 hidden units has captured the input distribution better.

The variance parameters of the visible units also influence the distribution capturing capabilities of the GBRBM. Usually, when the GBRBM is used to pre-train the first layer of an MLP the variances are simply set to one. However, variance learning is necessary when GBRBM is used to capture the joint density of the input data. Fig. 3.4 shows the density captured by GBRBM, with 10 hidden units, trained without variance parameter. Clearly, without variance learning, the GBRBM failed to capture the underlying joint density of the data.

From computational point of view, number of parameters to be estimated in GBRBM are $NM + N + 2M$ which is very high as number of neurons increase which results in data insufficiency. To overcome this problem, sparsity constraint [45] has been enforced on hidden layer activations through which many hidden neurons are forced to zero activations resulting in fewer parameters which is also helpful for better generalizability of the model. With very less number of parameters to be estimated, dependencies can be exploited using GBRBM thus avoiding the problem of data insufficiency. It is also observed that GBRBM captures distribution very

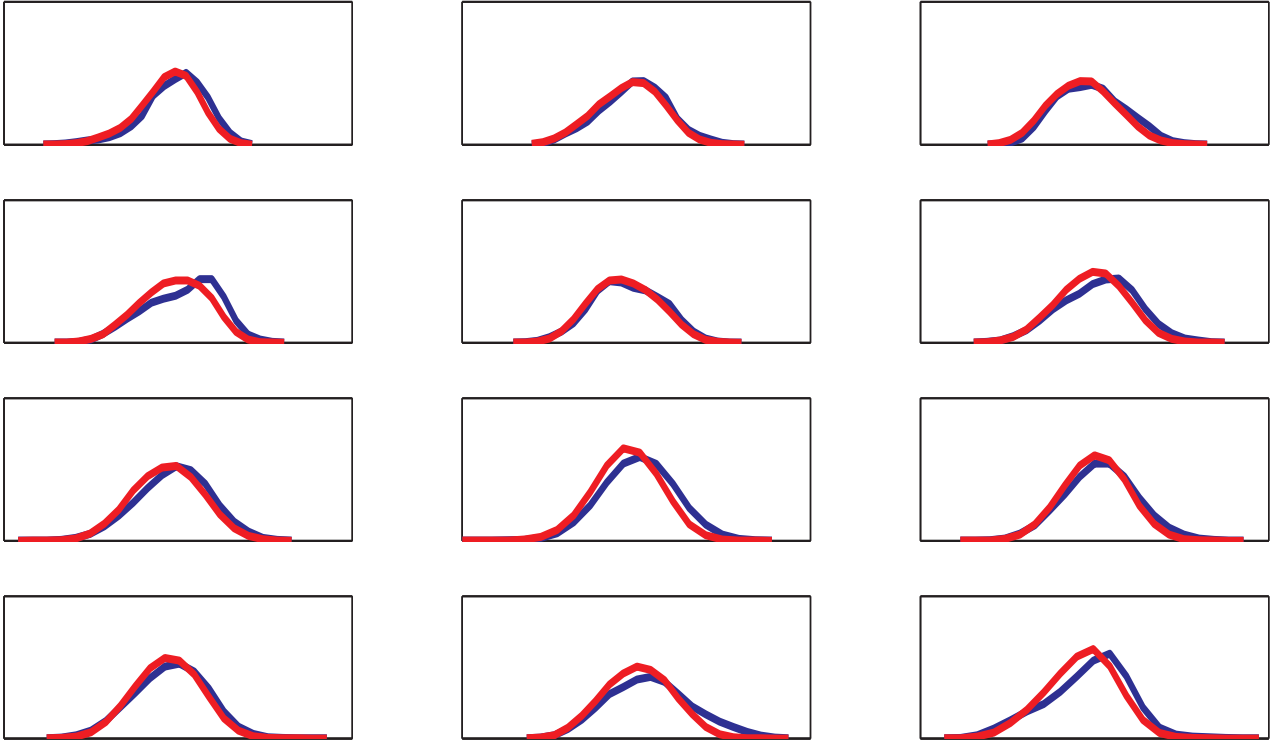


Figure 3.5: Comparison of original (red) and estimated (blue) marginal densities of first 12 dimensions of MFCC features

well compared to GMM in any case as can be seen from Fig. 3.1 and Fig. 3.3. But GBRBM has a limitation that it is very difficult to train to obtain optimal parameters.

3.2.1 STD using GBRBM Posteriors

A GBRBM, with 50 hidden units is trained, using 39-dimensional MFCC features, obtained from 5 hours of news data, to capture the density of the speech data in the acoustic space. The marginal densities of the original (blue) and estimated (red) MFCC features are shown in Fig. 3.5. The marginal densities of the first 12 MFCC features is shown to illustrate the effectiveness of GBRBM in capturing the joint density of the data. In this chapter, the state of the hidden units, at thermal equilibrium, is used as a feature for STD task. The probability that the j^{th} hidden neuron assumes a state of '1', given the state of the visible units (MFCC features) is computed as

$$P(h_j = 1 | \mathbf{v}) = \text{sigmoid} \left(\sum_{i=1}^V \frac{v_i}{\sigma_i} w_{ij} + b_j^h \right) \quad (3.5)$$

The posterior probability vector, representing the state of all the hidden units, is used to match the reference and query utterances. The distance matrices computed from GBRBM posteriors, with symmetric KL divergence, for matched and mismatched speaker conditions are shown in Fig. 4.2(e) and Fig. 4.2(f), respectively. In the case of mismatched speakers, the distance matrix computed from GBRBM posteriors helped to spot correct location of the query word as shown in Fig. 4.2(f), which is not the case using distance matrix computed from MFCC features as shown in Fig. 4.2(b). Hence the GBRBM posterior representation is better than MFCC features for STD task.

Our STD system is evaluated on 1 hour of read news data with 30 query words listed in Table. 4.4. The performance of STD system built with GBRBM posteriors is given in Table 3.4. GBRBM-NV denotes GBRBM trained without variance parameter learning. In this case, STD performance is no better than MFCC which denotes that density of input data is not captured. The performance of the GBRBM posteriors is slightly better than the GMM posteriors which can be attributed to ability of GBRBM to exploit the dependencies between

Table 3.3: Results of STD with various number of hidden neurons

Metric	GBRBM hidden neurons			
	30	50	70	90
$P@N$	50.83%	57.64%	56.58%	56.43%

Table 3.4: Performance comparison of STD systems built using different unsupervised posterior representations

Metric	MFCC	GMM	GBRBM-NV	GBRBM	GBRBM+GMM
$P@N$	45.68%	53.10%	45.84%	57.64%	59.91%

dimensions of input data. Since GMM and GBRBM are two different unsupervised data modeling techniques, the evidences from both these systems are combined linearly. The performance of the combined system is better than the performance of either of the systems alone.

3.3 Conclusion

In this chapter, unsupervised approaches GMM and GBRBM are explored to extract robust features for STD task. It is observed that using GMM, joint density of input data can be better captured with sufficiently high number of mixtures. But with number of mixtures, complexity of the model also increases enormously leading to data insufficiency. Through experiments it is found that for 64 mixtures GMM, approximately 7.5% improvement is observed compared to MFCC proving robustness of GMM posteriors. Further improvement in STD performance is achieved by using GBRBM posteriors. It is shown through experiments that optimal number of hidden units are required for better generalization of trained GBRBM model. Less number of hidden units leads to underfitting and memorization of training data occurred with more number of hidden units. Number of hidden units is chosen through experiments on speech data. It is learnt that variance parameter is crucial in capturing joint density and GBRBM posteriors extracted without variance learning are no good than MFCC in STD task. It is understood that GBRBM posteriors are more robust than GMM posteriors as around 4.5% improvement is observed compared to GMM posteriors. Further improvement is observed with the fusion of GMM and GBRBM posteriors indicating presence of complimentary information in them.

