EXAMPLE-SPECIFIC DENSITY BASED MATCHING KERNELS FOR VARYING LENGTH PATTERNS OF SPEECH AND IMAGES

A THESIS

submitted by

ABHIJEET SACHDEV
(S13004)

for the award of the degree of

Master of Science
(By Research)

School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
India - 175001

THESIS CERTIFICATE

This is to certify that the thesis titled **EXAMPLE-SPECIFIC DENSITY BASED MATCHING KERNELS FOR VARYING LENGTH PATTERNS OF SPEECH AND IMAGES**, submitted by **Abhijeet Sachdev**, to the Indian Institute of Technology, Mandi, for the award of the degree of **Master of Science (By Research)**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Date
Mandi, 175001

Dr. Dileep A.D.
Guide

Dedication

I dedicate my dissertation work to my family. A special feeling of gratitude to my loving parents whose words of encouragement motivated me always. My sister Ajeeta, has never left my side and is very special. I also dedicate this dissertation to my many friends and family who have supported me throughout the process.

I dedicate this work to my parents for being with me always in every situation of life.

Declaration

I hereby declare that this submission is my own work and to the best of my knowledge and belief, it contains no material previously published or written by another person for award of any degree in any University, except where due acknowledgment has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree in any university. This thesis is a presentation of my original research work.

Candidates name and signature

Acknowledgments

First and foremost, praises and thanks to the God, for blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my advisor Dr. Dileep A.D. for his continuous support in research, also for his great help, patience and motivation. His guidance helped me in all the time of research and writing of this thesis. I cannot imagine having a better advisor than him for my MS study.

Besides my advisor, I would like to thank Prof. C. Chandra Sekhar from IIT Madras for his always ready to help behavior and immense knowledge I have taken from him.

I would also like to say special thanks to Dr. Veena Thenkanidiyoor from NIT Goa for her guidance and support in my research work.

I am extremely thankful to Prof. Timothy A. Gonsalves, Director of IIT Mandi, for providing excellent research environment.

I would also like to give special thanks to Dr. Arnav Bhavsar for his always support and guidance throughout the course of my research.

I am thankful to Dr. Anil K Sao, Chairperson of School of Computing & electrical engineering for providing excellent facilities for conducting research work.

I am also thankful to the rest of my thesis committee: Dr. Padmanabhan Rajan and Dr. Pratibha Garg for their encouragement and insightful comments.

I also thank to Dr. Sarita Azad and Dr. Om Prakash Singh for their support in research work.

I would also like to say special thanks to Miss Shikha Gupta, my colleague, for her moral support and suggestions without which this journey would have been more difficult.

Last but not the least, I would like to thank my parents for their immense belief on me, support and patience.

Abhijeet Sachdev

Abstract

In thesis, we address some issues in classification of varying length patterns of speech and scene images represented as sets of continuous valued feature vectors using kernel methods. Kernels designed for varying length patterns are called as dynamic kernels. This thesis considers the matching based approaches for designing dynamic kernels.

The thesis first proposes the example-specific density based matching kernel (ESDMK) based support vector machine (SVM) classifier for varying length patterns. The proposed kernel is computed between a pair of examples, represented as sets of feature vectors, by matching the estimates of the example-specific densities computed at every feature vector in those two examples. The number of feature vectors of an example among the k nearest neighbors of a feature vector is considered as an estimate of the example-specific density. The minimum of the estimates of two example-specific densities, one for each example, at a feature vector is considered as the matching score. The ESDMK is then computed as the sum of the matching score computed at every feature vector in a pair of examples. The main issue in building the proposed kernel is choice of k, the number of neighbors. This thesis proposes to combine all the matchings obtained using the different values of k to compute pyramid match ESDMK. We propose to compute pyramid match ESDMK as the weighted sum of matches obtained by computing the ESDMKs at sequence of increasingly coarser neighbors. The proposed ESDMKs does not include spatial information in the images which is important for better matching of images. We propose the spatial ESDMK (SESDMK) to include the spatial information. We consider a fixed number of spatial regions in every scene image. An ESDMK for the local feature vectors in a particular region from the two examples is constructed. Then, the SESDMK is constructed as a combination of ESDMKs of all the

regions. The performance of the SVM-based classifiers using the proposed family of ESDMKs for sets of local feature vectors extracted from images and long duration speech is studied for scene classification, speech emotion recognition and speaker identification tasks and compared with that of the SVM-based classifiers using the state-of-the-art dynamic kernels.

