1. Introduction

Pattern classification is one of the most active research and application areas of artificial neural networks (ANN). This thesis addresses the application of ANN as a pattern classifier focusing on a specific topic in pattern recognition: automatic speech recognition (ASR), where the goal is to automatically produce a text transcription of spoken words. ASR is one area in the context of speech, which involves development of pattern classification models from speech data. Research in the field of ASR attained attention of researchers because of its fast growing demands in versatile applications. Speech being the most direct form of human communication, successful ASR systems can enhance the ease, speed, and effectiveness with which humans can interact with machines.

Though ASR is well developed for western languages, ASR in the area of Malayalam is relatively less investigated. The research work presented in this thesis is a small contribution towards the larger goal of developing a general purpose continuous speech recognition (CSR) system for Malayalam. Hidden Markov models (HMM), a statistical framework that supports both acoustic and temporal modeling, are extensively applied in state-of-the art ASR systems. However, HMMs have some drawbacks; the major one is poor discrimination power. Also, for reasons of efficiency, HMMs make a number of suboptimal assumptions while modeling the basic speech units. Consequently, the spectral and temporal correlations are poorly modeled.

The thesis first proposes an HMM based small vocabulary, speaker independent continuous speech recognizer for Malayalam. The system proposed is for the recognition of a predefined set of read sentences. The thesis then examines the integration of ANN into the HMM based system to alleviate some of the assumptions made by the HMM based methods and also to enhance the performance of the system.

2. Motivation

Many research groups have been working in the field of ASR, and in the past few years there have been proposed a number of ASR systems especially for European languages. Recently, there have also been various efforts towards ASR in Indian languages like Tamil, Telugu, Marathi and Hindi. Research in Malayalam speech recognition has started recently

and the area of CSR is relatively less investigated. Efficient voice interfaces in ones own native language can help even a common man to enjoy the benefits of information technology.

Inspired by the development of ASR systems for various Indian languages, the thesis first proposes a speaker independent Malayalam CSR system for a limited vocabulary task. The thesis then addresses methods to improve the recognition accuracy of the proposed HMM based system by incorporating phoneme posterior estimates computed using ANN approach.

When an ANN is used to solve a classification problem, the network can be trained either to provide the classification directly or to model the posterior probabilities of class membership. These probabilities can then be used in a subsequent decision making stage to arrive at a classification. This approach is very powerful and has been explored in this thesis to enhance the performance of the HMM based Malayalam CSR system. This hybrid HMM/ANN, take advantages of the complementary capabilities of connectionist networks (in particular their discriminative power) and of HMM models (in particular their capacity of handling time sequences).

3. Objective and Scope

HMM is a stochastic method which provides a rich and flexible mathematical framework for building recognition systems. The major advantages that support their application in the speech recognition area include their powerful learning and decoding methods for temporal sequences, good sequence handling capabilities and a rich mathematical framework. However, to take advantage of this representational power, the algorithms make numerous assumptions (eg:- uncorrelated features within an acoustic vector, first order Markov model assumptions for basic speech units) which are incorrect for human speech process. Consequently, the spectral and temporal correlations are poorly modeled.

Recognition accuracy is an important aspect of ASR systems. They must achieve a very high level of performance to be of general interest as a man-machine interface. Discriminative training leads to enhanced class separability in classification problems which gives reduced error rates and improved system performance. Improving discrimination in recognition systems is one of the fundamentally important research areas for ASR. In an HMM based

ASR system, a multivariate Gaussian mixture model (GMM) is typically used to represent the emission probability of each HMM state. The estimation of the parameters of the GMM is based on maximum likelihood (ML) criterion which assumes correctness of the models and is not discriminative. Also, maximizing the likelihood of the training data is not closely related to the typical evaluation criteria of the recognizer: the word error rate (WER). These issues with maximum likelihood estimation motivate an alternative form of estimating model parameters called discriminative training.

Neural networks classifiers are naturally discriminative. Several researchers have shown that the outputs of ANNs used in classification mode can be interpreted as estimates of a-posteriori probabilities of output classes conditioned on the input**.** Using Bayes rule, these state posteriors can be converted to likelihoods required by the HMM framework. The focus of the thesis is on methods to improve the posterior estimates from ANN which can enhance the recognition accuracy of the HMM based ASR system.

There are several possibilities of integrating ANN in the hybrid speech recognizer architecture based on the role of ANN. The thesis uses ANNs as front-ends for HMMs. In this hybrid model, HMMs perform the temporal modeling and ANNs perform the acoustic modeling. Although many types of neural networks can be used for classification purposes, our focus is on multilayer perceptron (MLP) which is the most widely studied and used neural network classifier for speech recognition.

