



Realtime Information System Based on Speech Recognition

Miss Himanshu¹, Sarbjit Kaur², Alpa Sharma³

¹M.Tech Student Department of Computer Science & Engineering, Modern Institute of Engineering & Technology, Haryana, Kurukshetra university, India

²Assistant Professor, Department of Computer Science & Engineering, Modern Institute of Engineering & Technology, Haryana, India

³M.Tech Student Department of Computer Science & Engineering, Modern Institute of Engineering & Technology, Haryana, Kurukshetra university, India

Abstract— *Speech recognition is the process of converting spoken words into text. In case of speech recognition the research followers are mainly using three different approaches namely Acoustic phonetic approach, Pattern recognition approach and Artificial intelligence approach. Speech recognition is the process of converting spoken words into text. One of the problems faced in speech recognition is that the spoken word can be vastly altered by accents, dialects and mannerisms.*

The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, database and performance evaluation.

Keywords— *Automatic Speech Recognition (ASR), Hidden Markov Model (HMM), NN(Neural Network), ANN (Artificial neural network), MLP(multi-layer perceptions).*

I. INTRODUCTION

Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine.

In simple words, speech recognition can be put together as the ability to take the audio format as input and then generate the text format from it as output.

Speech recognition [1] [2] involves different steps:

1. Voice recording
2. Word boundary detection
3. Feature extraction [3]
4. Recognition with the help of language models [4]

II. SPEECH RECOGNITION APPROACHES:

Speech recognition process deal with speech variability and account for learning the relationship between specific utterance and the corresponding word or word [5]. There has been steady progress in the field of speech recognition over the recent year with two trends [6]. First is academic approach and second is the pragmatic, include the technology, which provides the simple low-level interaction with machine, replacing with buttons and switches. A second approach is useful now, while the former mainly make promises for the future. There are three approaches to speech recognition [7] [8] [9].

A. Acoustic-phonetic approach [10][11][12][13]

Artificial Intelligence approach

Pattern recognition approach

1) Acoustic-Phonetic Approach :

In this speech recognition algorithm, the system tries to decode the speech signal in a sequential manner based on the observed acoustic features of the speech waveform and the known relations between acoustic features and phonetic symbols. Figure 1 shows a block diagram of the acoustic- phonetic approach to speech recognition. The first step in the process is the parameter measurement process, which provides an appropriate spectral representation of the speech signal. The next step in the processing is the feature detection stage where the spectral measurements are converted to a set of features that describe the acoustic properties of the various phonetic units. Finally, the recogniser tries to determine the best matching word or sequence of words.

2) Pattern Recognition Approach:

In this approach, the speech patterns are used directly without explicit feature determination and segmentation. The method has two steps-namely, training of speech patterns, and recognition of patterns by way of pattern comparison. In the parameter measurement phase, a sequence of measurements is made on the input signal to define the “test pattern”. The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern and reference pattern is computed. Finally the decision rule decides which reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase.

a) Template Based Approach:

Template based approach to speech recognition have provided a family of techniques that have advanced the field . A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate s words. Recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern.Each word must have its own full reference template;

b) Stochastic Approach:

Stochastic modeling [15] entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition.

The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions.A template based model is simply a continuous density HMM, with identity covariance matrices and a slope constrained topology. Although templates can be trained on fewer instances,the lack the probabilistic formulation of full HMMs and typically underperforms HMMs. Compared to knowledge based approaches; HMMs [16] [17] [18] [19] enable easy integration of knowledge sources into a compiled architecture. A negative side effect of this is that HMMs do not provide much insight on the recognition process. As a result,it is often difficult to analyze the errors of an HMM system in an attempt to improve its performance. Nevertheless,prudent incorporation of knowledge has significantly improved HMM based systems.

3) Artificial Intelligence Approach (Knowledge Based Approach):

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand.

III. PROBLEM

3.1 Problem In Generic Speaker Verification

The general approach to ASV consists of five steps: digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision, and enrollment to generate speaker reference models. A block diagram of this procedure is shown in Fig 1 [20].Feature extraction maps each interval of speech to a multidimensional feature space. (A speech interval typically spans 10 to 30 ms of the speech waveform and is referred to as a frame of speech).

