Automatic Speech Recognition System for Hindi Utterances with Regional Indian Accents: A Review

[1]Abhishek Thakur, [2]Rajesh Kumar, [3]Naveen Kumar
[1,2,3]Dept. of ECE, IGCE, Abhipur, Mohali, Punjab, India

Abstract

This paper presents a study of automatic speech recognition system for Hindi utterances with regional Indian accents. In paper [3] we have designed matlab based ASR and control system for eight English key words by using simple rule base. This rule base algorithm is the beginning stage for Key Word recognition. In paper [4] we have designed Design of Hindi Key Word Recognition System for Home Automation System Using MFCC and DTW. Features of the speech signal are extracted in the form of MFCC coefficients and Dynamic Time Warping (DTW) has been used as features matching techniques. The recognition results are tested for clean and noisy test data. Average accuracy for clean data is 97.50 % while that for noisy data is 91.25 %. We face problem in noise environment to detect correct utterance now we are going to review different papers and find out different techniques to design our ASR control system for Hindi Key Words using MFCC and DTW in noise environment.

Keywords

Hindi Key words Recognition; Mel Frequency Cepstral coefficient (MFCC); Dynamic Time Wrapping (DTW); Automatic Speech Recognition (ASR); Artificial Neural Network (ANN); Hidden Markov Model (HMM); Support Vector Machine (SVM).

I. Introduction

Most people don’t like to press an untold number of buttons to accomplish a task. All appliances or other daily consumer devices are not used to its fullest extent. So it is only natural that if people could “talk” to these machines their lives would be even more comfortable. Most consumer appliances today have some sort of electronic or computer control, this feature prepares the way to realize the goal of “talking” to these appliances. Due to the rapid development in this field all over the world speech recognition systems have been implemented in a variety of applications, most eminent automated caller systems, automated information systems, speech recognition systems converting speech to text. These devices perform various tasks from simple user voice command. Such a wide application area brings frequent usage of such systems also in noisy environment.

In Bell Labs speech recognition work began in 50,s speech recognition system Audrey System developed first ten English digs. Research work in this field made good progress, and as an important issue in conducting research in the late 60’s the early 1970s. Further speech recognition in the 1980s, the HMM model and Artificial Neural Network (ANN) are successfully used in speech recognition. 1988 FULEE Kai and others use the VQ/I IMM method to achieve speaker independent continuous speech recognition system-SPHINX, including 997 vocabulary. This is the first of the world speech recognition system, it is a high performance, non-specific, large vocabulary continuous speech recognition system. People finally breakthrough of the three major obstacles, including a large vocabulary, continuous speech and non specific and it identified the mainstream of statistical methods. Speech recognition system developed as product in many developed countries such as the Microsoft, United States, South Korea, as well as IBM, Apple, Japan, AT&T and other companies [2].

A. Speech Production Process

For efficient extraction of features, it is necessary to understand the speech production mechanism in human beings. Then the conversion of the message into a language code takes place in which the talker converts the message into sets of phoneme sequences corresponding to the sounds that make up the words. Along with that the talker also determines the duration, loudness of sounds and pitch associated with it. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate and to shape the vocal tract such that the proper sequence of speech sounds is produced. Speech signal classifies into three possible ways depending on how speech sound you articulate:

1. Voiced Excitation

The glottis is closed. The air pressure forces the glottis to open and close periodically thus generating a periodic pulse train (triangle shaped). This “fundamental frequency” usually lies in the range from 80Hz to 350Hz.

2. Unvoiced Excitation

The glottis is open and the air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.

3. Transient Excitation

A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately.