Keywords: *Varying length patterns, scene images, long duration speech, set of local feature vectors, support vector machine, dynamic kernels, example-specific density based matching kernel, scene classification, speech emotion recognition, speaker identification*

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

An important component in the classification task using support vector machines (SVM) is the computation of inner product operation [1]. When classes are non-linearly separable inner product operation is typically computed in high dimensional space. Evaluation of inner product in a high dimensional space is avoided by using an inner product kernel, $K(\mathbf{x}_m, \mathbf{x}_n)$, defined as $K(\mathbf{x}_m, \mathbf{x}_n) = \mathbf{\Phi}(\mathbf{x}_m)^T \mathbf{\Phi}(\mathbf{x}_n)$ [2]. Here $\mathbf{\Phi}(\mathbf{x}_m)$ and $\mathbf{\Phi}(\mathbf{x}_n)$ are the high dimensional feature vectors representation of the pair of examples $\mathbf{x}_m$ and $\mathbf{x}_n$ respectively. The commonly used kernels such as Gaussian kernel, polynomial kernel etc., are computed between the two examples represented in fixed dimensional space. Such kernels are called static kernels. However, the data such as images in which features are extracted from local homogeneous regions and speech are represented as varying length sets of local feature vectors. The focus of this work is on classification of varying length patterns of images and speech. The kernels computed for varying length patterns are called dynamic kernels [3]. This work focus on designing dynamic kernels for varying length patterns.

1.1 Scene images represented as sets of local feature vectors

A scene image can be defined as a semantically coherent human-scaled view of a real-world environment [4]. Scene image's content can be represented at several levels depending upon the kind of information being considered. For example the lowest level comprises of low level primitives such as color and

texture. The next level consists of semantic cues such as buildings, roads, streets etc. Our aim is to identify the class to which the given image belongs and to achieve this we need features in the image. The most important characteristic we require is that there must be a strong correlation between feature values and the corresponding class of the image. Once, the features are identified, a classifier can be constructed using classifiers like SVM. Low level features can be obtained from the complete image as a whole or from individual local regions of the image. The low-level visual features extracted by processing all the pixels in an image are used to obtain a d-dimensional global feature vector. Semantic variability of a scene image may not be well represented by a global feature vector. Thus it is necessary to go for local feature vectors to capture local semantic information in a better way. There exist two approaches to extract local feature vectors. In the first approach, image is divided into fixed sized blocks and then features are extracted as shown in Figure 1.1. While in the second approach, interest points are detected first, and then around the interest point, fixed size block is considered and feature vector is extracted as illustrated in Figure 1.2.

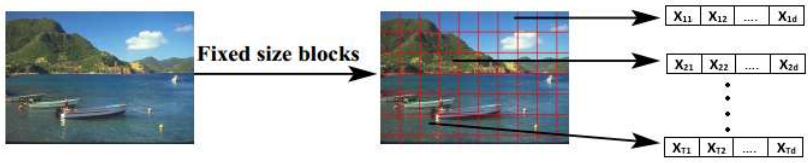


Figure 1.1: Illustration of representing an image by local feature vectors extracted from fixed size non overlapping blocks [4].