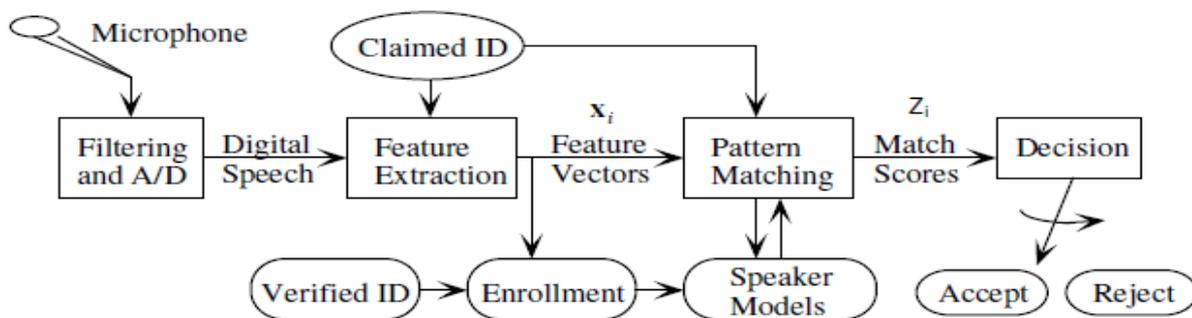


Fig 1: Generic speaker verification [20]

3.2 Speaker Recognition Problem

Regardless of which type of task, the problem may be further characterized as being text dependent or text-independent. In the text-dependent case, the train and test utterances are required to be a specific word or set of words; the system can then exploit the knowledge of what is spoken in order to better make a decision. For the text-independent case, there is no constraint on what is said in the speech utterances, allowing for generalization to a wider variety of situations. The dissertation work focuses on the text-independent speaker verification task. For each target (or hypothesis) speaker and test utterance pair, the system must decide whether or not the speaker identities are the same. In that case,their are two

types of errors occurs: false acceptance (or false alarm) and false rejection. A false accept occurs when the system incorrectly verifies an impostor test speaker as the target speaker. A false reject occurs when the system fails to verify a true test speaker as the target speaker. A trial refers to a target speaker and test utterance pair. In general, training data of target speaker may include one or more samples of speech, of variable length, the test data may also include varying lengths of speech samples. Our purposes, the train and test utterances will both be a single conversation side, that's typically 2.5-3 minutes of speech.

3.3 Objectives

The most significant factor affecting automatic speaker recognition performance is the variation in the signal characteristics (intersession variability and variability over time). Variations arise from the speakers themselves as well as from the recording and transmission channels, such as:

- Short-term variation due to the speaker's health and emotions
- Long-term changes due to aging
- Different microphones
- Different background noises (closed environment vs. open environment etc.)

IV. METHODOLOGY

4.1 Hmm Method

The HMMs (both unstructured, and structured) are trained according to the Baum-Welch algorithm given in. This is extremely similar to EM, but also takes state transitions into account. The idea is to train three quantities simultaneously: (1) the state transition matrix, A , (2) the observations probabilities, b_j , for each state j , and (3) the initial state probabilities, π_j . The observation probabilities are GMMs that calculate the probability that continuous feature vectors come from a given state. Parameter scaling is performed to prevent underflow. The Rabiner tutorial was followed explicitly; however, all coding was performed from scratch. Again, spurious singularities are avoided using variance limiting. The only difference between the U-HMM and S-HMM was during initialization. To make a U-HMM [21], the data can be clustered randomly according to k-means. Randomly initialized k-means segmentation is performed according to the number of desired states. This provides an initial segmentation of the audio frames into N states. After this segmentation, a GMM is trained upon each separate cluster according to the desired number of Gaussians, K . Using this method the initial observation Probabilities for the j^{th} state given the t^{th} observation, $b_j(O_t)$, are fully initialized. The state transition matrix, A , is initialized to equal probabilities (each entry is set to $1/N$). For the S-HMM, we initialize more carefully. Each state must occupy certain durations of each frame and the states must transition from one another in a sequential manner. To ensure that the training Data is divided into sequential, but similar segments. In this approach, each utterance is divided into N continuous regions of equal duration corresponding to the desired number of states. The initial GMM for each state is then trained on matching portions of each utterance (i.e., state one is trained on the first X frames of each utterance, state two is trained on the next X frames, and so on). The Viterbi algorithm is used on a single utterance of the data according to the current HMM model. The state assignments from the optimal Viterbi path are then used to update the state transition matrix and re-cluster the observations, assigning them to different states. All the GMM's are retrained on the new state clusters. This is performed iteratively until the state assignments from the Viterbi algorithm stop changing.