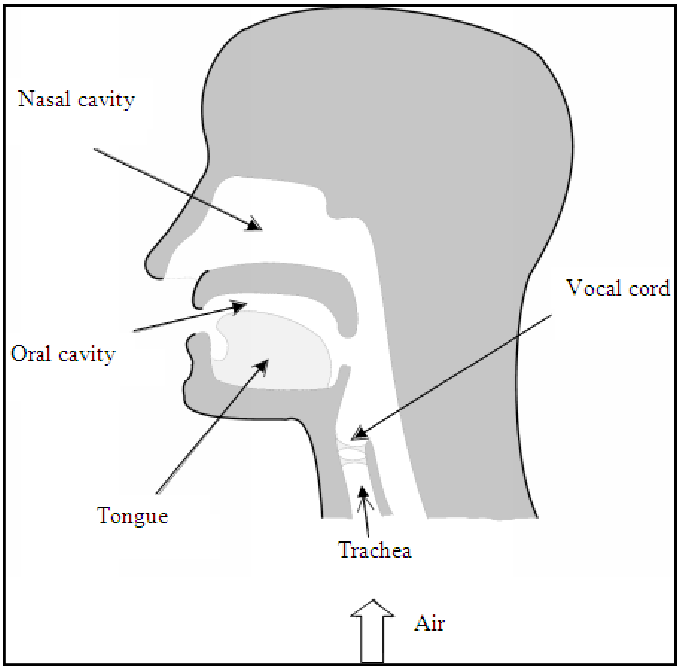

Fig. 1: Human Speech Production Process

With some speech sounds these three kinds of excitation occur in combination. The spectral shape of the speech signal is determined by the shape of the vocal tract. By changing the shape of the pipe (and in addition opening and closing the air flow through your nose) you change the spectral shape of the speech signal, thus articulating different speech sounds.

B. Classification of Speaker Recognition

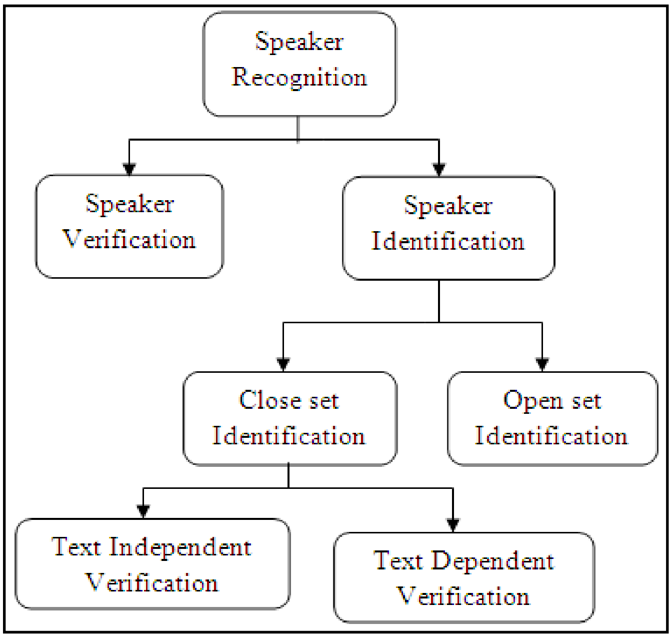

Fig. 2: Classification of Speaker Recognition System (Bibek Kumar Padhy, 2009)

Speaker recognition is of two types, speaker verification and speaker identification. Speaker verification task is to recognize the speaker voice. This process involves only binary decision about speaker voice. In speaker identification there is no such operation, only the input speech is given and the system finds out the matched samples from the known database. Speaker identification divided into two classes, e.g., open set and closed set speaker identification. In open set speaker identification the decision has to be made upon to whom the unknown speech sample match the most and if no satisfactory matching is found, the system also give result about the speech is not present in the data base. In closed set speaker identification the speaker's data has to be present inside the database. The system has to respond with the matched speaker.

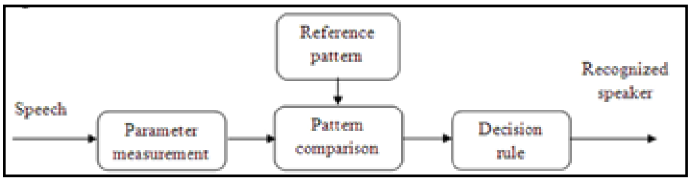

Fig. 3: Pattern Recognition Approach to Speaker Recognition

Fig. 3, depicts a systematic block diagram of a speaker recognition system using pattern recognition approach. The three basic steps in a pattern recognition model are (1) parameter measurement (in which a test pattern is created), (2) pattern comparison, and (3) decision making. The first step in a speaker recognition system, whether for identification or verification, is to build a model of the voice of each target speaker, as well as a model of a collection of background speakers, using speaker dependent features extracted from the speech waveform.