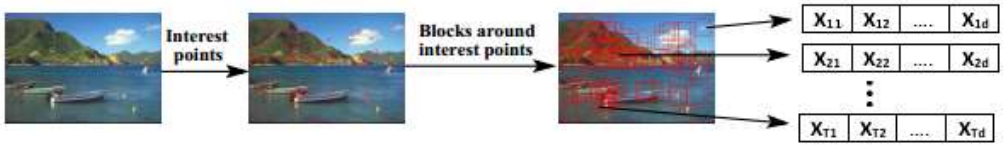


Figure 1.2: Illustration of representing an image by local feature vectors extracted from fixed size blocks that are placed around interest points [4].

Let $\mathbf{X}$ be the scene image and $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_T\}$ be the local feature vectors extracted from the image. The scene image is then represented as a set of

local feature vectors as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 ... \mathbf{x}_T\}$. When the images are of different sizes, the number of fixed size non-overlapping blocks in them will also be different resulting in different number of local feature vectors extracted from both the images. Such sets of different number of local feature vectors are called varying length patterns of images.

1.2 Speech signals represented as sets of local feature vectors

Short-time analysis of a speech signal involves performing spectral analysis on frames of about 20 milliseconds duration each and representing a frame by a real valued local feature vector. Acoustic modeling of subword units of speech such as phonemes, triphones and syllables, involves developing classification models for patterns extracted from speech segments of subword units. Even when two signals belong to utterance of the same class, their durations can be different and hence the number of local feature vectors obtained after short time analysis is different. Duration of speech signal of a subword unit is short and it is necessary to model the temporal dynamics and correlations among the features while developing the classification models for subword units. In such cases, it is necessary to represent a speech signal as a sequence of local feature vectors. The speech signal of an utterance with T frames is represented as a sequential pattern $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 ... \mathbf{x}_T\}$, where $\mathbf{x}_t \in R^d$ is the local feature vector for frame t.

In the tasks such as text-independent speaker recognition, spoken language identification and speech emotion recognition, a phrase or a sentence is used as a unit. The duration of an utterance of a phrase or a sentence is long. Preserving sequence information in a phrase or a sentence is not considered to be critical for these tasks. The phonetic content in the speech signal is not considered to be important for these tasks, and preserving the sequence information is not critical. In such cases, the speech signal of an utterance can be represented as a set of local feature vectors, without preserving the sequence information. Let $\mathbf{X}$ be the speech signal and $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_T\}$ be the d-dimensional local feature vectors extracted from different frames of long durational speech signal as shown in Figure 1.3. The long durational speech sig-

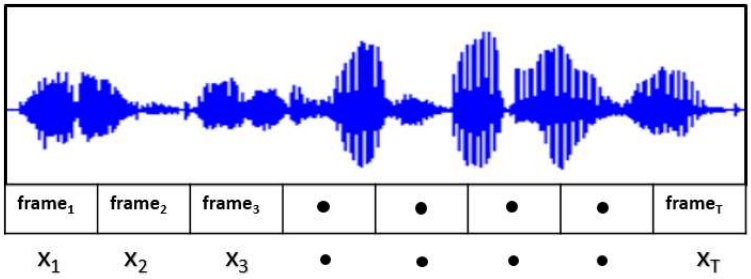


Figure 1.3: Illustration of representing an speech signal as a set of local feature vectors.

nal is then represented as a set of local feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 ... \mathbf{x}_T\}$. When the duration of two speech signals are different, the number of local features vectors extracted from them will also be different resulting in sets of varying cardinality. Such sets of different number of local feature vectors are called varying length patterns of speech. In this research work we focus only on the classification of long duration speech signals represented as varying length sets of local feature vectors.

1.3 Classification of varying length patterns of speech and images

Gaussian mixture models (GMMs) [5] are commonly used for classification of varying length patterns represented as sets of local feature vectors. The maximum likelihood (ML) based method is commonly used for estimation of parameters of a GMM for each class. The ML based method gives robust estimates of parameters only when sufficient training data is available. When the training data available for a class is limited, robust estimates of parameters can be obtained through the maximum a posteriori (MAP) adaptation of a class-independent GMM (CIGMM), which is also called as universal background model (UBM), to the training data of each class [6]. The CIGMM is a large size GMM built using the training data of all the classes. The ML method and the MAP adaptation method are non-discriminative training based methods because the estimation of parameters for each class is done independently. The discriminative training based large margin method has

been proposed for estimation of parameters of GMMs [7]. In this method, the parameters of the GMMs of all the classes are estimated simultaneously by solving an optimization problem to maximize the distance of training examples to the boundaries of the classes.