Chapter 4

Supervised Approaches

In supervised approaches, hidden patterns in the input data are captured with the help of labelled information. Generally, training in these approaches include optimal mapping of input samples to target class. A combination of generative and discriminative models have proven to be effective for complex classification tasks, like speech recognition [46]. Generative models estimate joint density of the input data, while discriminative models capture the boundaries between the classes. Examples for generative models include GMM, HMM, Restricted Boltzmann Machines (RBM) and Gaussian-Bernoulli RBM (GBRBM). Examples for discriminative models include Support Vector Machines (SVM), and Multi Layer Perceptron (MLP). In this work, HMM-MLP hybrid modelling is attempted for posterior extraction.

4.1 Posterior Extraction using Hybrid HMM-MLP

It was shown that a combination of HMM-MLP hybrid modelling is better suited for phoneme recognition, than either one of them alone [47]. In the HMM-MLP hybrid modelling, HMM is aimed at handling the varying length patterns, while the MLP is aimed at estimating the non-linear discriminant functions between the phoneme classes. In this approach, phoneme boundaries obtained from HMM are used to train MLP. The phonetic posteriorgrams, estimated from HMM-MLP hybrid modelling, can be used as a representation for STD task.

4.1.1 Speech segmentation using HMM

HMMs are doubly stochastic models, and can be used to model non-stationary signals like speech. Assuming that a state represents a specific articulatory shape of the vocal tract, the speech signal (observation sequence) can be represented by a sequence of states. Here, the state sequence is not known (hidden) to us. Through HMM training, parameters of each state are obtained to capture the statistical properties of speech signal in that state. Generally, each phoneme is modelled using a three-state left-right HMM, assuming that the phoneme is produced in 3 phases. For example, the production of a stop-consonant consist of three main phases, namely, closure phase, burst phase and transition phase into succeeding phoneme. More number of states can also be used for better modelling, but it requires large amount of data for training. In this evaluation, each phoneme is modelled as a three-state HMM with 64 Gaussian mixtures per state. Labelled speech data is used to estimate the parameters of the HMM, which include the initial probabilities, emission probabilities and state-transition matrices, using Baum-Welch algorithm [48]. The trained HMM models are used to align the manual transcriptions with the speech data, to obtain the phoneme boundaries. The phoneme boundaries obtained from forced alignment are used for training a Multi Layer Perceptron (MLP).

Table 4.1: Recognition accuracy for using different number(C) of phoneme grouping

C	HMM	MLP
6	77.88	81.87
15	70.6	76.02
25	69.51	74.24
45	62.68	69.11

4.1.2 Extraction of Phonetic Posteriors using MLP

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of inputs onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, as in Fig. 4.1, with each layer fully connected to the next one. Each node, in the hidden and output layers, is a neuron (or processing element) with a non-linear activation function. Sigmoid and hyperbolic tangent activation functions are typically used in the hidden layers, while softmax activation function is used in the output layer. The output of the softmax function can be interpreted as posterior probability of the class given the input frame. The weights of the network can be learnt, using back-propagation algorithm, by maximizing the cross entropy between the estimated posteriors and actual phoneme labels [49].

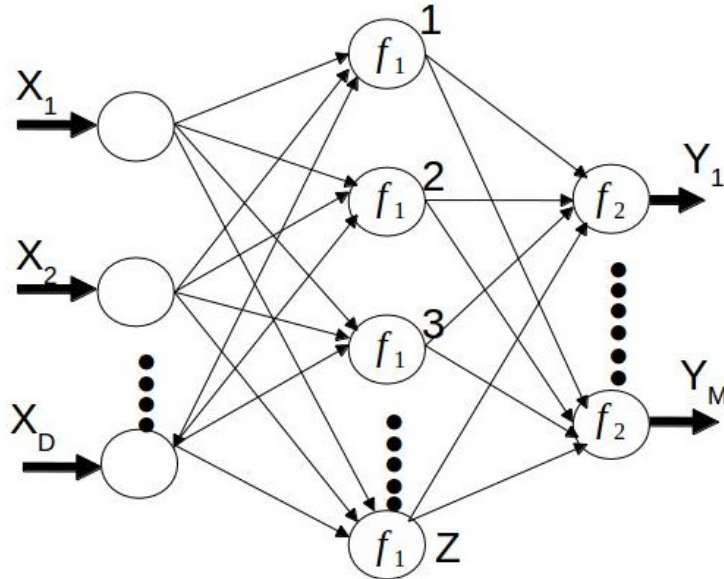


Figure 4.1: MLP Network. X and Y are the input and output vectors respectively. D , Z and M denote the number nodes at the input, hidden and output layer respectively. f_1 and f_2 denotes the activation functions at hidden and output layer respectively.

Unlike GMM, an MLP can be trained with higher dimensional correlated data. Hence, context information can be learnt using MLP by presenting concatenated speech frames as input. A context of 13-frames is used with 39-dimensional MFCC features to form a 507 ($39 * 13$) dimensional input feature vector. Phoneme labels obtained from the HMM forced alignment are used as output classes. An MLP with single hidden layer, having 1000 sigmoid units, is trained to map the input feature vectors to the phoneme label. The performance of HMM-MLP hybrid approach is evaluated on 5 hours of Telugu broadcast news data, in which 3 hours of data is used for training and the remaining 2 hours was used for testing the system. The performance of HMM-MLP hybrid approach is shown in Table 4.1 for different configurations of phoneme groupings as given in Table 4.2. The performance of HMM system alone is also given for comparison. The HMM-MLP hybrid system is consistently better than the HMM alone. As the number of classes increase, there is a gradual drop in the recognition accuracy.

Table 4.2: Grouping of the phonemes into different classes

45 classes	a	a:	i	i:	j	u	u:	e	e:	o	o:	v	f	ʃ	s	h	f	m	n	k	k ^h	g	g ^h	c	c ^h	z	j ^h	j ^h	ʒ	j	t	t ^h	d	d ^h	t	t ^h	d	d ^h	p	p ^h	b	b ^h	r	l	sil
25 classes	a	i	j	u	e	o	v	s	h	f	m	n	k	g	c	j	t	d	t	d	p	b	r	l	sil																				
15 classes	a	i	u	e	o	F(Fricatives)			N(Nasal)		G(Glottal)		P(Palatal)			R(Retroflex)		D(Dental)		B(Bilabial)		r	l	sil																					
6 classes	V(Vowel)						F(Fricatives)			N(Nasal)		C(Consonants)										T(Trill and Liquid)		sil(Silence)																					

Table 4.3: Average performance of STD obtained with different phoneme classes

Metric	6 classes	15 classes	25 classes	45 classes	raw MFCCs
$P@N$	44.05	77.65	80.13	72.36	45.68
$P@2N$	55.68	86.50	89.13	80.57	54.91
$P@3N$	59.17	88.10	90.75	82.39	60.81

4.1.3 STD using Phonetic Posteriors

The MFCC features extracted from every 25 ms frame of speech signal is converted into phonetic posterior representation using the trained MLP. Since phonetic posteriors are obtained from the MLP, trained with large amount of speech data collected from several speakers, they are more robust to speaker variability. Hence they are better suited for the STD, than the raw MFCC features. Phonetic posteriors of word "səmaik^hjə" spoken by two different speakers are shown in Fig. 2.1 which can be better matched. Speaker invariant nature of phonetic posteriors is illustrated, in Fig. 4.2(g) and 4.2(h). The distance matrices computed from MFCC features of reference and query utterances are shown in Fig. 4.2(a) and 4.2(b). The distance matrices computed from posterior features are shown in Fig. 4.2(g) and 4.2(h). In the case of matched speakers, there is a DTW path at correct location, indicating the presence of query word in the reference utterance, in distance matrices computed from both MFCC and posterior features. When the speakers do not match, the desired location of query word is successfully found in the distance matrix computed from the MLP posterior features, but not from the MFCC features. This shows the robustness of phonetic posteriors to speaker variability