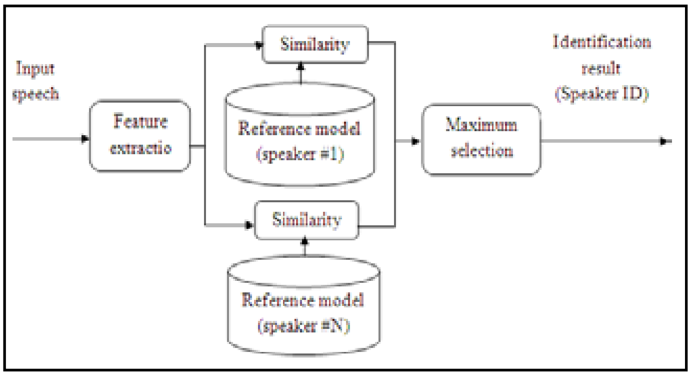
Input speech
Feature extractio
Similarity
Reference model (speaker #1)
Similarity
Reference model (speaker #N)
Maximum selection
Identification result (Speaker ID)

Fig. 4: Speaker Identification System

The process of speaker identification is divided into two main parts, e.g., the enrollment of Hindi key words and the identification of Hindi key words. During the enrollment of Hindi key words (training), speech samples are collected from the speakers, and they are used to train their models. In the identification of Hindi key words (testing), a test sample from an unknown speaker is compared against the speaker database [10]. In training and testing first step is common, i.e., feature extraction, where the speaker dependent features are extracted from the speech sample. The main purpose of this step is to reduce the amount of test data while retaining the speaker discriminative information. Then in the enrollment phase, these features are modeled and stored in the speaker database. In the identification step, the extracted features are compared against the stored models present in the speaker database. Based on results obtained from these comparisons the final decision about speaker identity is made. Fig. 4 and 5 gives overview of these components of a speaker recognition system for the verification and identification task.

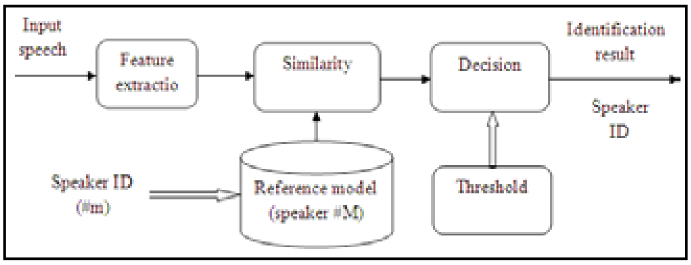

Fig. 5: Speaker Verification System

Speech recognition system is essentially a kind of pattern recognition system, including three basic units such as feature extraction, pattern matching, and reference model library. The unknown speeches is converted into electric signals through microphone, attached to the input of identification system, preprocessed first, then establish the model according to the characteristics of human speech sounds, analyze the input voice signal, and extract the desired characteristics. The speech recognition templates we need are acquired based on it.

II. Speech Recognition Technology

A. Detection of Hindi Key Words

The acoustic speech signal includes various kinds of information about speaker, e.g., "high level" properties such as dialect, context, speaking style, emotional state of speaker etc and also some "low level" properties such as pitch (fundamental frequency of the vocal cord vibrations), intensity, formant frequencies and their bandwidths, spectral correlations, short time spectrum and others

(Abid M. Jindani, 1998). The amount of data, generated during the speech production, is quite large while the essential characteristics of the generated speech changes quite slowly therefore, requires relatively less data to represent the characteristics of speech and the person who has spoken it. According to these matters feature extraction is a process of reducing data while retaining the speaker discriminative information of the speakers.