The discriminative model based approaches to classification of varying length patterns of speech and images represented as sets of local feature vectors includes the SVM based approaches [1]. There are two approaches to varying length pattern classification using SVMs depending on the type of kernel used. In the first approach, a varying length pattern is first mapped onto a fixed length pattern and then a kernel for fixed length patterns is used to build the SVM [8, 9]. In the second approach, a suitable kernel for varying length patterns is designed. The focus of this research work is to use dynamic kernel based SVMs for classification of varying length patterns of speech and images represented as sets of local feature vectors. Different approaches to design the dynamic kernels are as follows: (1) Explicit mapping based approaches [10–12] , where a set of local feature vectors is mapped onto a fixed dimensional representation and a kernel function is defined in the space of that representation. (2) Probabilistic distance metric based approaches [12, 13] , where a suitable distance measure for two sets of local feature vectors is kernelized. (3) Matching based approaches [14, 15] where a kernel function is defined by matching the local feature vectors in a pair of examples. In this research work, we address the issues in designing the dynamic kernels using the matching based approaches. We propose to build a family of example-specific density based matching kernels (ESDMK) based support vector machine (SVM) classifiers for varying length patterns of speech and images represented as sets of local feature vectors.

1.4 Objectives and scope of the work

The objective of this research work is to address the issues in designing and demonstrating ESDMK-based SVM classifiers for the classification of varying length patterns of speech and images represented as sets of local feature vectors.

In this research work, the proposed ESDMK is constructed between a pair of examples represented as sets of local feature vectors by matching

the estimates of example-specific densities computed at every feature vector in the two examples. For every feature vector in the pair of examples an example-specific density is obtained by computing its k nearest neighbors.

The main issue in obtaining a better ESDMK is the choice of k, the number of neighbors. This issue is addressed by combining all the matches obtained using different values of k using pyramid match principle.

The ESDMK designed using both the approaches does not include the spatial information in the scene images which is important for better matching of images. This issue is addressed by considering a fixed number of spatial regions in every scene image and constructing ESDMK for the local feature vectors in a particular region from the two examples.

The effectiveness of ESDMK-based SVMs is demonstrated using the studies on speech emotion recognition, speaker identification and scene classification tasks.

1.5 Organization of the thesis

The thesis is organized as follows:

In **Chapter 2** we provide an introduction to generative and discriminative approaches for classification of sets of local feature vectors.

In **Chapter 3** we introduce dynamic kernels. Commonly used dynamic kernels for varying length patterns of speech and images represented as sets of local feature vectors are presented.

In **Chapter 4** we propose the example-specific density based matching kernel (ESDMK) for sets of local feature vectors. We show that the capabilities of ESDMK in capturing the local information is better than the other dynamic kernels. The effectiveness of the SVM-based classifiers using the proposed ESDMK is studied for scene image classification, speech emotion recognition and speaker identification.

In **Chapter 5** we propose pyramid example-specific density based matching kernel (PESDMK). An ESDMK essentially matches the two sets of local feature vectors for a particular value of k. We propose to use pyramid match principle to enhance the matching ability of ESDMK by considering the increasing values of k.

In **Chapter 6** we propose spatial example-specific density based matching kernel (SESDMK) and pyramid spatial example-specific density based matching kernel (PSESDMK). Scene images carry a lot of spatial clues that can be used in the computation of similarity between a pair of scene images. Thus for scene image classification, spatial ESDMKs are proposed to incorporate spatial information present in the scene images while computing ESDMKs.

In **Chapter 7** we summarize the contributions of the present work. We also present some directions for further work.