Subsequence DTW is employed to match the posterior features extracted from the reference and query utterances, and detect the possible matching locations of query in the reference utterance. This method is evaluated on 1 hour of Telugu broadcast news data with 30 query words spliced from continuous speech data. The performance of the STD system obtained with different phoneme classes is given in Table 4.3. Even though the phoneme recognition accuracy increased with reducing the number of phoneme classes, it did not result in improved STD. There is a significant decrease in $P@N$ from 15 to 6 classes. Since several phonemes are grouped together into a single class, the number of false matches increases, and results in poor performance. The best performance was achieved with 25 phoneme classes, in which the aspirated and unaspirated stop constants produced at the same place of articulation are grouped together. Hence 25 phoneme classes are used for all further studies in this chapter.

The performance of the HMM-MLP system is much better than the performance of the unsupervised posteriors obtained from GMM and GBRBM. This is because the phonetic posteriors are obtained using a supervised approach, while the GMM and GBRBM posteriors are extracted without using labelled information.

The performance of the STD system, built using different posterior features, on 30 query words from Telugu language is presented in Table 4.4. Assuming that the probability of misclassification is same for all the syllables, miss rate of longer query words is less compared to smaller query words. On an average, this can be observed in the Table 4.4 for all representations. For longer query words, the performance is almost similar, with all the three representations, but for smaller query words HMM-MLP posterior features perform better than GMM and GBRBM posteriors.

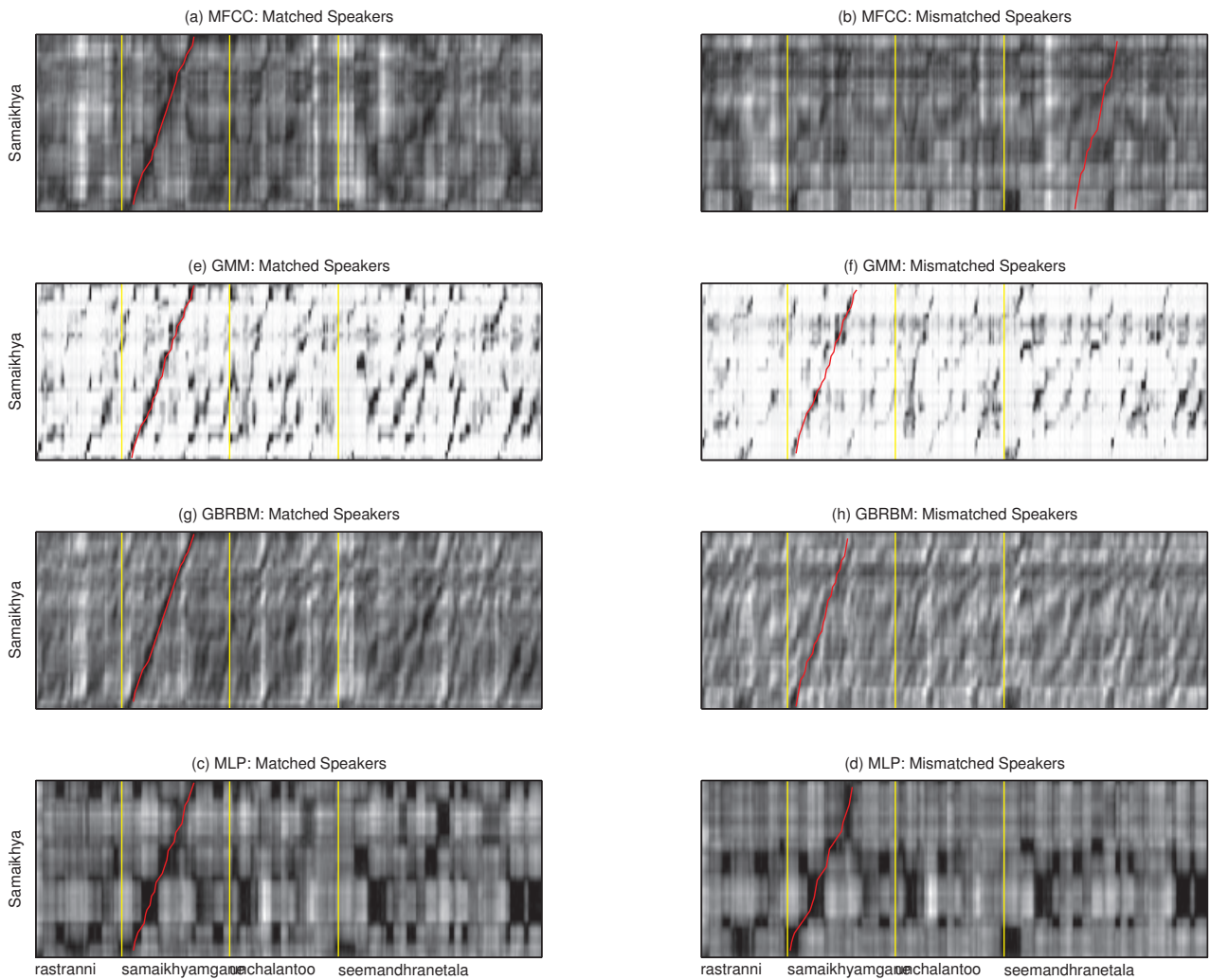


Figure 4.2: Illustration of effectiveness of posterior features for STD over MFCC. Distance matrix along with DTW path computed from (a) MFCC features for matched speakers, (b) MFCC features for mismatched speakers, (c) GMM posteriors for matched speakers, (d) GMM posteriors for mismatched speakers, (e) GBRBM posteriors for matched speakers, (f) GBRBM posteriors for mismatched speakers, (g) MLP posteriors for matched speakers, (h) MLP posteriors for mismatched speakers

Table 4.4: Query words (written in IPA) with their P@N(%) for Telugu language

Query word	HMM-MLP	GMM	GBRBM	Query word	HMM-MLP	GMM	GBRBM
prənbmuk ^h əɾji	99.75	100	100	digvijəjsing	82.78	50	60
teləŋga:nə	83.50	90	90	adjəks ^h urə:lu	76.96	50	50
səmarvəsəm	88.97	84.09	86.36	prəb ^h utvəm	71.43	48.57	57.14
səiləjəmat ^h	84.64	80	60	ad ^h ikarrulu	85.31	42.85	64.29
alpəpirdənəm	84.62	76.92	69.23	haidra:bə:d	83.57	42.85	71.43
parləment	88.24	76.47	76.47	kəm:tji	58.82	35.29	35.29
bəŋga:la:k ^h atəm	81.75	75	75	emikəlu	72.25	35.13	40.54
kəŋgɾes	78.00	74	78	erpat:tu	63.38	31.8	40.91
ra:ji:na:ma	85.19	70.37	59.26	vartarvənəm	67.00	30	50
ne:pət ^h jəm	62.69	69.23	76.92	vib ^h əjənə	71.43	28.57	33.33
pəncə:jəti	76.62	68.96	82.76	səmaik ^h jə	93.55	22.58	54.84
so:mija:ga:nd ^h i	90.83	66.66	83.33	dilli:	50.00	20.83	41.67
porliŋg	63.75	62.5	75	vi:varə:lu	80.00	20	59.91
kirənkumar:re:d̪i	95.53	57.14	85.71	ru:pə:ji	70.00	20	40
nirnejəm	83.33	55.55	63.89	məntri	32.73	14.28	12.70