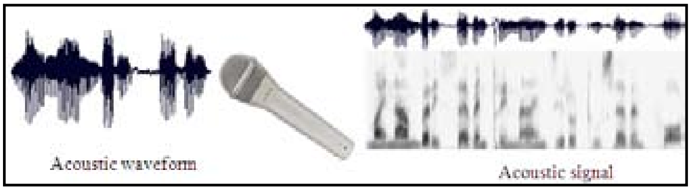

Fig. 6: Speaker Recognition Process

The steps required to make computers perform speech recognition are: Voice recording, word boundary detection, feature extraction, and recognition with the help of knowledge models. Word boundary detection or End Point Detection (EPD) is the process of identifying the start and the end of a spoken word in the given sound signal. While analyzing the sound signal, at times it becomes difficult to identify the word boundary. This can be attributed to various accents people have, like the duration of the pause they give between words while speaking. To generate the knowledge model one needs to train the system. During the training period one needs to show the system a set of inputs and what outputs they should map to. This is often called as supervised learning.

B. End Point Detection

The goal of end point detection is to identify the important part of an audio segment for further processing. Hence EPD is also known as “voice activity detection” or “speech detection”. EPD plays an important role in audio signal processing and recognition. There are two types of errors in EPD, which cause different effects in speech recognition, as follows.

- False Rejection: Speech frames are erroneously identified as silence/noise, leading to decreased recognition rates.
- False Acceptance: Silence/noise frames are erroneously identified as speech frames, which will not cause too much trouble if the recognizer can take short leading/trailing silence into consideration.

Principle

In speech recognition it is important to detect when a word is spoken. The system does detect the region of silence. Anything other than silence is considered as a spoken word by the system. The system uses energy pattern present in the sound signal and zero crossing rate to detect the silent region. Taking both of them is important as only energy tends to miss some parts of sounds which are important. In the above example, the volume threshold is determined as:

volTh=(volMax-volMin)/epdPrm.volRatio+volMin; (1)

Where epdPrm.volRatio is 10, and volMin and volMax are located at indices of 1/32*length(volume) and 31/32*length(volume), of the volume vector after sorting. The ratios or constants in the above four methods should be determined through labeled training data. It should be noted that wave files of different characteristics (recordings via unidirectional or Omni directional microphones, different sample rates, different bit resolutions, different frame sizes and overlaps) will have a different best thresholds. Create a new threshold by using linear combinations of these thresholds, etc.

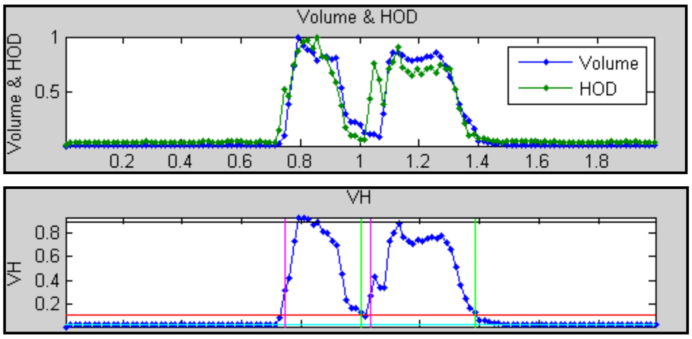

Fig. 7: Shows the Energy Pattern for the Same Word

From the above example, it is obvious that the leading unvoiced sound is likely to be misclassified as silence. Moreover, a single threshold might not perform well if the volume varies a lot. As a result, an improved method can be stated next:

1. Use an upper threshold tu to determine the initial end-points.
2. Extend the boundaries until they reach the lower threshold tl.
3. Extend the boundaries further until they reach the ZCR threshold tzc.

The above improved method uses only three thresholds, hence it is possible to use grid search to find the best values via a set of labeled training data. Now it should be obvious that the most difficult part in EPD is to distinguish unvoiced sounds from silence. One way to achieve this goal is to use high order difference of the waveform as a time domain features. It is obvious that the high order difference on the waveform can let us identify the unvoiced sound more easily for this case. Therefore we can take the union of high volume and high HOD regions to have most robust of EPD.

