Proposed Model of Speech Recognition using MFCC and DNN

Subodh Virkar[1],Archana Kadam[2],Shohaib Mallick[3], Nikhil Raut[4], Satyam Tilekar[5]
[1,3,4,5] Student, Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering,Nigdi,Pune, Maharashtra, India
[2]Professor, Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering, Nigdi, Pune, Maharashtra, India

Abstract **– Speech Recognition is the ability of the machine to identify the word or phrases in human language and convert it into machine understandable form. Speech Recognition allows you to supply input to an application together with your voice. Speech recognition systems aim to form human machine communication quickly and simply . The main focus of the project would be to convert the speech of a human into text. In this paper, we propose a system architecture that will fetch speech data, process it and give out an effective text outcome.**

For Demonstration, we use Polygon smoothing algorithm to pre-process the data, and then use MFCC for feature extraction and we did the comparative study of classification techniques like SVM models and DNN. Although the system provides classification using SVM, in comparison, our system proves to be more precise and accurate if we consider big data to be processed.

Over the last few years, deep neural networks (DNNs) have become increasingly popular in many areas including ASR, so we have carried out a detailed survey about Speech Recognition using DNN.

Keywords: Speech Recognition, Polygon Smoothing Algorithm, MFCC, SVM, DNN

1. INTRODUCTION

The field of speech remains open for research as a 100% effective system is yet to be developed. Speech recognition and speech understanding systems have made their way into mainstream applications and almost everybody has used a speech recognition device at one occasion or another. The foremost natural application of speech recognition is dictation; where speech recognition is accustomed to compose letters/e-mails and other documents. For several languages, such dictation software has been quite widely used for several years already, and it works alright. Additionally to dictation, the recognition of coherent natural speech is utilized in many applications where speech recognition is hidden from the user, as an example, in automatic transcription of audio archives, so on enable better organization and indexing. Application of this classification technique could be Voice recognition, classifying words for better use, error detection in English language, use of these classification techniques in image processing, etc. More advance speech recognition system can offer applications in military appliances, medical appliances, home security systems, etc. Speech recognition is an important asset to control and to solve crimes, to evolve search engines, for identification purpose, to protect smartphone or bank account. It plays very important role where security, safety are prior concern and also for human machine interaction. The goal is to create a speech recognition module which is more precise and has fewer errors than the existing system. Our proposed model will be a system which will take audio as input, pre-process it (remove noise, surround sound, etc.), process the data and give desired text output. We will use methods like Polygon smoothing algorithm, MFCC, DNN and softmax method (if needed).

1.1 LITERATURE SURVEY:

Number of people have collected data and analysed Speech recognition prior to this research. The data collected can be divided into three parts namely voice pre-processing, feature extraction, feature classification using different methods.

1. The paper [1] shows the use of polynomial smoothing technique (Savitzky-Golay).Then use the modify autocorrelation function (MACF) and modify of the typical magnitude difference function (W-AMDF).This above methods are wont to separate addressed voice waveform, consistent with the characteristics. Experimental results show that the using the tactic can effectively improve the sound files of effective pre-signal extraction accuracy, improve the efficiency of signal processing and reduced the pitch. The voice of each person has a fixed cycle. Therefore the voice file by changing the location of speech signal pitch is used to achieve the voice pitch, tone length, temperature changes.
2. The paper [2] considers pre-processing of voice signals for voice recognition system based on the use of artificial neural network. It uses the method of eigenvalues decomposition, which is based on eigenvalues analysis of the autocorrelation matrix or data matrix. This method provides better resolution and parameter estimation than other parametric methods, especially at low signal/noise ratio. To implement this approach, the voice message is separated into components phonemes which can distinguish peaks increase and decay level the signal. The pre-treatment is carried out in MATLAB using envelop () function, which returns upper and lower limits of input sequence.

3. In paper [3] [4] they have discussed that Feature Extraction is that the process of extracting important information from the recorded speech .MFCC, LPC is employed as a feature extraction technique. As observed within the past research on speaker recognition systems, accuracy of the system decreases when the amount of input voice samples increases.
   a. Mel Frequency Cepstral Coefficient (MFCC) :
      i.Framing and Windowing
      ii.Perform Full Fourier Transform on windowed signal.
      iii.Pass these signals through Mel Filter bank.
      iv.Taking log and performing Discrete Fourier Transform on the signals.
   b. Linear Predictive Coding: It is desirable to compress the signal for efficient transmission and storage. Digital signal is compressed and so transmitted for efficient utilization of channels on wireless media. For medium or

low bit rate coder, LPC is most generally used. The LPC calculates an influence spectrum of the signal. After Framing, windowing is completed to smooth the signal and take away the discontinuities using a hamming window. Twelfth-order autocorrelation coefficients are found and reflection coefficients are calculated using the Lavison-Durbon algorithm.