4. Description of the Research Work

The thesis first describes the design and development of the proposed small vocabulary, speaker independent, continuous speech recognizer for Malayalam based on the state-of-the art HMM approach. The basic speech unit selected is a phoneme. The system uses context-independent modeling for its phonemes. The Speech databases have been developed to provide adequate coverage to the phonemes included in the application. Mel-frequency cepstral coefficients method is used to extract acoustic features from the input signal. Continuous density HMM, which is used in this work to model phonemes, represents the general case where the emission probability density functions are continuous. The emission probability density is approximated using a Gaussian mixture density with diagonal covariance matrices. Training is performed by the Baum-Welch algorithm, which is based

on the maximum likelihood approach. For decoding, the Viterbi algorithm is used.

The method of maximum likelihood may not always lead to an optimal recognition performance in the speech recognition systems as the correlation between the likelihood and WER may be weak. Also, in practical situations, where the training data is limited, ML training may lead to unreliable estimate of the parameters. Therefore, it is preferable to employ a training scheme that explicitly aims at reducing the word error rate and that addresses the data sparsity problems. Hence, the thesis examines the prospect of integrating ANN, which optimizes parameters for their discriminative power rather than to maximize likelihood, with HMM. The advantages of the hybrid system like discriminative training which focuses on estimating model parameters that minimize the error rate, capability to incorporate contextual information etc. have been studied. Experimental studies have shown that the posteriors derived from MLP trained in multi-class classification mode enhance the performance of the HMM based system.

Many real applications translate into classification problems with a large number of classes and a huge number of data. ASR falls into this category and the high number of classes to be separated makes the boundaries between classes complex. A multi-class classifier cannot perform at a high level in such cases as the classification algorithm has to learn to construct a large number of separation boundaries. Many studies have demonstrated that an adequate decomposition of such real world problems into sub problems can be favorable to the overall computational complexity. Motivated by this fact, the thesis further examines the prospect of using a pairwise modeling approach, which is based on the 'divide-and-conquer' strategy, as a way to improve overall classifier performance. The proposed approach is simple since the binary decision is learned on fewer training examples. It also improves the generalization ability of the network because of the redundancy in the training data and is very useful in the case of very limited training material.

In this thesis, MLP training criterion for posterior probability estimation, by way of training the MLP in multi-class and pairwise pattern classification approaches have been explored in detail. Methods to combine the outputs of pairwise classifiers to obtain phoneme posteriors have also been discussed. The thesis then describes the advantages of the proposed pairwise classifier and how the integration of the pairwise classifier, for estimation of posterior probability, improves acoustic phonetic modeling and recognition accuracy.

Speech corpora for research should be developed in a well designed manner to provide adequate representation of textual/linguistic information, regional/dialect variations, speaking style and environment etc. Speech databases for research purposes are available for several western languages. A selected set of Indian academic and research institutions in a consortium mode (Linguistic Data Consortium for Indian Languages) have recently developed speech corpora for various Indian languages, but the licensing policy is yet to be finalized. So we have developed in-house speech databases for experiments reported in this thesis. The speech corpora developed consist of naturally and continuously read Malayalam sentences from both male and female speakers who speak various dialects of Malayalam. Two datasets for training and one for testing have been developed. The entire corpus consists of 255 sentences with 1275 words and a total of 7225 phonemes by 20 (9 male & 11 female) different speakers.

5. Conclusions

Starting from a baseline HMM based ASR system for continuous Malayalam speech recognition, models based on the hybrid approach –both multi-class and pairwise- have been developed. Various experiments have been conducted to study the performance of the proposed continuous speech recognizer for Malayalam. The performance of the system has been evaluated using the popular WER metric. Sentence recognition rate and phoneme recognition rate, two other metrics used to assess the performance of the CSR system, have also been used to evaluate the performance of the systems developed. Experiments conducted show that the use of discriminative criteria in training improves the performance of ASR systems significantly compared to using the conventional maximum likelihood criterion. The results support our hypothesis that a pairwise neural network in estimating the posterior probability yields recognition results that outperform the performance gains with both the HMM and multi-class based HMM/ANN hybrid.

Publications

1) Anuj Mohamed, K.N. Ramachandran Nair, HMM/ANN Hybrid Model for Continuous Malayalam Speech Recognition, Procedia Engineering, 30(2012), 616-622.
2) Anuj Mohamed, K.N. Ramachandran Nair, Continuous Malayalam Speech Recognition using Hidden Markov Models, in Proc. 1st Amrita-ACM-W Celebration on Women in Computing in India, 2010.