Speech Recognition: A Complete Perspective

Ashok Kumar, Vikas Mittal

Abstract: Speech is basic form of communication between human beings. Speech recognition is a process to convert speech sound to corresponding text. Speech recognition technology has been developed to a large extent in last few years. But still there exist many important research challenges e.g. speaker and language variability, environmental noise and the vocabulary size etc. The objective of this paper is to present a complete perspective on speech recognition describing various processes and summarizing various methods used in a typical speech system.

Index Terms: Speech recognition, Feature extraction, Modeling, Speech processing, Training & Testing.

I. INTRODUCTION

Speech recognition, popularly also known as Automatic Speech Recognition (ASR) is the process of converting speech signal to a sequence of words by means of an algorithm implemented as a computer program. Speech processing is one of the major fields of signal processing. Speech recognition area aims at to develop techniques for speech input to a machine [1]. The early computer systems were limited in scope and power. But the revolution in computer technology has evolved the field of automatic speech recognition. Now a day's it's easy to store huge database for speech recognition due to advancement in computer technology. Language is basic medium for communication, so it's better to expect human computer interfaces in native languages [2]. There are only few languages on which speech recognition systems have been developed. So a lot of scope is there to build speech recognizers in native languages [3]. Advancement in statistical modeling of speech has gained a widespread application in the field of speech recognition. Automatic speech recognition has reduced the human efforts in many fields, such as automatic call processing in telephone networks, data entry, voice dictation; query based updated travel information and reservation, natural language understanding and translators etc. Speech recognition technology has also an extensive use in telephone networks to automate and enhance the operator services. [4], [5], [6]. This paper highlights the basic building blocks of speech recognition systems, technological progression and problems in path of automatic speech recognition.

II. SPEECH RECOGNITION SYSTEM

The speech sound is captured using microphone to convert it in to electrical signal. The purpose of sound card inside the computer is to change analog signal into digital signal. Sound card has capabilities to store and play this speech signal. There are following building blocks for general speech recognition system [7].

A. Signal preprocessing
B. Feature extraction
C. Language model
D. Decoder
E. Speech Recognition

A. Signal Pre-processing

Speech signal captured by microphone, telephone etc. are analog in nature so it is required to be digitized as per Nyquist theorem. This theorem states that a signal is to be sampled more than twice the rate of highest frequency present in it. Generally sampling frequencies for speech signal are 8 KHz and 20 KHz. For telephonic speech signal it is recommended to have 8 KHz sampling rate while 16 KHz is generally used for normal microphones [8].

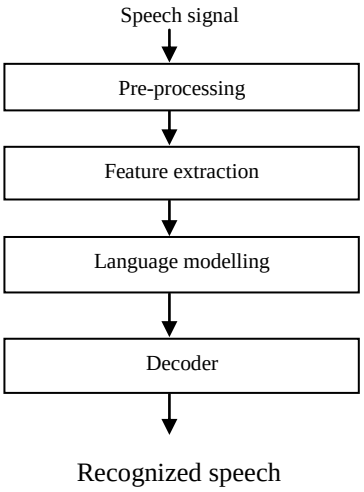

Fig. 1. Basic working of Speech Recognition System

B. Feature extraction

Feature extraction is used to find a set of properties that are stable and acoustically correlated to each other. So it is a type of parameterization of speech signal.

Such parameters can form the observation vectors. The goal of feature extractor is to identify relevant information for purpose of accurate classification.

Revised Manuscript Received on December 22, 2018.

Ashok Kumar, Electronics and Communication Engineering, NIT Kurukshetra, Kurukshetra, India.
Vikas Mittal, Electronics and Communication Engineering, NIT Kurukshetra, Kurukshetra, India.

Figure 2 shows the role of feature extraction in speech recognition.