Chapter 2

Approaches for Classification of Sets of Feature Vectors

Modern schemes for classification tasks generally fall into one of the two broad approaches, generative or discriminative. Both of these approaches are described in this chapter. In the first approach, generative models are built for each class independently and then Bayes' decision rule is applied to classify the example. This chapter introduces the Gaussian mixture model (GMM) as the generative model based approach for classification of sets of local feature vectors. Discriminative schemes are an alternative approach for classification tasks. Unlike generative approaches, these attempt to model either the class boundaries or posterior probabilities directly. Discriminative model based approach for classification of sets of local feature vectors described in this chapter includes support vector machines (SVM).

2.1 Generative model based approaches for classification of sets of local feature vectors

Generative model based approaches for classification of sets of local feature vectors uses the Bayes' classifier. In generative model based approaches for classification, underlying probability distribution function is built for each of the semantic class. The Bayes' decision rule is then applied to assign the

final label to the example. Let C be the total number of classes and $\mathbf{X}$ be an example represented as sets of local feature vectors, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_T\}$. The likelihood of $\mathbf{X}$ for class c, $p(\boldsymbol{X}|c)$, is used to compute the corresponding posterior probability using the Bayes' rule as follows:

$$y = \underset{c}{\operatorname{argmax}}\, p(c|\boldsymbol{X}) = \underset{c}{\operatorname{argmax}}\, \frac{p(\boldsymbol{X}|c)P(c)}{p(\boldsymbol{X})} \tag{2.1}$$

Here, $P(c)$ is the prior probability of class c and $p(\boldsymbol{X}|c)$ is the likelihood of example $\boldsymbol{X}$ for class c, $p(c|\boldsymbol{X})$ is the posterior probability of the example $\boldsymbol{X}$ belonging to semantic class c. The class conditional density is generally computed using a suitable model for the probability distribution of data. The main issue in the generative model based approaches is the choice of a model for the probability distribution. One of the most popular forms of distribution for modeling the underlying distribution is the Gaussian mixture model (GMM).

2.1.1 GMM-based classifier for sets of local feature vectors

Gaussian mixture models (GMMs) are the commonly used generative models for classification. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_{T_c}\}$ be the set of local feature vectors for an example $\mathbf{X}$ in the training dataset of class c, where T_c is the number of local feature vectors in $\mathbf{X}$. The likelihood of a local feature vector $\mathbf{x}_t$ being generated by the GMM for class c, λ_c, with K Gaussian components is given by

$$p(\mathbf{x}_t; \lambda_c) = \sum_{k=1}^{K} \pi_{ck} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck}) \tag{2.2}$$

where π_{ck} are the mixture weight and must satisfy

$$0 \leq \pi_{ck} \leq 1$$

and

$$\sum_{k=1}^{K} \pi_{ck} = 1$$

$\boldsymbol{\mu}_{ck}$ and $\boldsymbol{\Sigma}_{ck}$ are respectively the mixture weight, mean vector and covariance

matrix of the kth component of λ_c. Here $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck})$ is the multivariate Gaussian distribution for the kth component of λ_c given by

$$\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_{ck}|}} exp(\frac{-1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{ck})^T \boldsymbol{\Sigma}_{ck}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{ck})) \quad (2.3)$$

Training a GMM for class c involves the estimation of parameters, $\lambda_c = \{\pi_{ck}, \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck}\}$, $k = 1, 2, ..., K$, using the training data set of class c. Maximum likelihood (ML) method is commonly used for estimation of parameters of a GMM. The ML method estimates the parameters of a GMM such that the total likelihood is maximized. The log likelihood of an example, represented by a set of local feature vectors, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ...\mathbf{y}_T\}$, for λ_c is given by

$$\ln p(\mathbf{Y}|\lambda_c) = \sum_{t=1}^{T} \ln \sum_{k=1}^{K} \pi_{ck} \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck}) \quad (2.4)$$