4. Paper [5] has proposed a speech recognition model where feature extraction is done using linear predictive method and LPC parameter for every word is obtained. In the speech recognition section of the appliance, the Support Vector Machines (SVM) method is employed. These are the Soft Margin and Least Square SVM classifier. As a result, 91% accurate recognition success for the soft margin SVM classifier, 71% correct recognition of the least square SVM classifier has been achieved. But, in practice, changes in the pitch, microphone position, noise in the recording medium, changing parameters have been factors affecting the recognition success. In the speech recognition section of the application, finding the best parameters by performing a parameter scan on a large scale is one of the factors that can improve the recognition performance.

5. In Paper [6] the author has discussed Mel Frequency Cepstrum Coefficients feature extraction technique with support vector machine. SVM is used to distinguish the emotion of the sample given. An SVM classifier differentiates between anger, happiness, fear, sadness and updates the database as it goes. Author also discussed characteristics of Support Vector Machines in the segregation and classification of different aspects of speech that are extracted, such as amplitude, pitch etc. that are essential to understand the speaker's state. With this we can acquire a fully functional recognition system that can be used for security based systems.

6. Paper [7] discusses a detailed study and exploration on the algorithm of pronunciation error detection from three different perspectives: posterior probability calculation, hypothesis testing model and supervised classification. We mainly study the simpler non-interpolation processing method of pitch curve features in DNN, and embed the pitch into HMM-DNN acoustic model, which improves the system's ability to distinguish tone and stress. Then, binomial logistic regression model is used as the basic binary classifier, and the underlying shared network is a common feed forward neural network. They validate the performance of pronunciation error detection based on a common learning classifier in an English learning database.
7. Paper [8] presented a method of automatic annotation of speech corpora, using transcriptions from two complementary ASR systems. Our experiments showed that ASR systems, with HMM-GMM and DNN acoustic models, make significantly less identical errors compared to ASR systems both using HMM-GMM acoustic models. The method used in this paper aims at obtaining a high-quality annotated speech corpus in an automatic, unsupervised fashion. The newly obtained speech corpus is intended to be used to retrain the existing ASR systems, increasing the acoustic variability of the models and consequently boosting the transcription accuracy.

1.2 COMPARATIVE STUDY

One specific profit that neural network models have over SVMs is that their size is fastened. they're constant quantity models, whereas SVMs square measure statistics. That is, in a very DNN you have got a bunch of hidden layers with sizes h1 through hydrogen azide counting on the quantity of options, and bias parameters, and people compose your model. Against this, associate SVM (at least a kernelized one) consists of a collection of support vectors, elect from the coaching set, with a weight for every. within the worst case, the quantity of support vectors is precisely the quantity {of coaching|of coaching} samples (though that in the main happens with little training sets or in degenerate cases) and generally its model size scales linearly. In the linguistic communication process, SVM classifiers with tens of thousands of support vectors, every having many thousands of options, isn't extraordinary. If you would like to use a kernel SVM you have got to guess the kernel. However, DNNs square measure universal approximators with solely estimate to be done is the breadth (approximation accuracy) and height (approximation efficiency).

Speech recognition model with SVM	**Speech recognition model with DNN**
1.Requires a short number of samples for training.	1. Requires a large number of samples for training.
2. Effective for short size of data.	2. Effective for large size of data.
3.Good generalization ability even with a few training samples.	3. A poor generalization ability if trained with few training samples.
4. Training multiclass SVM is not easy.	4. Gives many number of outputs

1.2.1 Table of comparison

1.3 ALGORITHMS STUDIED

1.3.1 Pre-processing using ANN:

Here we are using MATLAB (envelope function) this algorithm improves the susceptibility of the neural network against the input data . This method is used when the voice is recorded live. The analysis is done not only for the code sequence but also for the phonemes. This indirectly simplifies the task of speech recognition

1.3.2 Mel Frequency Cepstral Coefficients (MFCC):

MFCC are the Mel Frequency Cepstral Coefficients. MFCC takes under consideration human perception for sensitivity at appropriate frequencies by converting the traditional frequency to Mel Scale. These coefficients represent the features like power, pitch, and vocal tract configuration present in the speech signal. It is the best feature extraction method that gives highest accuracy.

