CHAPTER 3

LITERATURE SURVEY

3.1 IMPORTANCE OF DISCRIMINATIVE APPROACH

Gaussian Mixture Modeling (GMM) and Hidden Markov Modeling (HMM) techniques have been successful in classification tasks. Maximum Likelihood Estimation and Expectation Maximization algorithm can be used to estimate the model parameters efficiently. However, a major drawback in this type of modeling technique is that the modeling is carried out in isolation i.e., the modeling technique, when modeling a class, does not consider information from other classes. In other words, out of class data is not used to adjust the model parameters that may lead to poorer performance of the classifier. This may increase the classification (or confusion) error. Further, in conventional GMM-based classifiers, the performance is, to a greater extent, directly proportional to the duration of the test utterances, which is another major drawback. Better classification accuracy can be achieved if the training technique is able to capture the unique features of a class, i.e., the features that discriminate a class from other classes, efficiently.

Many research works have been reported in the literature to increase the classification accuracy of a classifier by increasing the discriminative power of the classifier. Such techniques can be grouped into mainly two classes as follows:
1. Discriminating the classes in the feature level itself by identifying and removing the common features between two classes under consideration.

2. Adjusting the model parameters themselves such that two classes, in the feature space itself, are well separated.

3.2 LITERATURE SURVEY

3.2.1 Baseline systems used

In Reynolds and Rose (1995), Gaussian mixture model is introduced and evaluated for text-independent speaker identification. The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability to Gaussian mixtures to model arbitrary densities. The Gaussian mixture model is experimentally evaluated on a 49 speaker conversational speech database containing both clean and telephone speech. The experiments examine algorithmic issues such as model initialization, variance limiting, and model order selection. To compensate for spectral variability introduced by the telephone channel and handsets, robustness techniques such as long-term mean removal, difference co-efficients, and frequency warping are applied and compared. The experiments also examine the GMM speaker identification performance with respect to an increasing speaker population, and comparisons to other modeling techniques (uni-modal Gaussian model, vector quantization code book model, tied Gaussian mixture model, and radial basis function).

In Reynolds (1995a, Reynolds (1995b), high performance speaker identification and verification systems are presented based on Gaussian mixture speaker models. The identification system is a maximum likelihood
classifier and the verification system is a likelihood ratio hypothesis tester using background speaker normalization. The systems are evaluated on four publicly available speech databases: TIMIT, NTIMIT, Switchboard and YOHO. The different levels of degradations and variabilities found in these databases allow the examination of system performance for different task domains. Constraints on the speech range from vocabulary-dependent to extemporaneous and speech quality varies from (near-ideal), clean speech to (noisy), telephone speech.

The use of GMM for speaker identification was shown to provide good performance with several existing techniques. However, this criterion only utilizes the labeled utterances for each speaker model and very likely leads to a local optimization solution. Classification accuracy of any classification task can be increased by using discriminative training approaches. The discriminative algorithms used in the literature for speaker or speech recognition are described in the Sections 3.2.2 and 3.2.3.

3.2.2 Model-level discrimination

To improve the discriminative qualities of Gaussian mixture models, several approaches have been proposed. Universal Background Model-Gaussian Mixture Model (UBM-GMM) is a popular one among them. UBM is a base model from which all speaker models are adapted by a form of Bayesian adaptation. In Reynolds et al (2000), the GMM-UBM system is built around the optimal likelihood ratio test for detection, using simple but effective Gaussian mixture models for likelihood functions, a universal background model for representing the competing alternative speakers, and a form of Bayesian adaptation to derive hypothesized speaker models. The use of a handset detector and score normalization to greatly improve detection performance, independent of the actual detection system, was also described and discussed. Finally, representative performance benchmarks and system
behavior experiments on the 1998 summer-development and 1999 NIST SRE corpora are presented.