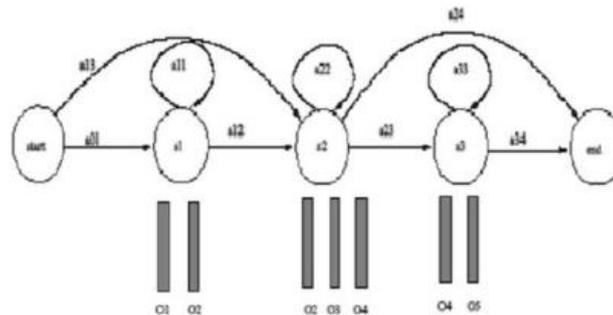


Figure2: Diagrammatic Representation of HMM

4.2 Neural Network

Although our previous experimented system has some advantages, a significant problem of very long time spending during recognition has obstructed the system in the practical implementation. There are several ways to overcome this problem. One is to use fewer references, which certainly effect system performance especially with large number of speakers to be recognized. What we expect is to use another recognition engine that provides optimal recognition rate and less processing time. Artificial neural network (ANN) [22] is one of our expectations due to its very fast processing. A neural network also is used to classify each frame as belonging to a specific speaker. The network has a three layered architecture and is trained using the back propagation algorithm. The number of the input nodes is equal to the size of the input vectors. The number of the output nodes is equal to the number of the registered to the system speakers. Finally, the number of the hidden nodes is chosen by the user. There were some researches proposing to use ANN in speaker recognition tasks. The speaker identification system for English language using ANN was implemented and found its advantages on both identification rate and fast processing. Multilayer Perceptron Multilayer Perceptron Neural Networks

are feed-forward and use the Back-propagation algorithm. We imply feed forward networks and Back-propagation algorithm (plus full connectivity). A typical topology of a fully connected feed forward network is shown in Fig 3. While inputs are fed to the ANN [23] forwardly, the ‘Back’ in Back-propagation algorithm refers to the direction to which the error is transmitted.

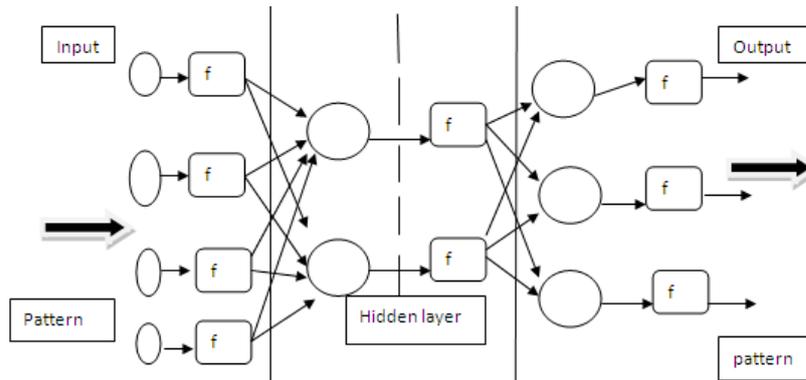


Fig 3: Architecture for FFNN for classification

Learning process in Back propagation requires providing pairs of input and target vectors. The output vector y of each input vector is compared with target vector d . In case of difference the weights are adjusted to minimize the difference. Initially random weights and thresholds are assigned to the network. The logistic function

$$F(x) = \frac{1}{1 + \exp(-x)} \dots\dots\dots 2.1$$

Which maps the real numbers into the interval $[-1 + 1]$ and whose derivative, needed for learning, is easily

$$\{ f'(x) = f(x)[1 - f(x)] \} \dots\dots\dots 2.2$$

The reason for its popularity is the ease of computing its derivative. Neural networks are adaptive statistical devices. This means that they can change iteratively the values of their parameters (i.e., the synaptic weights) as a function of their performance. These changes are made according to learning rules which can be characterized as supervised (when a desired output is known and used to compute an error signal) or unsupervised (when no such error signal is used). Back propagation [23] consists of measuring the error term between target output and the observed output.