1.3.3 Support Vector Machines (SVM):

Support Vector Machines (SVM) is a learning method for the solution of classification and regression problems, based on statistical learning theory and the least structural risk. Examples of application areas of SVM include handwriting recognition, face recognition, 3- dimensional object recognition, speech recognition, speaker recognition, speaker verification, text classification. SVM makes use of a hyperplane which acts sort of a decision boundary between the various classes. In SVM, we plot data points as points in an n dimensional space (n being the number of features you have) with the value of each feature being the worth of a specific coordinate. SVM are often want to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data.

1.3.4 Deep Neural Network (DNN):

A neural network approach for classification using features extracted by a mapping is presented. When the number of sample dimensions is much larger than the number of classes and no deviations are given but the means of classes, a mapping from class space to a new one whose dimensions is exactly equal to the number of classes is proposed. The vectors in the new space are considered as the feature vectors to be inputted to a neural network for classification. The property that the mapping does not change the separability of the original classification problem is given. Simulation results for speech recognition are presented.

2. PROPOSED ARCHITECTURE

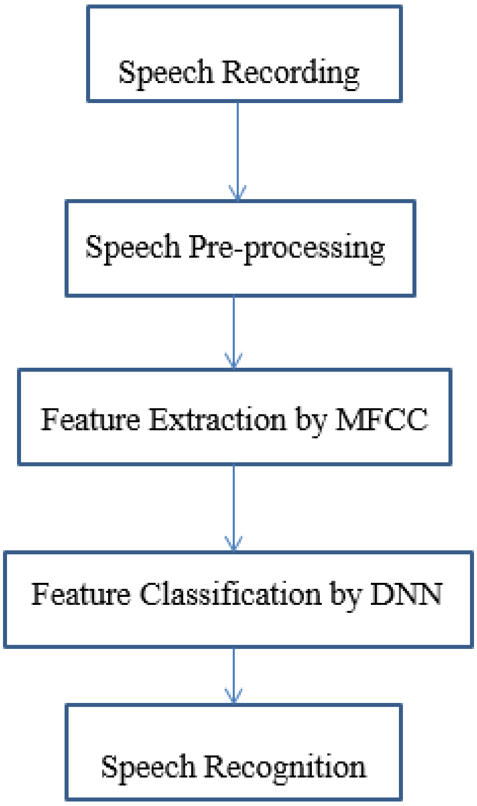

Figure 2.1. Block diagram of proposed architecture

As shown in above figure first we will be recording the voice of a person this voice signal is pre processed using ANN and we will get smooth and noise free voice signal which will be passed further for processing ,now the features from the voice signal is extracted using MFCC feature extraction algorithm and these coefficients are fed to DNN for classification.A neural network approach for classification using features extracted by a mapping is presented.

3. CONCLUSION

From the comparative study we have come to the conclusion that Neural Networks should be used instead of support vector machines (SVM) to handle large size of data and to get better accuracy with the help of features extracted from Mel Frequency Cepstral Coefficient (MFCC).

4. FUTURE SCOPE

It can be used in various regional languages because the existing system uses only English as the input language. Also voice biometric attendance can be taken in future using voice recognition system. Voice based games like chess which involves commands to character through voice. Banks are developing systems which allow transactions enabled through voice based security authentication. Speech recognition is also used for education purposes by people with disabilities also substantial research is being done in the field of military to implement voice recognition systems.

5. REFERENCES

[1] YANG Fan, LIU Ming-hui, XU Sun-hua, PAN Guo-feng “research on a new method of preprocessing and speech synthesis pitch detection” , 2010

[2] Gulmira K. Berdibaeva,Olen N .Bodin,Valery V. Kozlov, Dimitry I .Nefed’ev “pre-processing voice signals for voice recognition systems” , 2017

[3] Asma Mansour, Zied Lachiri ,”A comparative study in emotional speaker recognition in noisy environment”,2017

[4] Neha Chauhan,Tsuyoshi Isshiki, Dongju Li,”Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database”,2019

[5] Osman Eray , Sezai Tokat ,Serdar Iplikci ,”An Application of Speech Recognition with Support Vector Machines”,2018

[6] Ashwini Rajasekhar and Malaya Kumar Hota,”A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines”,2018

[7] Wu Ying,”English Pronuntiation Recognition And Detection Based on HMM-DNN”,IEEE 2019.

[8] Lucian Georgescu, Horia Cucu,”Automatic Annotation of Speech Corpora using Complementary GMM and DNN Acoustic Models.”,IEEE 2018.