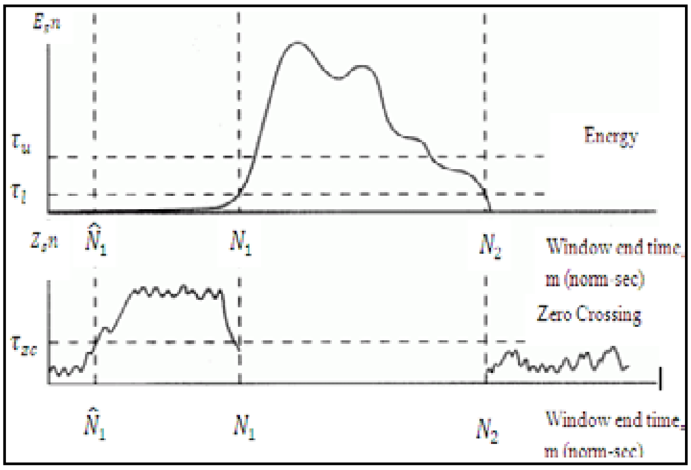

Fig. 8: Improved Methods for End Point Detection

C. Feature Extraction

Feature Extraction refers to the process of conversion of sound signal to a form suitable for the next stages to use. Feature extraction may include extracting parameters such as amplitude of the signal, energy of frequencies, etc. Recognition involves mapping the given input (in form of various features) to one of the known sounds. This may involve use of various knowledge models for precise identification. Knowledge models refer to models which help the recognition system. Mel-frequency cepstrum coefficients (MFCC) are well known features used to describe speech signal. They are based on the known evidence that the information carried by low frequency components of the speech signal is phonetically more important for human perception than carried by high-frequency components. Technique of computing MFCC is based on the short term analysis, and thus from each frame a MFCC vector is computed. MFCC extraction is similar to the cepstrum calculation except that one special step is inserted, namely the frequency axis is warped according to the Mel scale. Summing up, the process of extracting MFCC from continuous speech is illustrated in fig. 9.

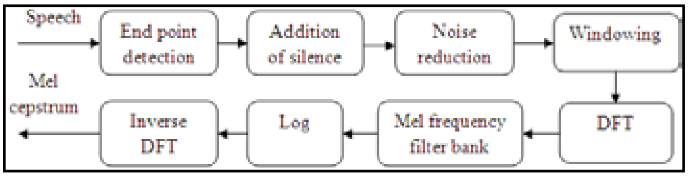

Fig. 9: Block Diagram of Mel Frequency Cepstral Coefficient (MFCC)

The speech signal can be represented as a “quickly varying” source signal convolved with the “slowly varying” impulse response of the vocal tract represented as a linear filter. Only the output (speech signal) is accessible and it is often desirable to eliminate the source part. The source and the filter parameters are convolved with each other in time domain. The cestrum is a representation of the signal where these two components are resolved into two additive parts. It is computed by taking the inverse DFT of the logarithm of the magnitude spectrum of the frame. The conversion from time domain to frequency domain changes the convolution to the multiplication. On that use of logarithm replaces all the multiplication steps by the addition. The inverse DFT is applied on it, which operate individually on quickly varying and slowly varying parts of speech signal.

D. Feature Matching (DTW) Scores

Dynamic time warping is an efficient algorithm to find a non-linear alignment path between two sequences that optimizes their distance. It is obtained by time scaling one of the signals non linearly so that it aligns with the other. It is an extremely efficient time series similarity measure. It minimizes the effects of shifting and distortion in signals allowing elastic transformation in time. It is one of the earliest approaches to isolated word recognition. For keyword recognition, a prototype of the keyword is stored as a template and compared to each word in the incoming speech utterance.

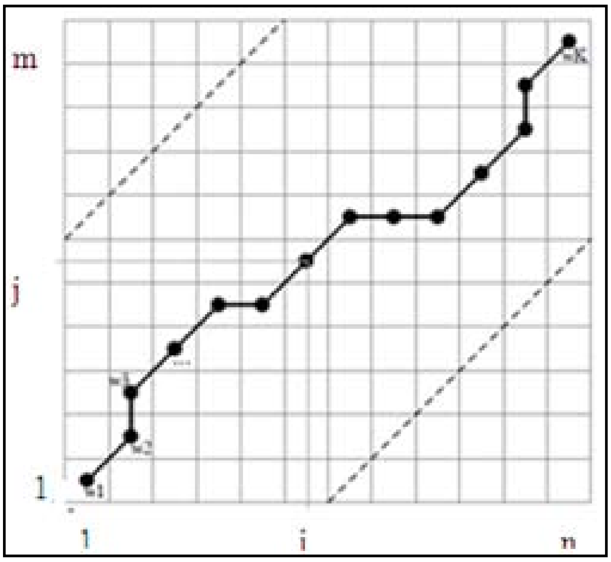

Fig. 10: Dynamic Time Warping.

