Speech Recognition using MFCC and Neural Networks

Nidhi Srivastava
Amity Institute of Information Technology
Amity University
Lucknow, India
nidhinovy@gmail.com

Prof.(Dr.) Harsh Dev
PSIT College of Engineering
Lucknow, India
doctortharshdev@rediffmail.com

Prof.(Dr.) S.Qamar Abbas
Ambalika Institute of Management & Technology
Lucknow, India
qrat_abbas@yahoo.com

Abstract — The most common mode of communication between humans is speech. As this is the most preferred way, humans would like to use speech to interact with machines also. That is why, automatic speech recognition has gained a lot of popularity. Many approaches for speech recognition exist like Dynamic Time Warping (DTW), Hidden Markov Model (HMM). This paper shows how Neural Network (NN) can be used for speech recognition and also investigates its performance in speech recognition. Feed-Forward Neural Network with back propagation algorithm has been applied. For the feature extraction of speech Mel Frequency Cepstrum Coefficients (MFCC) has been used which gives a set of feature vectors of speech waveform. Earlier research has shown MFCC to be more accurate and effective than other feature extraction techniques in the speech recognition. The work has been done on MATLAB and experimental results show that system is able to recognize words at sufficiently high accuracy.

Index Terms—Speech Recognition; MFCC; Neural Networks;

I. INTRODUCTION

Speech is clearly one of the most important communication methods available between humans, and it is the primacy of this medium that motivates research efforts to allow speech to become a viable human-computer interaction (HCI) [1]. So, Automatic speech recognition (ASR) is viewed as an integral part of future human-computer interfaces, that are envisioned to use speech, among other means, to achieve natural, pervasive, and ubiquitous computing. However, although ASR has witnessed significant progress in well defined applications like dictation and medium vocabulary transaction processing tasks in relatively controlled environments, its performance has yet to reach the level required for speech to become a truly pervasive user interface. Indeed, even in “clean” acoustic environments, state of the art ASR system performance lags human speech perception by up to an order of magnitude [2].

In this paper, for speech recognition first feature extraction was carried out followed by the speech classification. The paper is divided into five sections. Section II describes the feature extraction techniques of speech waveform, section III describes the classification process, section IV shows the experimental setup and results and section V gives the conclusion.

II. FEATURE EXTRACTION

Every speech has different individual characteristics embedded in utterances. The basis of speech recognition lies in feature extraction as it distinguishes one speech from the other. But extracted feature should meet some criteria while dealing with the speech signal such as:

a. Easy to measure extracted speech features.

b. It should not be susceptible to mimicry.
c. It should show little fluctuation from one speaking environment to another.

d. It should be stable over time.

e. It should occur frequently and naturally in speech. [3]

Various feature extraction techniques exist like Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) etc. In this paper we have used MFCC for feature extraction as previous research has shown this technique to be the better than other techniques.

**A. MFCC**

The Mel-frequency Cepstral Coefficients (MFCCs) introduced by Davis and Mermelstein is perhaps the most popular and common feature for SR systems. For speech recognition purposes and research, MFCC is widely used for speech parameterization and is accepted as the baseline. This may be attributed because MFCCs models the human auditory perception with regard to frequencies which in return can represent sound better. [4,5]. They are derived from a mel-frequency cepstrum (minimize-of-spectrum) where the frequency bands are equally spaced on the mel scale, which approximates the human auditory system’s response more closely than the linearly-spaced frequency bands used in the normal cepstrum [6]. The block diagram of MFCC as given in [7] is shown in Fig. 1.

![Block diagram of MFCC](image)

We have used VOICEBOX speech processing toolbox [8] for our experiment. The function melcepst from this toolbox is used for calculating the MFCC coefficients.

**III. SPEECH CLASSIFICATION**

Various authors have used different methods like DTW, HMM, ANN etc. for classification of the speech.

Dynamic Time Warping (DTW) algorithm is one of the oldest and the most important algorithms in speech recognition. DTW algorithm is for measuring similarity between two time series which may vary in time or speed. DTW is a dynamic programming technique in which the entire problem is divided into a small number of steps each requiring a decision to be made based on the local distance measures [9, 10].

The most flexible and successful approach to speech recognition so far has been Hidden Markov Models (HMMs). HMM is the popular statistical tool for modeling a wide range of time series data. In Speech recognition area HMM has been applied with great success to problem such as part of speech classification [11].

In this paper we have used Artificial Neural Network (ANN) for speech classification. An artificial neural network consists of a potentially large number of simple processing elements (called units, nodes, or neurons), which influence each other’s behavior via a network of excitatory or inhibitory weights. Each unit simply computes a nonlinear weighted sum of its inputs, and broadcasts the result over its outgoing connections to other units.

A Neural Network is a feed-forward artificial neural network that has more than one layer of hidden units between its inputs and its outputs. Each hidden unit typically uses the logistic function to map its total input from the layer below to the scalar state that it sends to the layer above [14].

The architecture of network used for a two-layer feed-forward network algorithm has been given in Fig 2. An elementary neuron with X1, X2 … inputs has been shown. Each input is weighted with an appropriate value wij [12].
Backpropagation, also known as Error Backpropagation or the Generalized Delta Rule, is the most widely used multi layer feed-forward supervised training algorithm for neural networks. The BPNN consists of three or more fully interconnected layers of neurons which can be trained by using the Backpropagation Algorithm (BP). The BP training can be applied to any multilayer NN that uses differentiable activation function and supervised training (Wasserman, 1989). BPN is an iterative algorithm designed to minimize the mean square error between the actual input and desired output [13].

IV. IMPLEMENTATION & RESULT
The Experiment has been conducted using MATLAB 2008b with Neural Network toolbox. In this study we took 1 male and 1 female voice and two words. Each one of them recorded each word thirty times. So, in total we had 120 voice samples. These samples were recorded in MATLAB. The sampling frequency for all recording was 44100 Hz. Of these files 60 samples were used for training sets and 60 samples were used for testing sets.

We calculated 12 coefficients of MFCC which was passed as input to the NN.

Finally, Neural Network toolbox of MATLAB was used to create, train and simulate the networks and mean square error was used to evaluate its performance. As already mentioned NN consists of neurons. These neurons can use any differentiable transfer function / to generate their output. The transfer function used in our case for hidden layers is tan-sigmoid, and for the output layer is linear.

Fig. 3 shows the MSE in the training phase. Mean Squared Error (MSE) is the average squared difference between outputs and targets. Lower values are better. Zero means no error. In our case the network error reaches almost zero as is clear from the below figure.
Trained data was different from the tested data. When the trained network was simulated against tested data it gave good result and could easily recognize the words. More than 90% accuracy was achieved. Getting the larger training data will improve the result.

V. CONCLUSION

In this paper, we have used MFCC and Neural Network for speech recognition. The whole experiment has been implemented on MATLAB R2008b using Neural Network toolbox and it successfully recognizes speech. The simulation shows high accuracy in result. Further, improvement can be made in this method which will yield more accurate and precise result.

REFERENCES


[12] R. B. Shinde, Dr. V. P. Pawar “Vowel Classification based on LPC and ANN”, IJCA, Volume 50 – No.6, July 2012.