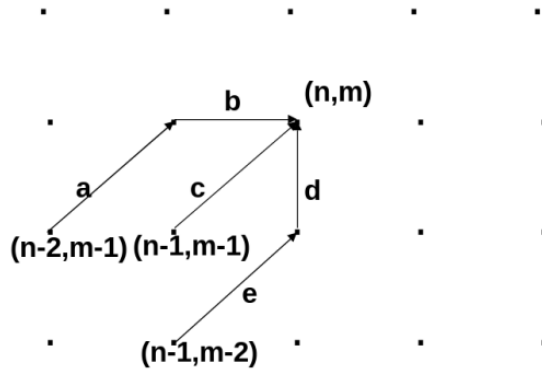


Figure 4.3: Subsequence DTW path with local weights a b c d e

Table 4.5: Comparison of STD performance, with queries spliced from continuous speech and queries recorded in isolation

Metric	$P@N$	$P@2N$	$P@3N$
Spliced from read data	80.49	88.61	90.10
Isolated recordings	56.02	66.70	69.66

4.1.4 Effect of speaking mode

The experiments, reported in the previous section, were conducted on 30 queries spliced from continuous read speech. In this section, the performance of the STD system is evaluated on the query words recorded from 20 native Telugu speakers in an isolated manner. It is observed that the duration of the query words recorded in isolated manner is almost double the duration of those spliced from continuous speech [50]. Since the query words are recorded in a different environment, there is channel mismatch between the reference and query words. Both these factors (duration and channel mismatch) lead to significant drop in the STD performance. In order to mitigate the effect of duration mismatch, experiments are conducted using different constraints on the warping path, as shown in Fig. 4.3. The weights (a , b , c , d , e) can be used to modify the shape of the warping path. A vertical path can be favoured by decreasing d and e , whereas a horizontal path can be favoured by decreasing a and b . A diagonal path can be favoured by decreasing c . In the case of isolated queries, whose duration is much longer, a vertical path should be favoured. The best performance with isolated queries was obtained with weights (2, 3, 2, 1, 1). In order to normalize the channel effects, cepstral mean subtraction and variance normalization are performed for every utterance. The average performance of STD, for both spliced and isolated queries, is given in Table 4.5. The performance of the STD system with isolated queries is almost 25% less than that of with the spliced queries.

4.2 Conclusion

In this chapter, phonetic posterior extraction for STD task is presented by using HMM-MLP hybrid model. MLP is trained on the phoneme boundaries obtained using trained HMM. Phoneme recognition accuracy of HMM-MLP model is better than HMM model alone. Phonetic posteriors extracted from MLP are supplied to matching algorithm, Subsequence DTW. Optimal number of classes is chosen through experiments. Independence of phoneme recognition accuracies and STD accuracies is observed. Effect of speaking mode of query on STD performance is also studied. As the durations of recorded query and queries spliced from continuous speech vary greatly, matching plays important role. By careful selection of path weights in DTW, STD performance is improved for recorded queries. Approximately 24% difference in performance is observed in STD performance when speaking mode changed. Our future efforts will be on enhancing the STD performance for recorded queries.

Chapter 5

Analysis of Features from Analytic Representation of Speech using MP-ABX Measures

Feature extraction is a prominent task in speech processing and the capability of features to capture relevant information is a major factor deciding the efficiency of numerous speech processing algorithms. Hence, analysis of the information content in features becomes important. Evaluation of adeptness of features are usually carried out by examining their performances in recognition tasks. Recognition systems now a days employ some form of machine learning/modelling techniques. For example, same MFCC features were used to extract GMM and GBRBM posteriors for STD task where superior performance of GBRBM posteriors can be attributed to effective GBRBM modelling. Also, they may use supervised learning with labelled speech database, which is known to deliver faithful efficiency as in the case of MLP posteriors for STD task. In such cases, the effectiveness in terms of scores attained by the recognition system cannot be attributed to features alone. Hence this cannot be a legitimate procedure for evaluating features. We need much simpler strategies directly dealing with features themselves, rather than employing complex modelling processes.

The ABX measures were proposed as a suitable means for analysing features extracted from speech signals. ABX tasks consist of context or speaker dependant discrimination tasks between pairs of speech stimuli. These measures were introduced in speech processing to study auditory perception of different sounds by both ears using listening tests [51]. The ABX measures were also used to study effects of accents within language on perception by children with speech difficulties [52]. Objective evaluation of features using ABX measures was presented in [53, 54, 55] where dynamic time warping (DTW) was employed to calculate similarity between pairs of speech stimuli. In [53] similarity measures between features were evaluated for discriminating words. In [54], significance of each intermediate stage in the extraction of MFCC/PLP is analysed using ABX measures. The noise robustness of various features was studied using ABX measures in [55].

In this chapter, we analyse the performance of features derived from complex analytic domain representation of speech using MP-ABX tasks. The FDLF coefficients extracted from analytic magnitude and IF features extracted from analytic phase are considered. These features are tested using ABX tasks set and their performances are compared with that of MFCC features. It is observed that the magnitude based features- MFCC and FDLF are demonstrating better phoneme discrimination capability irrespective of speaker variance than phase based features. Whereas, the IF features extracted from analytic phase presented better speaker discrimination ability irrespective of phoneme variance than FDLF. Also the existence of complementary information in these features is explored by feature concatenation and examining resultant ABX scores.

The rest of the chapter is organized as follows: Section 5.1 explains the ABX measures used in this study. Section 5.2 discusses analytic features extracted from speech which are being studied in this paper. The evaluation of analytic features with respect to ABX tasks is presented in Section 5.3. A detailed discussion

Table 5.1: Example triplets for MP-ABX tasks. SP stands for speaker

Task	A	B	X	Answer
PaC	/ba/ SP1	/ga/ SP1	/gu/ SP1	B
PaT	/ba/ SP1	/ga/ SP1	/ba/ SP2	A
TaP	/ba/ SP1	/ba/ SP2	/ga/ SP1	A
CaT	/ba/ SP1	/ga/ SP1	/gu/ SP2	B

on the nature of information present in different features, effects of feature extraction parameters, existence of complementary information etc. are presented in Section 5.4. In Section 5.5, we summarize the results of analysis of features.

5.1 MP-ABX discrimination tasks

We have adopted the minimal pair ABX (MP-ABX) discrimination tasks explained in [54] for our study. Pattern discrimination using MP-ABX tasks employ three speech stimuli- A, B and X triplet, which are consonant-vowel (CV) pairs and measure the discrimination capability between such CV pairs. In each task A and B differ by one phoneme or speaker, and X is chosen such that it is close to A or B. We have considered four discrimination tasks, namely, Phoneme across Context (PaC), Phoneme across Talker (PaT), Context across Talker (CaT) and Talker across Phoneme (TaP). In PaC task, all the stimuli in triplet are spoken by the same speaker, in order to evaluate the robustness of features across context. A and B differ by one phoneme and test stimulus X is chosen to have one phoneme in common with either A or B. To measure the invariance of features across speakers, PaT task is chosen. In this task, A and B are spoken by same speaker, but X by a different speaker. These two tasks intend to show the effectiveness of features in conveying linguistic information in speech signals irrespective of speaker as well as context, demonstrating their efficiency for speech recognition applications. In TaP task, discrimination ability of features in distinguishing speakers is measured, examining their usefulness to speaker recognition applications. A and B are spoken by two different speakers, but have same phoneme sequence. X consist of a different phoneme sequence, but spoken by the same speaker as of either A or B [54].