As shown in fig. 10, the feature vector for the reference keyword and input keyword are arranged along the two sides of the grid. In this case, the reference template of length n is arranged along the horizontal axis and the test word of length m along the vertical axis. Each block in the grid is the distance between corresponding feature vectors. The best match between these two sequences can be computed from the path through the grid which minimizes the total cumulative distance between them as shown by the dark line in fig. 10. Total distance between the test and the reference data is the sum of distance along the path. From fig. 10, it is apparent that the number of possible paths through the grid grows exponentially with the length of the word. Applying some constraints (Sakoe, et al.1978), the possible paths can be limited to a certain limit making the computation feasible. These constraints basically limit the possibility of infinite paths making computation more efficient.

For connected word recognition, where it is hard to precisely indicate the word boundary, before mentioned boundary condition is changed to cover some range of possible beginnings and endings of a word. In speech recognition, we have to classify not only single vectors, but sequences of vectors.

DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. To align two sequences using DTW, an n-by-m matrix where the (ith, jth) element of the matrix contains the distance $d(q_i, c_j)$ between the two points qi and c_j is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation as shown in “Eq. (2)”.

$$d(q_i, c_j) = d(q_i, c_j)^2 \qquad (2)$$

Each matrix element (i, j) corresponds to the alignment between the points qi and cj. Then, accumulated distance is measured by “Eq. (7)” [2].

$$D (i, j) = \min[D(i-1, j-1),D(i-1, j),D(i, j -1)]+ d(i, j) \qquad (3)$$

When we combine these MFCCs they produce a unique voice print for that word. With the help of dynamic time wrapping algorithm we can find out minimum global distance between reference and stored vectors. These features are extracted and store in a vector for comparison with refarance key word in DTW. DTW algorithm used to find out minimum global distance between stored and reference key word. The speech recognition system is essentially a pattern recognition system, including feature extraction, pattern matching and reference model library.

E. Related Search Technology

When basic morphemes are acquired, we can search according to the relation between morphemes, and determine which morphemes is a sense group and how to arrange the morphemes. Using correlation to determine the meaning group is very important, because the words are not arbitrary collocation but have rules which called grammar, including grammar of written and spoken [6]. The grammar is rules of speech recognition, so it is very important. Association mode with the grammatical constraints has some species such as probability of related word appears and context repetition rate etc. Connected word recognition means HMM stored in the system is directed at isolated word, but the speech is composed by these words. Since the technique is a connected sequence, namely according to pronunciation sequence to find its best matching reference module word, we should consider how to solve some problems such as sometime we even know word length range but we do not know the word number in sequence, or in addition to the first and the end point of a sequence, each word boundary position we does not know.

Representative speech recognition methods include dynamic time warping (DTW), hidden Markov model (HMM), vector quantization (VQ), artificial neural network (ANN), support vector machine (SVM) and so on. The article focuses on two methods of hidden Markov model (HMM) and artificial neural network (ANN).

F. Hidden Markov Model (HMM)

The HMM model parameters represent the time varying characteristics of the voice signal. It consists of two interrelated stochastic processes common to describe the statistical characteristics of the signal. One of which is hidden (unobserved) finite state Markov chain, and the other is the observation vector associated with each state of the Markov chain stochastic process (observable). Reveal characteristics of the hidden Markov chain depends on the signal characteristics can be observed [11]. In this way, a certain period of time varying signals such as voice characteristics described by the random process corresponding to the symbols of state observation. Signal described by the hidden Markov chain transition probability changes with time. HMM model in a state j under the corresponding observed values by a set of probability bik, k = 1, 2, & M to describe it is one of the M discrete countable observations, and thus known as the discrete the HMM. When the observed value of a continuous random variable X, its corresponding observed values in the state j observed by a probability density function bj(X), which became continuous HMM. Continuous HMM using the Baum Welch algorithm to estimate model parameters applied in the estimation of, A parameter, but the description in the estimation of bj(X) parameter must be a certain limit can be established. Current most widely used is the Gaussian bj (X) it can be represented using the following formula [8]: Among them, the N(X, jk, jk) for multi-dimensional Gaussian probability function, jk mean vector jk side difference matrix, k is the bj(X) the number of mixed probability, cj(X)is the combination coefficient, and HMM is a more complete expression of acoustic model of the voice and it uses statistical methods of training the underlying acoustic model and the upper voice model into the unified voice recognition search algorithm can obtain better recognition results, and can be used for continuous speech recognition, but the drawback is the need to be very sophisticated calculations and a longer training sequence[7].

