A Survey On Speech Recognition System Implementation Methodologies

Ch. Ramaiah

Department Of Computer Science and Engineering
V.R.Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

Abstract

Speech recognition system recognizes the speech samples. There are many speech recognition systems implemented based on well known algorithms. Generally speech recognition systems have two parts the first part is feature extraction and second part is classification. There are so many algorithms for feature extraction and classification. Traditionally MFCC, LPC etc used for feature extraction phase. HMM, SNN, Neural Networks etc are used to classification.

Here this paper brings the survey of various methodologies to implement speech recognition system and their back bone algorithm. The main aim of this review is to develop an efficient speech recognition system to provide good recognition by selecting good algorithms for feature extraction and classification.

Keywords- ***MFCC*, HMM, SNN.**

I. INTRODUCTION

The speech recognition field has so many years old. Many interesting advances and developments have been going since the earliest speech recognizer at Bell Labs in the early 1950's. The development of ASR (Automatic Speech Recognition) increased gradually until the invention of Hidden Markov Models (HMM) in early 1970's. Many of the professional's contribution were to make use of ASR technology to what can be seen nowadays of various advancements in fields like multi-modal, multi-lingual/cross-lingual ASR using statistical techniques such as HMM, SVM, neural network.

The methodology of speech recognition system methodology is represented as below

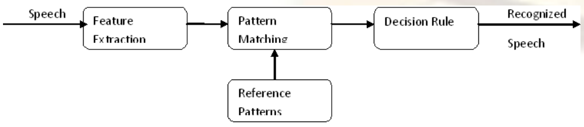

Fig1: Representation of speech recognition.

Feature extraction means extracting features from speech samples. Preprocessing, windowing, filtering are involved in this strategy. Training and recognition steps are occurred later.

This speech recognition system will be useful in so many speech based applications. In many fields like Banking, Security, Scientific and Voice based applications are the applicable areas for these speech recognition systems. Speech recognition

II. RELATED WORK

In Bangla speech recognition system [1] LPC and ANN are key algorithms. Here the implementation consists two parts they are speech signal processing and speech pattern recognition. First we look at the construction of feature extractor. It has three modules like sound recording, pre-emphasis filter and speech coding. Here the feature extractor uses a standard LPC Cepstrum coder, which converts the incoming speech signal into LPC Cpestrum feature space. A vector of 12 Linear Predicting coding Cepstrum coefficients is calculated from each data block using Durbin's method and the recursive expressions developed by Furi.

Pattern recognition is based on artificial neural network (ANN). Recognizer has designed to 10 digits and each digit input to the recognizer of size of 72 features, feature vector. Multilayer Perception Approach (MLPA) is used here. It is a new layered approach consists 72 input nodes, variable number of hidden nodes and 10 output nodes. The network has been utilized the sequence training method and its state is stored internally every time its weights adjusted in order to avoid the inconsistency of weights lead to infinite. The layered network stops the training process when the test set satisfies the condition checked at the end of each epoch of training set. After the completion of training the whole state of layered network is stored so that it can retrieve the state in the recognition process.

Artificial Neural Network adaptation

An *ANN* was used to classify the feature vectors of new speech. The outputs of network are the triphones obtained from baseline system [1],[9]. The identity of a triphone often depends not only on the spectral features at one point in time, but it also depends on how the features change over time **,** so the inputs to the network consist of the features for the frame to be classified, **as** well **as** the features for frames at -60, -30, 30, and 60 msec relative to the fiame to be classified (for a total of 195 input values).

1) Present all training data to the network, and its outputs are calculated and compared with the targets to generate an error.

2) For each time step, the error is backpropagated to find gradients of errors for each weight and bias. Because the contributions of weights and biases to errors via the delayed recurrent connection are ignored, this gradient is actually an approximation.
3) Use this gradient to update the weights with the backprop training function.