In all the above tasks, either speaker or context is kept invariant, which is a controlled condition and cannot be demanded in real world scenarios. CaT task is proposed to evaluate the robustness of features across different contexts and different speaker identities. In this task, the triplet is chosen as in PaC task, but X is spoken by a different speaker. This task assumes significance in all speech recognition applications where features are required to be robust across speakers as well as context. The summary of all tasks utilized in this work is given in Table.5.1.

To evaluate each discrimination task, we follow the procedure described in [54]. The aim of evaluating each task is to find out the stimuli closest to test stimuli X. We compute dissimilarity between the pairs of phoneme sequences: A-X and B-X using DTW. Cosine similarity is used as distance metric in DTW which is recommended by [53]. An error will be identified when X shows more similarity to the stimuli which is not the correct answer. Mean error is computed by averaging the number of errors across speakers and number of trials using different triplets.

5.2 Analytic features of speech

The complex analytic domain representation of a continuous time signal $s(t)$ is given by [56]:

$$s_a(t) = s(t) + js_h(t) \quad (5.1)$$

where $s_h(t)$ is the Hilbert transform of the real signal $s(t)$, which is expressed as $s_h(t) = \mathcal{F}^{-1}\{S_h(j\Omega)\}$. \mathcal{F}^{-1} denotes inverse Fourier transform and $S_h(j\Omega)$ is obtained as:

$$S_h(j\Omega) = \begin{cases} +jS(j\Omega) & \Omega < 0 \\ -jS(j\Omega) & \Omega > 0 \end{cases} \quad (5.2)$$

where $S(j\Omega)$ is the Fourier transform of $s(t)$ [57]. The analytic signal $s_a(t)$ can be expressed in polar form as

$$s_a(t) = a(t)e^{j\phi(t)} \quad (5.3)$$

where $a(t)$ and $\phi(t)$ are the time-varying magnitude and phase of the analytic signal, respectively. If $s(t)$ is a narrowband signal, $a(t)$ and $\phi(t)$ can be perceived as amplitude modulated (AM) and frequency modulated (FM) components of $s(t)$ [56]. Natural signals, including speech, cannot be represented by single time varying AM-FM component as they are mostly wideband signals. The multiband AM-FM demodulation of wideband speech signals was described in [58] and was utilized for formant tracking in [59, 60] using filterbank analysis. Features obtained from AM-FM demodulation were employed for speech recognition [61, 62], speaker recognition [63, 64, 65] etc. A narrowband component extracted from speech signal in Figure.5.1(a) is shown in Figure.5.1(b), together with its AM and FM components in Figure.5.1(c) and Figure.5.1(d) respectively.

INCLUDE CAPTION FOR FIGURE F

5.2.1 Frequency domain linear prediction

Feature extraction from AM component using autoregressive modelling of temporal envelope of speech was discussed in [66, 67] which is the frequency domain dual of conventional linear prediction (LP) analysis. While LP analysis captures autocorrelations from time domain and models the magnitude spectral envelope in frequency domain, the FDLP analysis captures spectral autocorrelations and models temporal envelope. Analysis of temporal resolution of FDLP feature extraction was done in [68] and noise robustness capabilities were studied which showed significant improvements in phoneme recognition performance [69]. The temporal envelope obtained from a 16th order FDLP analysis of the signal in Figure.5.1(b) is shown in Figure.5.1(f), which demonstrates credible approximation of the original envelope of the signal in Figure.5.1(c).

5.2.2 Instantaneous frequency

The direct computation of analytic phase suffers from phase wrapping problem. It is not possible to draw meaningful inferences from analytic phase directly (see Figure.5.1(d)). But the time derivative of analytic phase, the IF is free from phase wrapping and can be computed unambiguously. IF is the frequency of a sinusoid which locally fits a signal and wideband signals cannot be represented accurately with single sinusoid. Hence IF computation is meaningful only when the signal under consideration is decomposed into narrowband components. The IF computed from a narrowband component of speech shows variations in its frequency, around the centre frequency of corresponding narrowband filter (See Figure.5.1(e)). These variations correspond to formant transitions of speech. The post processing and extraction of features from IF is explained in [65] and are explored further using ABX tasks in this paper.

5.3 Evaluation of analytic features

The ABX triplets for MP-ABX tasks were obtained from TIMIT database by segregating required CV sequences from continuous speech. The phonetic level transcription available with the TIMIT database was used to extract these CV sequences. Each of the MP-ABX tasks was performed upon an approximate number of 5000 triplets spoken by 190 male and 90 female speakers.

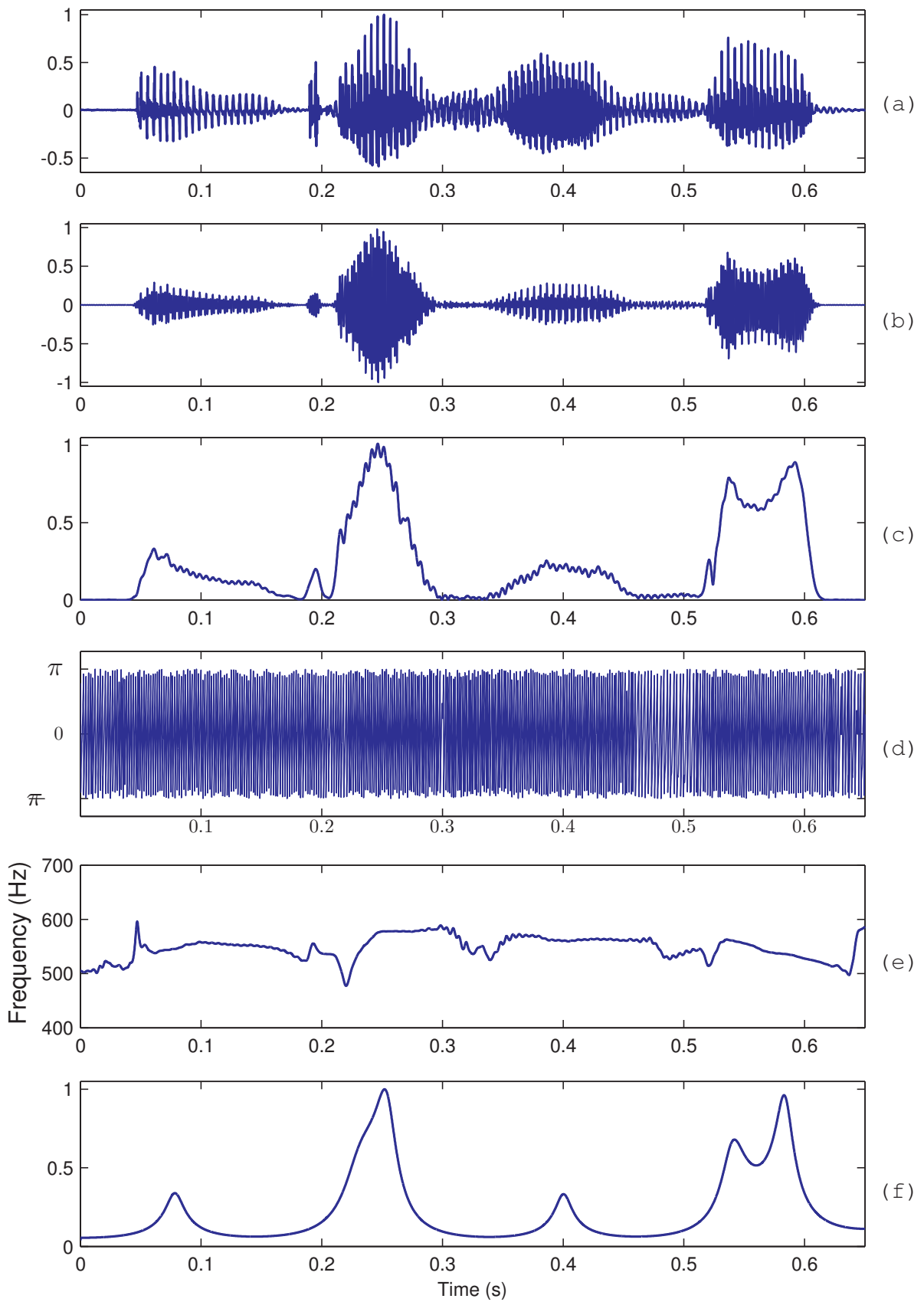


Figure 5.1: Figure illustrating AM-FM components of a speech signal (a) Segment of speech signal, (b) Filtered narrowband component ($f_i=440\text{Hz}$), (c) Magnitude envelope, (d) Phase Component, (e) Smoothed IF and (f) FDLP Temporal Envelope.