In Del Alamo et al (1996), a novel discriminative training procedure for a Gaussian Mixture Model (GMM) Speaker Identification System is described. The proposal is based on the segmental Generalized Probabilistic Descent (GPD) algorithm formulated to estimate the GMM parameters. Two major innovations over similar formulations of segmental GPD training are proposed. The first one is that, a misclassification measure based on an individual representation of competing speakers, that explicitly allows to take into account different learning strategies for correctly or incorrectly classified speakers. The second one is, an empirical loss function to control the training procedure convergence, with a likelihood-based selection of correctly or incorrectly classified competing speakers. A comparison between the proposed method and the traditional GPD algorithm is also presented.

In Bahl et al (1986), A method for estimating the parameters of hidden Markov models of speech recognition is described. Parameter values are chosen to maximize the mutual information between an acoustic observation sequence and the corresponding word sequence. Recognition results of the proposed Maximum Mutual Information (MMIE) based method is compared with maximum likelihood estimation method.

In Markov et al (2001), the Maximum Normalized Likelihood Estimation (MNLE) algorithm and its application for discriminative training of HMMs for continuous speech recognition is presented. The objective of this algorithm is to maximize the normalized frame likelihood of training data. Instead of gradient descent techniques usually applied for objective function optimization in other discriminative algorithms such as Minimum Classification Error (MCE) and Maximum Mutual Information (MMI),
Markov et al (2001), used a modified Expectation-Maximization (EM) algorithm which greatly simplifies and speeds up the training procedure. Evaluation experiments showed better recognition rates compared to both the Maximum Likelihood (ML) training method and MCE discriminative method. In addition, the MNLE algorithm showed better generalization abilities and was faster than MCE.

In Ben-Yishai and Burshtein (2004), a discriminative training algorithm for the estimation of Hidden Markov Model (HMM) parameters is presented. This algorithm is based on an approximation of the Maximum Mutual Information (MMI) objective function and its maximization in a technique similar to the expectation-maximization (EM) algorithm. The algorithm is implemented by a simple modification of the standard Baum–Welch algorithm, and can be applied to speech recognition as well as to word-spotting systems. Three tasks were tested: Isolated digit recognition in a noisy environment, connected digit recognition in a noisy environment and word-spotting. In all tasks a significant improvement over maximum likelihood (ML) estimation was observed.

In Markov and Nakagawa (1998), a new discriminative training method for Gaussian Mixture Models (GMM) and its application for the text-independent speaker recognition is described. The objective of this method is to maximize the frame level normalized likelihoods of the training data. In contrast to other discriminative algorithms, the objective function is optimized using a modified Expectation-Maximization (EM) algorithm which greatly simplifies the training procedure. The evaluation experiments using both clean and telephone speech showed improvement of the recognition rates compared to the Maximum Likelihood Estimation (MLE) trained speaker models, especially when the mismatch between the training and testing conditions is significant.
In Chen and Soong (1994), an N-best candidates based discriminative training procedure for constructing high performance HMM speech recognizers is proposed. The algorithm has two distinct features. The first one is, N-best hypotheses are used for training discriminative models, and the second one is, a new frame-level loss function is minimized to improve the separation between the correct and incorrect hypotheses. The N-best candidates are decoded based on tree-trellis fast search algorithm. The new frame-level loss function, which is defined as a halfwave rectified log-likelihood difference between the correct and competing hypotheses, is minimized over all training tokens. The minimization is carried out by adjusting the HMM parameters along a gradient descent direction. Two speech recognition applications have been tested, including a speaker independent, small vocabulary (ten Mandarin Chinese digits), continuous speech recognition, and a speaker-trained, large vocabulary (5000 commonly used Chinese words), isolated word recognition. Significant performance improvement over the traditional maximum likelihood trained HMM’s has been obtained.