Maximum Likelihood Linear Regression (MLLR) speaker adaptation

Based on expectation-maximization (EM) technique MLLR computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data If a small amount of data is available it is better to compute a global adaptation transform applied to every Gaussian component in the model set [9]. However as more adaptation data becomes available, Gaussian components can be grouped and specific transformation can be computed for everyone. By way of a regression class tree, the set of transformations to be estimated can be chosen according to the amount and type of adaptation data that is available. Similar to tree-based state tying, the tying of each transformation across a number of mixture components makes it possible to adapt distributions for which there were no observations at all.

The regression tree is constructed using a centroid splitting algorithm, which uses a Euclidean distance measure to grow the binary regression class tree. Each leaf node of the tree specifies a particular mixture component cluster. This algorithm proceeds as follows
1. Select a terminal node to split.
2. Calculate the mean and variance from the mixture components clustered at this node.
3. Create two children whose means are initialized to the parent mean perturbed in opposite direction from each other by a fraction of the variance.
4. Using a Euclidean distance measure, each component at the parent node is assigned to one of the children whose mean is closer to the component.
5. Based on the component assignments, calculate the new means for the children. Go to step 1) until there is no change in assignments from one iteration to the next or the requested number of terminals has been achieved.

There is another speaker dependant system [2]. A database has been created that contains 1000 total voice patterns for 10 digits for each digit there are 100 repititions.500 are used for training and 500 are used for testing same as Abdul Ahad has used. For English words there are 1200 voice patterns 40 for each word.900 repetitions are used for training and 300 are used for testing. All of these recorded sound files are of one speaker. This is mono speaker database. Features are extracted using MFCC. There are 67x39 values after applying MFCC with two Deltas for dynamicity of voice. These 67x39 values are reduced to only 39 values by getting only maximum values from each column of 67x39 MFCC feature vector. Numbers of input neurons to the neural network are 39. Different numbers of hidden layers are tested. Using one hidden layer there is best solution with 19 neurons. Output layer have 10 neurons for digits. On the basis of this technique 30 English words are also trained and recognized. By using full features 67x39 it takes much time for training. But with the use of 39 values training is complete within one hour. Here learning algorithm is back propagation.

The model of neural network used is given in figure2. The learning strategy of this type of neural network occurred is called supervised learning since it tells what to learn. It is up to the network to carry out how to learn process.

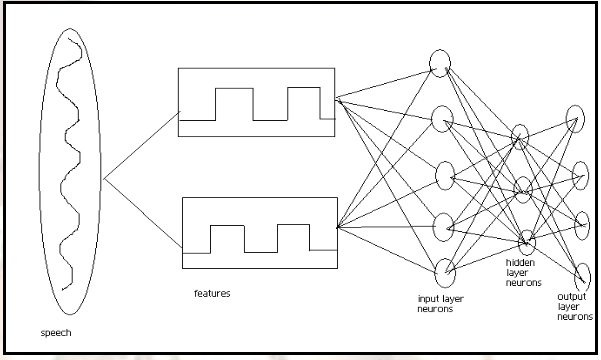

Fig 2: Neural network model for speech system.

K-Nearest Neighbor (KNN) classification

It is a very simple, yet powerful classification method [3]. The key idea behind KNN classification is that similar observations belong to similar classes. Thus, one simply has to look for the class designators of a certain number of the nearest neighbors and sum up their class numbers to assign a class number to the unknown.

In practice, given an instance $\boldsymbol{y}$, KNN finds the k neighbors nearest to the unlabeled data from the training space based on the selected distance measure. In our case, the Euclidean distance is used. Now let the k neighbors nearest to $\boldsymbol{y}$ be $Nk\ \boldsymbol{(\ y\ )}$ and $c(\boldsymbol{z})$ be the class label of $\boldsymbol{z}$.

It has been noticed that the success of any automatic speech recognition system [4] requires a combination of various techniques and algorithms, each of which performs a specific task for achieving the main goal of the system. Therefore, a combination of related algorithms improves the accuracy or the recognition rate of such applications. Figure 3 shows the architecture of the HMM based English digits speech recognition system.
Here total system consists multiple phases.
They can be represented by below figure.
Phase 1: reading of input signal (speech sample).
Phase 2: Extracting the features of speech samples.
Phase 3: Conversion of features into phonetic based words.