Then a class label y for $\mathbf{Y}$ is assigned using the following decision rule

$$y = \underset{c}{\text{argmax}} \ \ln p(\mathbf{Y}|\lambda_c) + \ln P(c) \quad (2.5)$$

A GMM-based classifier used for speaker identification in [16] uses speech utterances from 49 speaker classes which is a subset of speakers in the KING speech database [17]. A frame size of 20 ms and a frame shift of 10 ms are used for feature extraction from the speech signal of an utterance. Every frame is represented using a 12-dimensional feature vector of cepstral mean normalized [18] Mel frequency cepstral coefficients (MFCC). A 50-component GMM is considered for each speaker. The GMM-based classifier achieved 94.5% identification accuracy. This performance is compared with that obtained using the vector quantization (VQ) based method using 100 codevectors and a radial basis function (RBF) neural network model with 800 basis functions. It is shown that speaker identification using the GMM-based classifier attained a better accuracy compared to the classifier that uses the VQ-based method and a RBF model. In [19], a GMM-based classifier is used for speech emotion recognition using the emotional speech database for Basque recorded by the University of the Basque. This database contains speech utterances belonging to six emotional categories, namely, anger, fear,

surprise, disgust, joy and sadness. A 39-dimensional MFCC feature vector is extracted from a frame of 25 ms with a shift of 10 ms from the speech signal of an utterance. A speech emotion recognition accuracy of 98.4% is achieved using a GMM-based classifier that uses a 512-component GMM to model an emotion class.

One of the limitations of maximum likelihood (ML) method for GMM parameter estimation is that, it yields robust estimates of the parameters only when sufficient number of examples are available in training data. When only a limited amount of training data is available for a class, robust estimates of parameters can be obtained through maximum a posteriori (MAP) adaptation of a class-independent GMM (CIGMM) to the training data of each class [5, 12]. The adaptation provides a tight coupling between a class model and CIGMM. The CIGMM, also called as universal background model (UBM), is a large GMM built using the training data of all the classes. A class-dependent GMM is obtained by adapting the UBM to the data of a class. The MAP adaptation method is commonly used for adaptation. The adaptation is carried out using the expectation maximization (EM) method. The first step of the EM method estimates the sufficient statistics for each component in the UBM such as mixture weight, mean, and variance using the training data of a class. In the second step, these new estimates of sufficient statistics are combined with the UBM parameters using a data-dependent mixing coefficient. The data-dependent mixing coefficient is designed so that the components with high counts of data from the class rely more on the new sufficient statistics for final parameter estimation, and the components with low counts of data from the class rely more on the old sufficient statistics for final parameter estimation. Given a UBM and the set of local feature vectors extracted from all the examples in the training data of a class c, $\mathbf{D}_c = \{\mathbf{x}_{c1}, \mathbf{x}_{c2}, ..., \mathbf{x}_{cT_c}\}$, the probability of $\mathbf{x}_{ct}$ being generated by the component k of UBM, $\gamma_k(\mathbf{x}_{ct})$, is computed as

$$\gamma_k(\mathbf{x}_{ct}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_{ct}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_{ct}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{2.6}$$

where π_k is the mixture coefficient of the component k, and $\mathcal{N}(\mathbf{x}_{ct}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the normal density for the component k with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The effective number of feature vectors belonging to the

component k is given as

$$T_{ck} = \sum_{t=1}^{T_c} \gamma_k(\mathbf{x}_{ct}) \tag{2.7}$$

The weighted mean of the examples belonging to the kth component is obtained as

$$E_k(\mathbf{x}_{ct}) = \frac{1}{T_{ck}} \sum_{t=1}^{T_c} \gamma_k(\mathbf{x}_{ct}) \mathbf{x}_{ct} \tag{2.8}$$

The variance of the effective feature vectors belonging to the kth component is obtained as

$$E_k(\mathbf{x}_{ct}^2) = \frac{1}{T_{ck}} \sum_{t=1}^{T_c} \gamma_k(\mathbf{x}_{ct}) \mathbf{x}_{ct}^2 \tag{2.9}$$

where $\mathbf{x}_{ct}^2$ is the shorthand notation for $\text{diag}(\mathbf{x}_{ct}\mathbf{x}_{ct}^T)$ [5].