Minimum Classification Error (MCE) approach for speaker verification is proposed in Liu et al (1994). In this approach Liu et al (1994), all the competing speakers are used to evaluate the score of the anti speaker and formulates the optimization criterion such that speaker recognition error rate on the training data is directly minimized. They also proposed a normalized score function which makes the verification formulation consistent with the minimum error training objective. They show that speaker recognition performance is significantly improved when discriminative training is incorporated. Since all the competing speakers are used to evaluate the score of the anti speaker, it is not practical for verification test over a large population.
In Hong and Kwong (2004), Maximum model distance algorithm for GMM is described for speaker identification task. This approach (Hong & Kwong 2004) tries to maximize the distance between each model and a set of competitive speakers models. TIMIT corpus is used to evaluate this proposed training approach. The results show that the identification performance can be improved greatly when the training data is limited.

In Lie et al (1995), use of discriminative training to construct hidden Markov models of speakers for verification and identification is studied. As opposed to conventional maximum likelihood training which estimates a speaker’s model based only on the training utterances from the same speaker, a discriminative training approach is used which takes into account the models of other competing speakers and formulates the optimization criterion such that speaker separation is enhanced and speaker recognition error rate on the training data is directly minimized. The optimization solution is obtained with a probabilistic descent algorithm.

The Gaussian mixture model – Universal background model (GMM–UBM) system is one of the predominant approaches for text-independent speaker verification, because both the target speaker model and the impostor model (UBM) have generalization ability to handle “unseen” acoustic patterns. However, since GMM–UBM uses a common anti-model, namely UBM, for all target speakers, it tends to be weak in rejecting impostors’ voices that are similar to the target speaker’s voice. To overcome this limitation, Chao et al (2009) proposed a discriminative feedback adaptation (DFA) framework that reinforces the discriminability between the target speaker model and the anti-model, while preserving the generalization ability of the GMM–UBM approach. This is achieved by adapting the UBM to a target speaker dependent anti-model based on a minimum verification
squared-error criterion, rather than estimating the model from the scratch by applying the conventional discriminative training schemes.

In Kwong et al (2000), an Improved Maximum Model Distance (IMMD) is proposed for HMM based speech recognition task. However, the MMD approach regards all competitive models as having the same importance when considering their contributions to the model re-estimation procedure. This is not completely practical since some competitive models might not be the real competitors if its likelihood is much lower than that of the labeled model. To determine which approach offers the best performance, different competitors should also be paid different levels of attention according to their competitive ability against the labelled model. Experimental results showed that a significant reduction in errors could be achieved with this new approach when compared with the maximum model distance criterion.

In Miyajima et al (2001), a new frame work for designing a feature extractor in a speaker identification system based on the Discriminative Feature Extraction (DFE) method is presented. In order to find the frequency scale appropriate for accurate speaker identification, a mel-cepstral estimation technique using a second-order all-pass warping function is applied to the feature extractor; the frequency warping and the text independent model parameters are jointly optimized based on a Minimum Classification Error (MCE) criterion.

In Srikanth and Murthy (2010), GMMs have been built for each speaker discriminatively based on the available positive and negative examples for each speaker. In this approach (Srikanth & Murthy 2010), speaker models are trained by moving the mean values of the mixture components in such a way as to maximize the likelihood of speaker data while also minimizing the likelihood of negative examples for the speaker.
The effectiveness of this approach on classification accuracies on speaker recognition tasks is tested on the NTIMIT database and NIST SRE 2003 corpora. The results indicate improvements in the performance of the system built using this new approach when compared to the traditional GMM-based speaker recognition systems.

3.2.3 Feature-Level Discrimination

A new selective training method is proposed by Arslan & Hansen (1999), which controls the influence of outliers in the training data on the generated models. The resulting models are shown to possess feature statistics which are more clearly separated for confusable patterns. The proposed selective training procedure is used for hidden Markov model training, with application to foreign accent classification, language identification, and speech recognition. The resulting error rates are measurably improved over traditional forward-backward training under open test conditions. The proposed method is similar in terms of its goal to maximum mutual information estimation training, however it requires less computation, and the convergence properties of maximum likelihood estimation are retained in the new formulation.