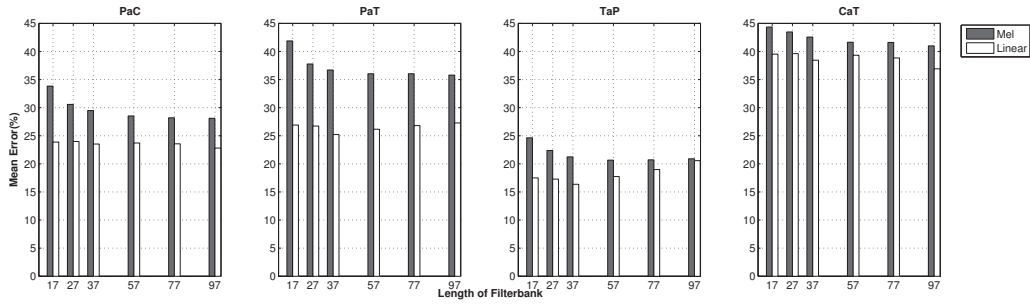


Figure 5.2: Effects of lengths and spacing of filterbank on Mean errors of MP-ABX tasks.

The speech signals were sampled at a frequency of 16kHz. The required features- MFCC, FDLP and IF were extracted from 25 ms long segments of CV sequences shifted by 10 ms. Standard 39 dimensional MFCC were extracted using mel filterbank of length 47. Similarly, 39 dimensional FDLP features were extracted using a 47-filters mel filterbank. The length of filterbank used for IF feature extraction is studied further, as IF is strictly a property of narrowband signal, where filter bandwidths and spacing crucially affect its effective computation.

In each MP-ABX task, dissimilarity scores between A-X and B-X pairs were obtained by performing DTW on their features. The outcome of each trial was decided by comparing the dissimilarity scores from the A-X and B-X pairs. Mean error score (in %) was calculated by averaging the number of failed trials over the entire set of triplets under consideration.

Analysis of filterbank parameters in IF based feature extraction is done with respect to MP-ABX tasks. The mel spacing and linear spacing of filters in the filterbank is explored initially. The comparison between mel and linear filterbanks with different lengths, is shown in Figure. 5.2. The performance of mel filterbank is consistently less than that of linear filterbank at all lengths. The frequency resolution of mel filterbank decreases considerably at high frequency (HF) ranges and thus the bandwidths of filters with higher centre frequencies are much high compared to those at low frequency range. Thus filtered components from HF region of speech signal lose their narrowband property, resulting in inefficient computation of IF and consequently poor performance. Hence linear filterbank is chosen for IF based feature extraction.

The effects of lengths of filterbank are studied using MP-ABX tasks, the results of which is shown in Figure. 5.2. When the number of filters is very less, their bandwidths have to be increased to span the entire speech spectral range under consideration. Thus the filters will cease to be narrowband, making the IF computation inefficient, resulting in poor performance. On the other hand, when there exist exceedingly large number of filters in the bank, each filter will have minimal bandwidths insufficient to capture the frequency variations around its centre frequency. This will result in grave information loss upon IF computation, delivering increased errors. The length of filterbank is fixed as 37, based on the results of this analysis.

In [65], discrete cosine transform (DCT) is applied over the IF values computed from different frequency bands and first few DCT coefficients are retained, to account for redundancies within. The variation of mean errors with respect to number of DCT coefficients retained is shown in Figure. 5.3. We have observed that the number of DCT coefficients to be retained should be greater than or equal to number of filters to obtain appropriate mean errors. This implies to the possibility of very less or no redundancy between IF computed from different frequency bands. As IF shows the local deviation in frequency about the centre frequency, the chances of existence of redundancies within information across neighbouring frequency bands is very less. As the extend of redundancy is very less, the DCT based compression results in information loss. Hence we have omitted the use of DCT and used the average values of IF over segments of speech from 37 frequency bands to obtain 37 dimensional feature vectors.

The MP-ABX tasks are performed on MFCC, FDLP and IF features and the mean errors for all tasks are shown in Table. 5.2. It can be seen that MFCC and FDLP are performing almost equivalently in PaC and PaT tasks, showing their effectiveness in conveying linguistic information in speech signals, making them suitable for speech recognition applications. FDLP is performing exceptionally well in CaT task, implying that it is

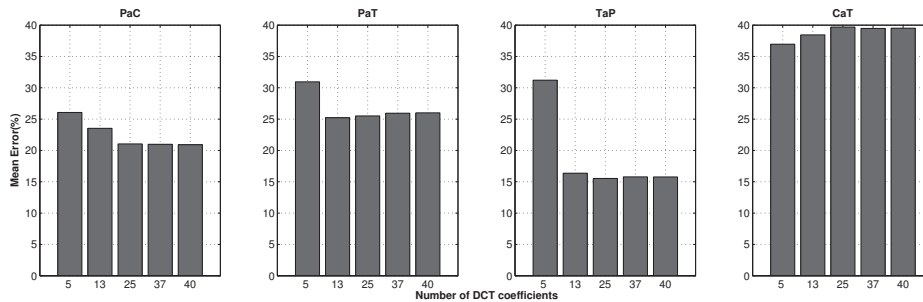


Figure 5.3: Effect of number of DCT coefficients retained from IF values on mean errors of MP-ABX tasks.

Table 5.2: Comparison of Mean Errors (%) for various features

Feature	PaC	PaT	CaT	TaP
MFCC	17.67	19.69	35.57	19.62
FDLP	21.97	18.81	23.35	33.41
IF	20.99	25.95	39.48	15.78
MFCC+FDLP	15.58	16.3	23.4	20.91
FDLP +IF	16.72	17.12	23.84	17.52
MFCC+IF	17.13	19.43	35.94	15.3

robust across different contexts irrespective of the speaker identity. On the other hand, IF features clearly outperformed the magnitude based features in TaP, which shows its stronghold in demonstrating speaker specific information in speech signals. Fusion experiments, combining pairs of features were carried out in order to explore the existence of complementary information within them. The combination of features was done by linear fusion of DTW similarity matrices. It can be seen from Table.5.2 that combination of features improved the performances in MP-ABX tasks by a remarkable margin, establishing the reciprocation of information between pairs of features. The mean errors computed here are inferior to those in [54] as the CV pairs are not isolated recordings, but obtained from continuous speech. This degradation in performance with stimuli obtained from continuous speech was already reported in [55].