Phase 4: Performing the viterbi search to get optimal state sequence.
Filter bank consists two address spaces one is for linear which for lower than 1000Hz and second for logarithmic which is for higher than1000Hz. Mels can be calculated by using the below forula[7]

$$\mathbf{Mel(f) = 2595 * log10(1+f/700)}$$

We can observe the mfcc procedure through below block diagram.

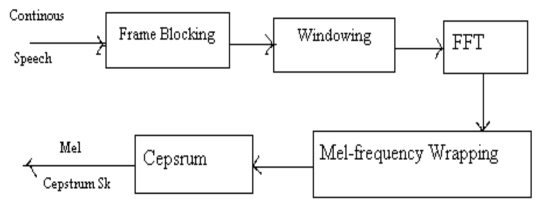

Fig 3: block diagram of MFCC.

In the final step, the log mel spectrum has to be converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT).

Here Hidden Markov Models (HMM) are used to train the features that are extracted through MFCC.

HMM is one of the powerful and dominating statistical approaches. It has a set of states Q, a set of transition probabilities A, a set of observation likely hoods B, a defined start state and end state (s), and a set of observation symbols.

There is another system based on multiple neural networks in which there are four main phases [12]. They are speech signal adoption, pre treatment, extraction of characteristic parameters, the recognition of PNN neural network and system identification. Speech training samples was generated spectrogram by spectrogram algorithm. Then, it is input into the PCNN [12] to obtain the characteristic parameter of PCNN. Finally the PNN is trained by using these characteristic parameter of PCNN. The sole icon of reflection input spectrogram is gained by specific speech signal when every parameters of PCNN cannot change. After these parameters is recognized, the result is showed that this system is increased the speech recognition.

III. CONCLUSION

Good Combination of feature extraction techniques and classification algorithm leads best result in speech recognition area. This study of methodologies for speech recognition systems enables us to implement speech recognition system with high recognition rate.

REFERENCES

[1] Paul A.K, Dilpankar Das, MD Mustafa Kamal “Bangla Speech recognition system using LPC and ANN”.

[2] S. M. Azam, Z.A. Mansoor, M. Shahzad Mughal, S. Mohsin “Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network”.

[3] Tsang-Long Pao1, Wen-Yuan Liao2, Yu-Te Chen3 “Audio-Visual Speech Recognition with Weighted KNN-based Classification in Mandarin Database”.

[4] Ahmad A.M.Abushariah, Teddy S.Gunawan , Othman O.Khalifa “English Digits Speech Recognition System Based on Hidden Markov Models”.

[5] L., Rabiner, and B. H., Juang. Fundamentals of Speech Recognition. Prentice Hall, NJ, USA, 1993.

[6] M. A. M. Abu Shariah, R. N. Ainon, R. Zainuddin, and O. O. Khalifa, “Human Computer Interaction Using Isolated-Words peech Recognition Technology,”

[7] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman“Speaker Identification Using Mel Frequency Cepstral Coefficients”.

[8] Nattanun Thatphithakkul1, Boontee Kruatrachue1, Chai Wutiwiwatchai2, Sanparith Marukatat2, and Vataya Boonpiam2, Non-members “Robust Speech Recognition Using KPCA-Based Noise Classification”.

[9] Xuelin Cheng, Han Wang, Zongge Li “Speech Adaptation Using Neural Networks for Connected Digit Recognition”.

[10] Mahmoud I. Abdalla and Hanaa S. Ali “Wavelet-Based Mel-Frequency Cepstral Coefficients for Speaker Identification using Hidden Markov Models”.

[11] Bai Maorui , Feng Mingming , Zheng Yuzheng “Speech Recognition System Using a Wavelet Packet and Synergetic Neural Network”.

[12] Bo Lu,Jing-jing Wu,Yu Wang , Jin-ping Li* “A Speech Recognition System Based On Multiple Neural Networks”.

[13] Anup Kumar Paul1, Dipankar Das2 , Md. Mustafa Kamal3 “Bangla Speech Recognition System using LPC and ANN”.

[14] Bai Maorui, Feng Mingming , Zheng Yuzheng “Speech Recognition System Using a Wavelet Packet and.