In Nagarajan and O' Shaughnessy (2007), a discriminant measure, using a product of Gaussian likelihoods, to estimate the amount of bias is proposed. By adjusting the complexity of the models, they show that this bias can be neutralized and a better classification accuracy can be achieved. The experiments are carried out on the OGI-MLTS telephone speech corpus on a language identification task. The results show that a better classification accuracy can be achieved without any degradation in the performance of any of the individual classes. Since the bias removal method is based on likelihoods, it can be utilized in any of the HMM/HMM-based classifiers.
Chi-Sang Jung et al (2010) proposed a new feature frame selection method based on the normalized minimum-redundancy and maximum -relevancy (NmRMR) criterion, which minimizes redundant information between selected feature frames but maximizes mutual information between speaker models and test feature frames. If the proposed criterion is also able to extract distinctive characteristics of speaker, it can be used as an effective feature frame selection method for speaker recognition systems. It is verified by experiments that the method proposed by Chi-Sang Jung et al (2010) produces consistent improvement, especially in a speaker verification system. It is also robust against variations in acoustic environment.

In Espy-Wilson et al (2006), speaker identification system using a set of features used to characterize speaker-specific information is proposed. A small set of low-level acoustic parameters that capture information about the speaker’s source, vocal tract size and vocal tract shape is described. The features consists of four formants (F1, F2, F3, F4), the amount of periodic and aperiodic energy in the speech signal, the spectral slope of the signal and the difference between the strength of the first and second harmonics. A Gaussian mixture model based text independent speaker identification system is created using this speaker-specific low level acoustic features. Performance of the system using low level acoustic feature set is compared with conventional GMM based speaker identification system using MFCC features.

In Kwon and Narayanan (2007), a simple method that employs only feature vectors that are deemed to contribute to discrimination is described. To overcome decision errors that arise due to model overlap, speaker models are trained to separate the data and select only useful feature vectors for more accurate speaker identification. Experimental results showed that this approach improve the speaker identification performance in
overcoming some of the difficulties arising when speaker models appear overlapped in a given feature space. The method is hence useful for detecting speakers from short segments in speech indexing applications as well as for improved performance for rapid speaker identification.

To avoid playback of recorded voice of the genuine speaker, a text prompted speaker verification task using HMM and Multilayer Perceptron (MLP) is described in (Delacretaz & Hennebert 1998). The set of context-independent phoneme HMMs is used to provide a segmentation of the speech signal into phonemes with a simple Viterbi forced alignment. The feature vectors, labeled with the corresponding phonemes, are then used to train MLPs, one per phoneme and per speaker. The discriminative power of the most frequently appearing phonemes was investigated. However, those phonemes are not unique to the particular speaker.

In another approach (Campbell et al 2006), GMMs themselves are used as feature vectors, called supervectors, to train Support Vector Machines (SVM) for speaker and language recognition tasks. Gaussian mixture models (GMMs) have proven extremely successful for text-independent speaker recognition. The standard training method for GMM models is to use MAP adaptation of the means of the mixture components based on speech from a target speaker. Recent methods in compensation for speaker and channel variability have proposed the idea of stacking the means of the GMM model to form a GMM mean supervector. In Campbell et al (2006), two new SVM kernels based on distance metrics between GMM models are described.

Better classification accuracy can be achieved if the training technique can be made to capture the unique features of a class, i.e., the features that discriminate one class from other, efficiently. In this thesis, we carry out research to improve the performance of the classification task, specifically speaker recognition task, by using unique characteristics of a
class at the feature-level and at the phoneme-level, the details of the research work is described in the subsequent chapters.

3.3 SUMMARY

This chapter describes the importance of discriminative approach in the classification task. The survey of different discriminative approaches used the literature to increase the discriminative power of the classification tasks are presented.