G. Artificial Neural Network (ANN)

Artificial neural network ANN (based Artificial Neural Networks), analogous to the way biological nervous systems process information, using a large number of simple processing units connected in parallel to form a complex information processing system. This system has the training, highly parallel, rapid judgment, fault tolerance features applies voice signal processing. Speech recognition neural networks are usually divided into two categories, a class of neural networks or neural networks with the traditional HMM the DP combination of hybrid network, the other is the establishment of the auditory neural network model based on human auditory physiology, psychology research. Neural network model that more commonly used and has the potentiating of speech recognition mainly include single layer perception model, multi-layer perception model, Kohonen self organizing feature map model, radial basis function neural network , predictive neural network etc. In addition, in order to make the neural network reflects the dynamic of the speech signal time-varying characteristics, delay neural network, recurrent neural network and so on. Artificial neural network technology in voice recognition applications mainly the following aspects:

- Reduce the modeling unit, generally in the phoneme modeling to improve the recognition rate of the entire system by improving the recognition rate of phonemes.
- Depth study of the acoustic model, the auditory model, the brain operation mechanism, the introduction of context information, in order to reduce the impact of changes in voice more than the speech signal.
- Extracted from the speech signal in a variety of features, a hybrid network model (HMM + NN), and apply a variety of knowledge sources (phonemes, vocabulary, syntax and meaning of the word), for voice recognition to understand the research, to improve system properties [5].

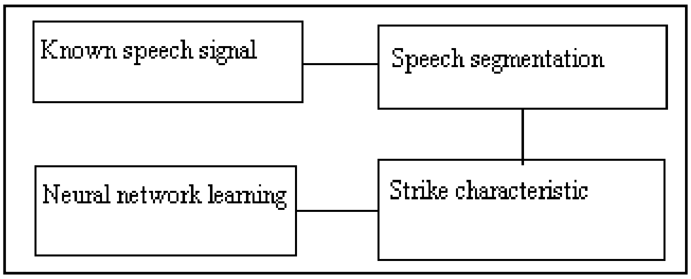

Fig. 11: Artificial Neural Network Learning System

Speech recognition using artificial neural network technology, including e-learning process and the speech recognition process, shown in fig. 11. The network learning process is to known speech signal as a learning sample, self-learning neural network, and ultimately a set of connection weights and bias. The speech recognition process is to test the voice signal as network input, the recognition results obtained through the network of associations. The key of these two processes is to strike a speech characteristic parameters and neural network learning.

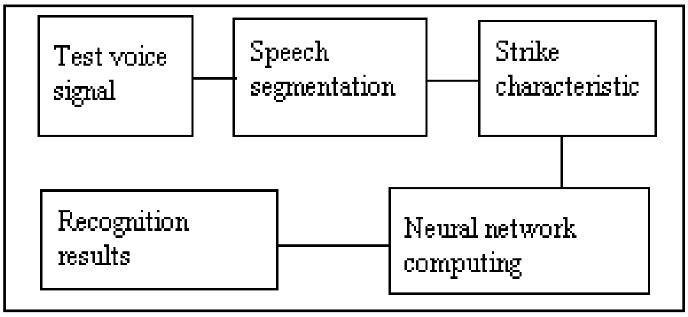

Fig. 12: Artificial Neural Network Speech Recognition System

The application of artificial neural networks in the field of speech recognition has been greatly developed in recent years, artificial neural networks in speech signal processing can be divided into the following areas: firstly, improve the performance of artificial neural networks. Secondly, artificial neural network has been developed method combines a hybrid system. Thirdly, explore the use of newly emerging or widespread concern mathematical methods constitute the unique nature of the neural network, and applied to the field of speech signal processing [6]. The application of artificial neural networks in speech recognition has become a new hotspot. Artificial neural network technology has been successfully applied to solve pattern classification problems, and was shown to have enormous energy, we can predict that in the last decade, artificial neural network-based speech recognition system products will appear in the market, people will adjust their own way of speaking to accommodate a variety of recognition system.