V. EXPERIMENT ANALYSIS

5.1 Neural network training

An artificial neural network (ANN), often just called a neural network (NN), is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. The particular model used in this technique can have many forms, such as multi-layer perceptions or radial basis functions. The MLP is a type of neural network that has grown popular over the past several years. MLP's are usually trained with an iterative gradient algorithm known as back propagation. According to the Mel frequency the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included.

The below waveform represents the matching of speech signal of the speaker1, with the speech of the same speaker whose speech was earlier stored in the database. The implementation of this project is done in MATLAB and the results can be seen in a GUI. The GUI takes the filename of the speaker as input and gives the name of the speaker as output. The GUI basically contains a button named “record” when this button is pressed the speech is recorded and stored in database. And it also contains a plot field in which the output wave is plotted. After giving the input file name we have to press “record” then the wave will be displayed in field. The snapshots of the GUI when providing inputs and when results are displayed are shown below. It also contains a button “play” to play the record voice.

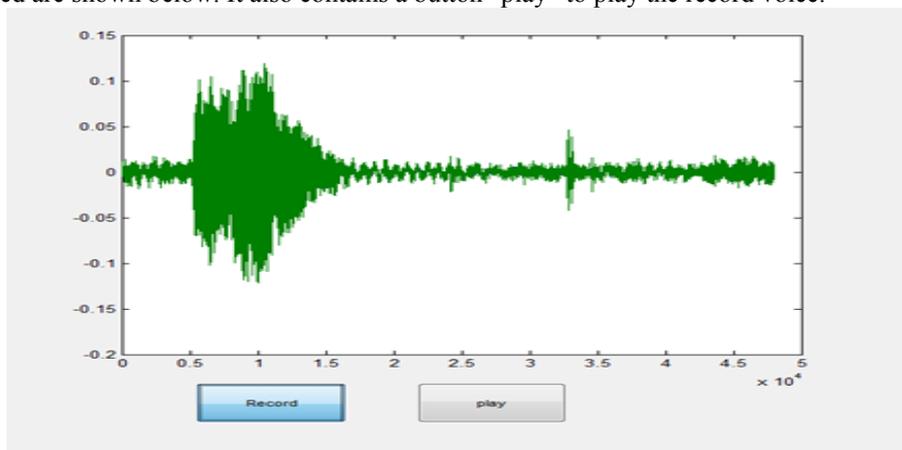


Fig 4: Recorded voice

Input signal were collected in computer using microphones on sample rate 16000 and 16 bit resolution. After recording train the data using some parameter. In neural network the average classification rate for training data and test data was obtained as as shown in Fig.7.

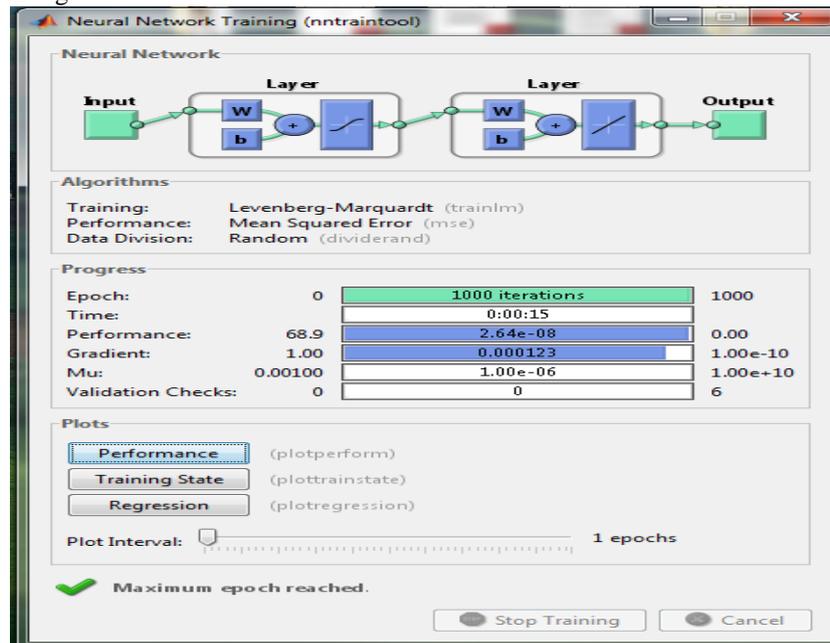


Fig 5 : Neural network training

In this paper, we have chosen to use a back propagation neural network since it has been successfully applied to many pattern classification problems including speaker recognition and our problem has been considered to be suitable with the supervised rule.