These new sufficient statistics obtained from the training data are used to update the old UBM sufficient statistics for the component k and obtain the adapted parameters for the component k as follows:

$$\hat{w}_{ck} = [\beta_k^{(w)} T_{ck}/T_c + (1 - \beta_k^w)\pi_k]\tau \tag{2.10}$$

$$\hat{\boldsymbol{\mu}}_{ck} = [\beta_k^{(\boldsymbol{\mu})} E_q(\mathbf{x}_{ct})/T_c + (1 - \beta_k^{\boldsymbol{\mu}})]\boldsymbol{\mu}_k \tag{2.11}$$

$$\hat{\boldsymbol{\sigma}}_{ck}^2 = \beta_k^{(\boldsymbol{\nu})} E_q(\mathbf{x}_{ct}^2)/T_c + (1 - \beta_k^{\boldsymbol{\nu}})(\boldsymbol{\sigma}_k^2 + \boldsymbol{\mu}_k^2) - \hat{\boldsymbol{\mu}}_k^2 \tag{2.12}$$

The adaptation coefficients $\{\beta_{ck}^{(\pi)}, \beta_{ck}^{(\boldsymbol{\mu})}, \beta_{ck}^{(\boldsymbol{\nu})}\}$ control the balance between the old and new estimates. The scale factor, τ, in Equation 2.10 is used to ensure that the sum of all the mixture weights is unity. The data-dependent adaptation coefficient β_k^{ρ}, $\rho \in \{\pi, \mu, \nu\}$ used in above equations is defined as

$$\beta_{ck}^{(\rho)} = \frac{T_{ck}}{T_{ck} + r^{(\rho)}} \tag{2.13}$$

where $r_(\rho)$ is a fixed relevance factor for parameter ρ.

This GMM-UBM system is the state-of-the-art system for classification task. In [6], a 2048-component UBM was built using the 1998 NIST SRE development data and is adapted to each of the 640 speaker classes in the training data from 1999 NIST SRE corpus. The resulting GMM-UBM system was evaluated for speaker verification task using the test data from 1999 NIST SRE corpus. It is also shown in [6] that only mean adaptation is sufficient to achieve good speaker recognition performance when the available data is less.

The ML method and the adaptation methods are non-discriminative training based methods because the estimation of parameters for each class is done independently. Recently, discriminative training based large margin (LM) method has been proposed for estimation of parameters of GMMs [7, 20]. In this method, the parameters of GMMs of all the classes are estimated simultaneously by solving an optimization problem to maximize the distance of training examples to the boundaries of the classes. In [21] GMM-based Bayesian classification is performed for image classification using the large margin based method for estimation of parameters of GMM. The LMGMM-based Bayesian classifier is built by refining the parameters of the EMGMM for each class using the large margin technique. In [20], a large margin GMM is used for speaker recognition task. From every frame of size 20 ms with the shift of 10 ms, 50 linear frequency cepstral coefficients (LFCC) were extracted. A 256-component UBM is built using the local feature vectors of all the utterances from the 2004 NIST SRE corpus. Then a MAP adapted GMM is built for each of the 50 male speakers belonging to the 2006 NIST SRE corpus. The large margin GMM-based system is built by refining the parameters of the adapted GMM for each speaker class. It was shown that the large margin based GMM classifier achieved a speaker identification accuracy of 77.6% which was better than the accuracy of the MAP adapted GMM-based system (73.3%).