III. Conclusion and Future Work

The main contribution of this study is that it presents the idea of automatic speech recognition system for Hindi utterances with regional Indian accents. The study shows that these approaches can be used to increase accuracy of noisy test data for automatic speech recognition system with regional Indian accents. Our effort will be directed toward developing the more appropriate and convenient method.

References

[1] J. Rajnoha, P. Pollak,"ASR Systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness", Radioengineering, Vol. 20, No. 1, April 2011.

[2] Ren Tianping,"Application of speech recognition technology [J]. Henan Science and Technology, 2005.

[3] Er. Abhishek Thakur, Neeru Singla,"Design of Matlab-Based Automatic Speaker Recognition and Control System", International Journal of Advanced Enginggring Sciences and Technologies, Vol. 8, Issue No. 1, pp. 100-106, 2011.

[4] Er. Abhishek Thakur, Neeru Singla, V.V. Patil,"Design of Hindi Key Word Recognition System for Home Automation System Using MFCC and DTW", In: International Journal of Advanced Enginggring Sciences and Technologies, Vol. 11, Issue No. 1, pp. 177-182, 2011.

[5] Yin Peng, Li Tao, Wang Haibing.Intelligent neural network system composed of the principle in speech recognition. Mini-Micro Systems, 2000, 21(8), pp. 836-839.

[6] Jiang Ming Hu,"In the Yuan Baozong, Lin Biqin. Neural networks for speech recognition research and progress", Telecommunications Science, 1997, 13(7), 1-6.

[7] Zhang Ping, Zhang Qiong,"Based on HMM and BP neural network for speech recognition", Cross-century, 2008.

[8] L A Liporace,"Maximum Likelihood for Multivariate Observation of MarkovSources", IEEE.Trans. IT, 1982, 28(5), pp. 729-734

[9] K. X. Huang, A. Acero, H. Wuenon,"Spoken Language Processing: A Guide to Theory", Algorithm and System Development, Pearson, 2005.

[10] Abid M. Jindani,"Speaker Independent Real-Time Speech Recognition System", pp. 1-69, August, 1998.

[11] Bengt J. Borgstrom,"HMM-Based Reconstruction of Unreliable Spectrographic Data for Noise Robust Speech Recognition", IEEE Transactions on Audio and Language Processing, Vol. 18, No. 6, pp. 1612-1623 August 2010.

Abhishek Thakur is working as Assistant Professor at Indo Global College Of Engineering, Abhipur, Mohali, Punjab. He has completed his M.Tech from Rayat Inst. Of Engg. & Information Tech., Punjab, India and B.Tech from Dr. J.J. Magdum College of Engineering, District Kolhapur, Maharashtra, India. He has 3 years of academic experience and 1 year Industry experience. He has authored research papers in reputed International Journals, International and National conferences. His areas of interest are Image & Speech Processing, Wireless Communication.

Rajesh Kumar is working as Associate Professor at Indo Global College Of Engineering, Mohali, Punjab. He is pursuing Ph.D from NIT, Hamirpur, H.P. and has completed his M.Tech from GNE, Ludhiana, India. He completed his B.Tech from HCTM, Kaithal, India. He has 11 years of academic experience. He has authored many research papers in reputed International Journals, International and National conferences. His areas of interest are VLSI, Microelectronics and Image & Speech Processing.

Naveen Kumar is Assistant Professor at Indo Global College of Engg., Mohali, Punjab, India. He is pursuing M.E. from National Institute of Technical Teachers' Training & Research (NITTTR), Chandigarh. India. He has completed B.Tech from SVIET, Mohali (Punjab), India in the year 2009. He has 2 years of academic experience. His areas of interest are Wireless & Mobile communication, Antenna Design.