5.4 Discussions

The tasks PaC and PaT represent phoneme discrimination capability of features regardless of context and speakers respectively. The CaT also denotes phoneme discrimination capability of features, in a more real world sense. On the other hand, TaP unveils speaker recognition capability of features. As it is already presented in Table.5.2, the magnitude based features show faithful performances in PaC and PaT tasks, claiming their suitability to speech recognition tasks. A closer inspection to the PaC scores from MFCC and FDLP reveals that MFCC is delivering lesser error than FDLP. As MFCC are short time spectral features, they are expected to be effective in identifying phonemes from short time frames as context will not play any significant role over short durations. But FDLP are extracted effectively from long time temporal envelopes, which can impart slight extend of context dependency into it and results in lesser performance than MFCC. In case of PaT task, both MFCC and FDLP are performing equivalently because speaker dependency will not get passed on to FDLP features as context dependency did. For CaT task, the scores attained from FDLP outperforms those from MFCC. FDLP succeeded in identifying phonemes irrespective of context and speaker, suggesting that speaker variance is more crucial than context variance in adversely affecting speech recognition. MFCC performed significantly superior to FDLP in TaP task, indicating its ability to convey relevant information for speaker recognition applications.

The IF features behaved almost complementary to FDLP in all tasks, demonstrating lesser context dependency and higher speaker dependency in phoneme recognition PaC and PaT tasks. Also, it is consistently poor in PaC, PaT and CaT tasks in comparison with MFCC. But, its speaker discrimination ability is distinctly

visible from TaP task, where it delivered remarkable performance than MFCC and FDLP. This suggest that the magnitude characteristics of speech showcase pools of linguistic information, where as its phase characteristics convey mostly speaker specific information.

To analyse the presence of complementary information in features, fusion experiments were performed. The magnitude based features from Fourier domain and analytic domain were combined and MP-ABX tasks were carried out. It was observed that the short term spectral information and long time temporal information in MFCC and FDLP respectively, enacted effectively in bringing down the mean errors from their individual scores in all tasks regarding speech recognition. But the speaker discrimination ability of FDLP features is not evident, as it failed to improve the performance in TaP task, when combined with MFCC.

The FDLP and IF features were combined together to explore the magnitude and phase characteristics of speech in complex analytic domain. The complementary nature of information in analytic magnitude and phase as exhibited by FDLP and IF features succeeded in overcoming context and speaker dependencies in PaC and PaT tasks. The inability of IF features to dwindle down the adverse effects of context and speaker variance together in phoneme recognition made the performance of fusion suffer in CaT task. Similarly, the ineffectiveness of FDLP features in speaker discrimination obstructed the fusion from performing better in TaP task. Thus combination of MFCC and FDLP yielded the best performance in phoneme discrimination tasks and IF feature performed remarkably in speaker discrimination task.

5.5 Conclusions

Analysis of features of speech extracted from magnitude and phase of the complex analytic representation was carried out using MP-ABX tasks. The frequency domain linear prediction (FDLP) coefficients and instantaneous frequency (IF) features were obtained from analytic magnitude and analytic phase respectively. Their performances with respect to phoneme and speaker discriminative MP-ABX tasks based on CV pair speech stimuli were evaluated and compared with those of conventional MFCC features. It was observed that the FDLP features are efficient in conveying linguistic information in speech signals and upon combination with MFCC delivered exceptional phoneme discrimination performances. On the other hand, IF features had a stronghold in manifesting speaker specific characteristics of speech and outperformed MFCC, FDLP and their combination in speaker discrimination task. This study suggests the importance of magnitude of speech in conveying linguistic details and phase of speech in revealing speaker characteristics.

Chapter 6

Conclusions and Future Work

In this thesis, we have presented the development of a spoken term detection system for Telugu Language. Subsequence DTW is employed to search for a query word in the reference utterance. The representation of reference and query utterances plays a crucial role during the search. In this work, three different representation techniques are investigated, namely phonetic posteriors, GMM posteriors and GBRBM posteriors. The phonetic posteriors, obtained from HMM-MLP phoneme recognizer, requires large amount of manually labelled data. On the other hand, the GMM posteriors and the GBRBM posteriors can be obtained from unlabelled speech data. It is observed that the performance of phonetic posteriors is much better than the performance of the GMM and GBRBM posteriors. However, its application is limited since it requires labelled data. Future efforts will focus on improving the unsupervised feature representation techniques, using sequence and context information. Analysis of features derived from analytic signal representation by using MP-ABX tasks is also presented. It was observed that magnitude based features FDLP and MFCC carry more linguistic information while phase based feature IF contains speaker specific characteristics.

References

- [1] J. Foote. An Overview of Audio Information Retrieval. *Multimedia Syst.* 7, (1999) 2–10.
- [2] A. J. K. Thambiratnam. Acoustic keyword spotting in speech with applications to data mining. Ph.D. thesis, Queensland University of Technology 2005.
- [3] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington. Results of the 2006 spoken term detection evaluation. In Proc. SIGIR Special Interest Group on Information Retrieval Workshop, Amsterdam, Netherlands, volume 7. 2007 51–57.
- [4] J. Tejedor, D. T. Toledano, X. Anguera, A. Varona, L. F. Hurtado, A. Miguel, and J. Cols. Query-by-Example Spoken Term Detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion. *EURASIP Journal on Audio, Speech and Music Processing* 23, (2013) 1–17.
- [5] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier. The spoken web search task at MediaEval 2012. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, British Columbia, Canada. 2013 8121–8125.
- [6] X. Anguera, L. J. Rodríguez-Fuentes, I. Szöke, A. Buzo, F. Metze, and M. Peñagarikano. Query-by-example spoken term detection on multilingual unconstrained speech. In INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore. 2014 2459–2463.
- [7] N. Alcaraz Meseguer. Speech analysis for automatic speech recognition. Master’s thesis, Norwegian University of Science and Technology 2009.
- [8] M. R. Hasan, M. Jamil, M. Rabbani, and M. Rahman. Speaker identification using Mel frequency cepstral coefficients. volume 1. 2004 565–568.
- [9] M. Weintraub. LVCSR log-likelihood ratio scoring for keyword spotting. In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95, Detroit, Michigan, USA, volume 1. 1995 297–300 vol.1.
- [10] K. Ng and V. W. Zue. Subword-based Approaches for Spoken Document Retrieval. *Speech Commun.* 32, (2000) 157–186.
- [11] R. Rose and D. Paul. A Hidden Markov Model based keyword recognition system. In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90, Albuquerque, New Mexico, USA. 1990 129–132 vol.1.
- [12] T. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU Merano/Meran, Italy. 2009 421–426.
- [13] X. Anguera. Speaker independent discriminant feature extraction for acoustic pattern-matching. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan., 2012 485–488.