2.2 Discriminative model based approaches for classification of sets of local feature vectors

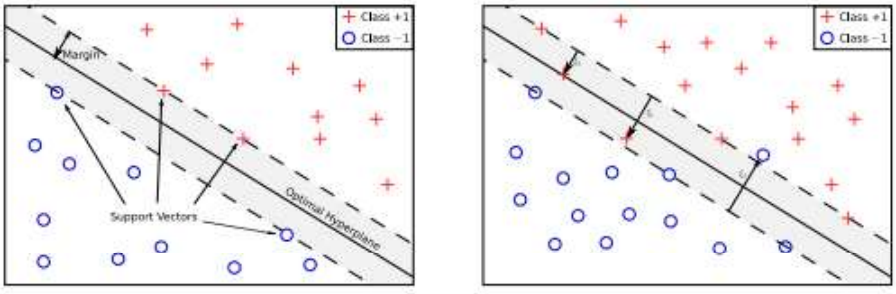
An alternative to the generative model based approaches for classification is the discriminative model based approaches. The discriminative model based approaches build the class boundaries by directly discriminating the data of each class from the data of the remaining classes. The discriminative model based approach used in this thesis is the support vector machine (SVM) for classification of sets of local feature vectors.

2.2.1 Support vector machine for classification of sets of local feature vectors

The support vector machine (SVM) [1] is a binary discriminative classifier that has been found to yield good performance on a wide range of machine learning tasks. SVMs are distance based classifiers that operate by finding a linear decision boundary according to a maximum-margin criterion. Consider a training dataset, $\mathbf{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, where each $\mathbf{x}_i$ is a vector of d elements and has an associated binary label $y_i = \omega$ where $\omega \in \{+1, -1\}$. When the training set is linearly separable, it is possible to locate a separating hyperplane within this space such that all training examples are correctly classified. This hyperplane is defined by weight vector $\mathbf{w}$ and bias b, a test example $\mathbf{x}$ may be classified according to the Equation 2.14

$$y^x = sign(\mathbf{w}^T\mathbf{x} + b) \tag{2.14}$$

Equation 2.14 is invariant under a positive rescaling of the hyperplane parameters. Thus, in order to obtain a unique solution it is necessary to introduce additional constraints. For SVMs this is achieved by defining canonical hyperplanes on either side of the decision hyperplane. For a fixed value of $\mathbf{w}$ and b these are defined by those $\mathbf{x} \in \mathbf{D}$ that form solutions to $\mathbf{w}^T\mathbf{x} + b = +1$ and those $\mathbf{x} \in \mathbf{D}$ that form solutions to $\mathbf{w}^T\mathbf{x} + b = -1$. Training examples are then constrained to lie outside this region. This arrangement is depicted in Figure 2.1(a) for two-dimensional data. Under these conditions the size of the margin can be calculated using the following expression



(a) hard margin (b) soft margin

Figure 2.1: Optimal SVM hyperplane for (a) linearly separable data using a hard margin and (b) non-linearly separable data using a soft margin with slack variables.

$$Margin = \frac{1}{\mathbf{w}^T\mathbf{w}} \tag{2.15}$$

The maximum-margin decision boundary is therefore defined by the parameters $\mathbf{w}$ and b that maximize equation 2.15 such that all training examples lie outside the margin. This yields the following quadratic optimization problem, known as the (hard margin) primal SVM problem.

$$\begin{aligned} \underset{\mathbf{w},\mathbf{b}}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1 \forall i \end{aligned} \tag{2.16}$$

In many situations, particularly when dealing with noisy data, it is not possible to linearly separate the training set. To allow SVMs to be trained in such conditions, the margin constraints, $y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1 \forall i$, are often relaxed to allow some training examples to be misclassified. A slack variable ξ_i is introduced for each training example $\mathbf{x}_i$ to provide a measure of the training error associated with the example. For each training example $\mathbf{x}_i$, the slack variable ξ_i is non-negative, and is equal to the distance by which $\mathbf{x}_i$ violates the original margin constraints. For correctly-classified training examples, $\xi_i = 0$. This is known as the soft margin case and is depicted in Figure 2.1(b). To avoid increasing the margin at the expense of misclassifying the training examples, the objective function is then altered to additionally