VI. CONCLUSION

The recognition accuracy of current speaker recognition systems under controlled conditions is high. However, in practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are necessary.. However, many research problems remain to be addressed, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores. Early applications of the technology have achieved varying degrees of success.

VII. FUTURE WORK

The promise for the future is significantly higher performance for almost every speech recognition technology area, with more robustness to speakers, background noises etc. This will ultimately lead to reliable, robust voice interfaces to every telecommunications service that is offered, thereby making them universally available. Text-dependent speaker recognition exists in the form of operational systems, but accurate text-independent speaker recognition remains a target.

REFERENCES

- [1] Ripul Gupta (2011), "Speech Recognition for Hindi," M.Tech Thesis, IIT Bombay.
- [2] Abhisek Paul(2011), "Speech Recognition in Hindi," M.Tech Thesis, National Institute of Technology, Rourkela
- [3] Q. Zhu and A. Alwan (2003), "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Computer Speech and Language*, vol. 17, no. 4, pp.381-402.
- [4] F. Jelinek, B. Meriardo, S. Roukos, and M. Strauss I, "A Dynamic Language Mode for Speech Recognition," IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598
- [5] Anusuya and Katti(2009), "Speech Recognition by Machine: A Review," *International Journal of Computer Science and Information Security*, Vol.6, No. 3, pp.181-205
- [6] Abdul Kadir K, (2010), "Recognition of Human Speech using q-Bernstein Polynominals," *International Journal of Computer Application*, Vol.2 - No.5, pp.22-28.
- [7] Reddy, R. (1976), "Speech Recognition by Machine: A Review," in proceedings of IEEE transaction, Vol. 64, No. 4, pp. 501-531.
- [8] Gaikwad, Gawali and Yannawar (2010), "A Review on Speech Recognition Technique," *International Journal of Computer Application*, Vol.10, No.3, pp.16-24.
- [9] Rohini B Shinde and V P Pawar (2012), "A Review on Acoustic Phonetic Approach for Marathi Speech," *Recognition. International Journal of Computer Applications* 59(2): 40-44.

- [10] Friesen, L. M., Shannon, R. V., Bas, kent, D., and Wang, X. (2001), "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.* 110(2), 1150–1163.
- [11] A. Mohamed, G. Dahl, and G. Hinton (2012), "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22.
- [12] L. Deng (2003), "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*, pp. 115–134. Springer-Verlag, New York
- [13] L. Deng and D. Yu (2007), "Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition," in *Proc.ICASSP*, pp. 445–448.
- [14] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, 17(1-2):91-108, August 1995.
- [15] Lan Wang, Ke Chen, and Huisheng Chi, "Capture Interspeaker Information With a Neural Network for Speaker Identification", *IEEE Transactions on Neural Networks*, 13(2):436-445, March 2002.
- [16] Toshihiro Isobe, Jun-ichi Takahashi, "A New Cohort Normalization Using Local Acoustic Information For Speaker Verification" *ICASSP*, 1999.
- [17] Sadaoki Furui, "Cepstral Analysis Technique for Automatic Speaker Verification" *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(2):254-272, APRIL 1981.
- [18] Zheng, N., Lee, T., 2007. Discrimination power of vocal source and vocal tract related features for speaker segmentation. *IEEE Trans. Audio, Speech Language Process.* 15 (6), 1884–1892.
- [19] Beigi, H. S. & Maes, S. S. (1998). Speaker, channel and environment change detection, *Proceedings of the World Congress on Automation (WAC1998)*.
- [20] Joseph P. Campbell "Speaker Recognition: A Tutorial" *IEEE*, VOL. 85, NO. 9, September 1997.
- [21] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (2004), "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *IEICE Trans. Inf. Syst. (Japanese edition)*, vol. J87-D-II, no. 8, pp. 1565–1571.
- [22] Müller, K.-R., Mika, S., Raetsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks* 12 (2), 181–201.
- [23] Moonasar, V., Venayagamoorthy, G., 2001. A committee of neural networks for automatic speaker recognition (ASR) systems. In: *Proc. Internat. Joint Conf. on Neural Networks (IJCNN 2001)*, Washington, DC, USA, July 2001, pp. 2936–2940.