- [14] Y. Zhang and J. R. Glass. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriors. In *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU Merano/Meran, Italy*. 2009 398–403.
- [15] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li. An acoustic segment modeling approach to query-by-example spoken term detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan. 2012 5157–5160.
- [16] R. R. Pappagari, S. Nayak, and K. S. R. Murty. Unsupervised spoken word retrieval using Gaussian-Bernoulli restricted Boltzmann machines. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, Singapore. 2014 1737–1741.
- [17] I. Szöke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapso, and J. Cernocký. Comparison of keyword spotting approaches for informal continuous speech. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal. 2005 633–636.
- [18] C. Chelba, T. J. Hazen, and M. Saralar. Retrieval and browsing of spoken content. *IEEE Signal Processing Mag* 25, (2008) 39–49.
- [19] M. Saraclar and R. Sproat. Lattice-Based Search for Spoken Utterance Retrieval. In *North American Chapter of the Association for Computational Linguistics*, Boston, Massachusetts. 2004 129–136.
- [20] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language* 14, (2000) 373 – 400.
- [21] D. Hakkani-Tur and G. Riccardi. A general algorithm for word graph matrix decomposition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '03)*, Hong Kong, volume 1. 2003 I-596–I-599 vol.1.
- [22] C. Chelba and A. Acero. Position specific posterior lattices for indexing speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, USA. 2005 443–450.
- [23] D. R. H. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish. Rapid and accurate spoken term detection. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium. 2007 314–317.
- [24] P. Yu, K. Chen, C. Ma, and F. Seide. Vocabulary-Independent Indexing of Spontaneous Speech. *IEEE Transactions on Speech and Audio Processing* 13, (2005) 635–643.
- [25] D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, and M. Saraclar. Effect of pronunciations on OOV queries in spoken term detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Taipei, Taiwan*. 2009 3957–3960.
- [26] T. Ge and Z. Li. Approximate substring matching over uncertain strings. *Proceedings of the VLDB Endowment* 4, (2011) 772–782.
- [27] G. Aradilla, J. Vepa, and H. Bourlard. Using Posterior-Based Features in Template Matching for Speech Recognition. In *International Conference on Spoken Language Processing*, Pittsburgh, PA, USA. 2006 .
- [28] P. Fousek and H. Hermansky. Towards ASR Based on Hierarchical Posterior-Based Keyword Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2006*, Toulouse, France, volume 1. 2006 I–I.
- [29] Y. Zhang, R. Salakhutdinov, H.-A. Chang, and J. Glass. Resource configurable spoken query detection using deep Boltzmann machines. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan. 2012 5161–5164.

- [30] W. Shen, C. M. White, and T. J. Hazen. A comparison of query-by-example methods for spoken term detection. Technical Report, DTIC Document 2009.
- [31] D. J. Berndt and J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In U. M. Fayyad and R. Uthurusamy, eds., KDD Workshop. AAAI Press, 1994 359–370.
- [32] A. Park and J. Glass. Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing* 16, (2008) 186–197.
- [33] M. Müller. Information Retrieval for Music and Motion. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.
- [34] X. Anguera, R. Macrae, and N. Oliver. Partial sequence matching using an Unbounded Dynamic Time Warping algorithm. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Sheraton Dallas Hotel, Dallas, Texas, USA. 2010 3582–3585.
- [35] M. ANGUERA. Method and system for improved pattern matching 2014. EP Patent App. EP20,120,382,508.
- [36] G. Mantena and X. Anguera. Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, Vancouver, British Columbia, Canada. 2013 8515–8519.
- [37] Y. Lin, T. Jiang, and K. Chao. Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis. *J. Comput. Syst. Sci.* 65, (2002) 570–586.
- [38] Y. Zhang, K. Adl, and J. R. Glass. Fast spoken query detection using lower-bound Dynamic Time Warping on Graphical Processing Units. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Kyoto, Japan. 2012 5173–5176.
- [39] Y. Zhang and J. R. Glass. A Piecewise Aggregate Approximation Lower-Bound Estimate for Posteriorgram-Based Dynamic Time Warping. In INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy. 2011 1909–1912.
- [40] Y. Zhang and J. R. Glass. An inner-product lower-bound estimate for dynamic time warping. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, Prague, Czech Republic. 2011 5660–5663.
- [41] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. John Wiley & Sons, 2012.
- [42] R. R. Pappagari, K. Rout, and K. S. R. Murty. Query word retrieval from continuous speech using GMM posteriorgrams. In International Conference on Signal Processing and Communications (SPCOM), 2014, Bangalore, India. 2014 1–6.
- [43] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science* 313, (2006) 504–507.
- [44] M. A. Carreira-Perpinan and G. E. Hinton. On contrastive divergence learning. In Proceedings of the tenth international workshop on artificial intelligence and statistics. 2005 33–40.
- [45] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area V2. In J. Platt, D. Koller, Y. Singer, and S. Roweis, eds., Advances in Neural Information Processing Systems 20, 873–880. Curran Associates, Inc., 2008.
- [46] M. Hochberg, S. Renals, A. Robinson, and G. Cook. Recent improvements to the ABBOT large vocabulary CSR system. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Detroit, Michigan, USA, volume 1. 1995 69–72 vol.1.

- [47] H. Bourlard and N. Morgan. Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. In *Adaptive Processing of Sequences and Data Structures*, 389–417. Springer, 1998.
- [48] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* 77, (1989) 257–286.
- [49] S. Haykin. *Neural Networks: A Comprehensive Foundation*. 2nd edition. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.
- [50] K. Rout, R. R. Pappagari, and S. R. M. Kodukula. Experimental Studies on Effect of Speaking Mode on Spoken Term Detection. In *National Conference on Communications 2015 (NCC-2015)*. Mumbai, India, 2015 .
- [51] J. E. Cutting. Different speech-processing mechanisms can be reflected in the results of discrimination and dichotic listening tasks. *Brain and Language* 1, (1974) 363–373.
- [52] L. Nathan and B. Wells. Can children with speech difficulties process an unfamiliar accent? *Applied Psycholinguistics* 22, (2001) 343–361.
- [53] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky. Rapid Evaluation of Speech Representations for Spoken Term Discovery. In *INTERSPEECH 2011*. 2011 821–824.
- [54] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux. Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013*. Lyon, France, 2013 1–5.
- [55] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux. Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise. In *INTERSPEECH 2014* .
- [56] L. Cohen. *Time-frequency analysis: theory and applications*. Signal processing series. Prentice Hall, Inc., Upper Saddle River, NJ, USA., 1995.
- [57] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck. *Discrete-time signal processing*. Signal processing series, 2nd edition. Prentice Hall, Inc., Upper Saddle River, NJ, USA., 1999.
- [58] P. Maragos, J. Kaiser, and T. Quatieri. Energy separation in signal modulations with application to speech analysis. *Signal Processing, IEEE Transactions on* 41, (1993) 3024–3051.
- [59] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *The Journal of the Acoustical Society of America* 99.
- [60] A. Rao and R. Kumaresan. On decomposing speech into modulated components. *Speech and Audio Processing, IEEE Transactions on* 8, (2000) 240–254.
- [61] D. Dimitriadis, P. Maragos, and A. Potamianos. Robust AM-FM features for speech recognition. In *IEEE Signal Processing Letters*, volume 12. 2005 621–624.
- [62] H. Yin, V. Hohmann, and C. Nadeu. Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency. In *Speech communication*, volume 53. 2011 707–715.
- [63] M. Grimaldi and F. Cummins. Speaker identification using instantaneous frequencies. In *IEEE Transactions on audio speech and language processing*, volume 16. 2008 1097–1111.
- [64] T. Thiruvaran, E. Ambikairajah, and J. Epps. Speaker identification using FM features. In *11th Australian International Conference on Speech Science and Technology*. 2006 148–152.
- [65] K. Vijayan, V. Kumar, and K. S. R. Murty. Feature Extraction from Analytic Phase of Speech Signals for Speaker Verification. In *INTERSPEECH 2014*. 2014 1658–1662.

- [66] M. Athineos and D. P. Ellis. Frequency-domain linear prediction for temporal features. In IEEE Workshop on Automatic Speech Recognition and Understanding. 2003 261–266.
- [67] M. Athineos and D. P. W. Ellis. Autoregressive modeling of temporal envelopes. In IEEE transactions on signal processing, volume 55. 2007 5237–5245.
- [68] S. Ganapathy and H. Hermansky. Temporal resolution analysis in frequency domain linear prediction. In J. Acoust. Soc. Am, volume 132. 2012 .
- [69] S. Thomas, S. Ganapathy, and H. Hermansky. Recognition of reverberant speech using frequency domain linear prediction. In IEEE Signal Processing Letters, volume 15. 2008 681–684.