سومین کنفرانس فناوری اطلاعات و دانش
۶ تا ۸ آذرماه ۱۳۸۶

Using Radial Basis Probabilistic Neural Network for Speech Recognition

Nima Yousefian, Morteza Analoui
Computer Department, Iran University of Science &Technology
Tehran, Iran
ni_yousefiyan@comp.iust.ac.ir, analoui@iust.ac.ir

***Abstract*— Automatic speech recognition (ASR) has been a subject of active research in the last few decades. In this paper we study the applicability of a special model of radial basis probabilistic neural networks (RBPNN) as a classifier for speech recognition. This type of network is a combination of Radial Basis Function (RBF) and Probabilistic Neural Network (PNN) that applies characteristics of both networks and finally uses a competitive function for computing final result. The proposed network has been tested on Persian one digit numbers dataset and produced significantly lower recognition error rate in comparison with other common pattern classifiers. All of classifiers use Mel-scale Frequency Cepstrum Coefficients (MFCC) and a special type of Perceptual Linear Predictive (PLP) as their features for classification. Results show that for our proposed network the MFCC features yield better performance compared to PLP.**

I. INTRODUCTION

A major problem in speech recognition system is the decision of the suitable feature set which can accurately describe in an abstract way the original highly redundant speech signal. In non-metric spectral analysis, mel-frequency cepstral coefficients (MFCC) are one of the most popular spectral features in ASR. In parametric spectral analysis, the LPC mel-cepstrum based on an all-pole model is widely used because of its simplicity in computation and high efficiency [1].

Another popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction [2]. PLP was originally proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency sub-band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel for example from a telephone line[3]. RASTA-PLP outperforms PLP for recognition of channel-distorted speech.

Many pattern classifiers have been proposed for speech recognition. During the last several years, Gaussian Mixture Models (GMMs) became very popular in Speech Recognition systems and have proven to perform very well for clean wideband speech [4]. However, in noisy environment or for noisy band-limited telephone speech the performance of GMM degrades considerably.

Another well known contemporary classification technique, Vector Quantization (VQ) is tested with our dataset and showed absolutely high recognition rate between other classifiers. So we decided to design a new probabilistic neural network (PNN) that uses a competitive function for its transfer function such as VQ networks. Results show that this network overcomes all other pattern classifiers in recognition of Persian one digit numbers dataset.

PNNs are known to have good generalization properties and are trained faster than the back propagation ANNs. The faster training is achieved at the cost of an increased complexity and higher computational and memory requirements [5].

II. SYSTEM CONCEPT

A. Dataset

A sequence of 10 isolated digits (0, 1, 2, …, 9) Voices from 35 different speakers were recorded in Computer Department of Iran University of Science and Technology. All of voices saved as wave files with audio sample size as 16 bit and audio sample rate 16 KHZ with mono channel. So there are 350 wave files. We divided them into two separate parts, 20 speakers (200 wave files) for training and 15 remaining speakers (150 wave files) for testing. So the ratio of train to test is 4:3.

B. Features Extraction

The goal of feature extraction is to represent speech signal by a finite number of measures of the signal. This is because the entirety of the information in the acoustic signal is too much to process, and not all of the information is relevant for specific tasks. In present ASR systems, the approach of feature extraction has generally been to find a representation that is relatively stable for different examples of the same speech sound, despite differences in the speaker or various environmental characteristics, while keeping the part that represents the message in the speech signal relatively intact.