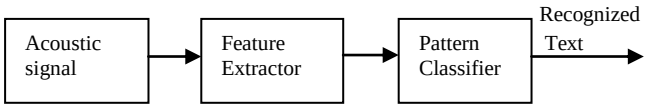

Fig. 2. Role of feature extraction in speech recognition

Linear predictive Cepstral coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC) and PLP are some of the commonly used feature extraction techniques [9]. The most widely used feature extraction techniques are discussed below:

a. Linear Predictive Coding (LPC)

It is one of the best signal analysis techniques to estimate the basic parameters of speech. The basic concept behind this technique is that present speech samples can be approximated by past samples by following a linear combination. A unique set of feature vectors can be obtained by minimizing the sum of squared differences between actual speech samples and predicted values [3]. Figure 3 shows the steps to extract the feature in LPC.

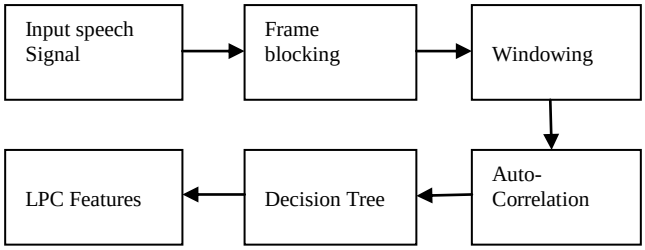

Fig. 3. Steps to extract LPC Features

b. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC features are widely used in speech recognition, due to logarithmic positioning of its frequency bands. It approximates the human auditory, more as compared to the other techniques. In order to extract the MFCC coefficients the signal at input, pass through the hamming window to minimize the discontinuities of signal.

Then DFT is used to generate Mel filter bank. The width of triangular pulses varies according to the Mel frequency warping.

Finally Inverse Discrete Fourier Transform (IDFT) is used for calculation of Cepstral coefficients. Mel frequency can be computed by following formula [3],

$$\text{Mel}(f) = 2595 \log_{10}(1+f/700)$$

Where f is frequency and Mel (f) is Mel frequency

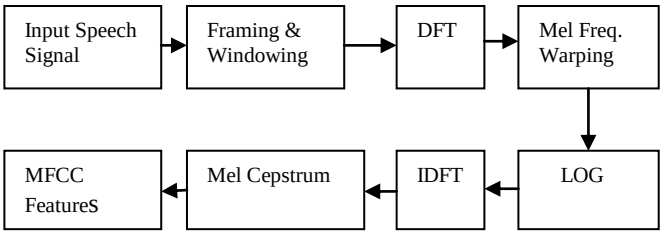

Fig. 4. steps involved in MFCC feature extraction.

c. Perceptual Linear Prediction (PLP)

Perceptual Linear prediction (PLP) demonstrated an improvement over the LPC due to its three principal characteristics derived from the psycho-acoustic properties of human hearing like spectral resolution of critical band, equal loudness curve and intensity loudness power law [10].

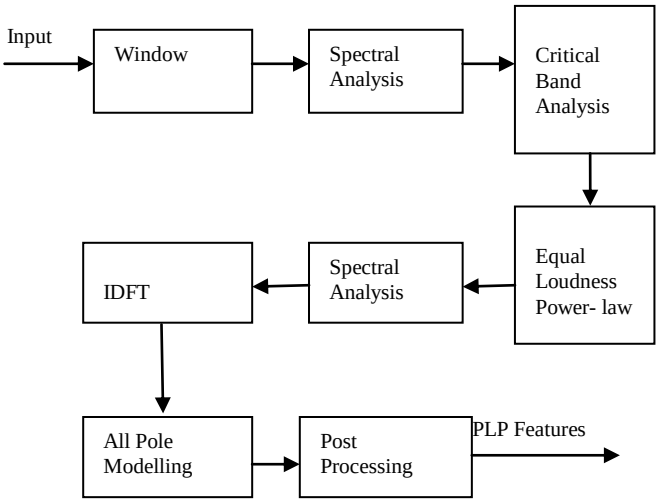

Fig. 5. Block diagram of PLP feature extraction approach

C. Language modeling

Language modeling is used to find the correct word sequence by predicting nth words using (n-1) preceding words. Language modeling is of various types.

- Uniform model: where occurrence of each word is equally probable.
- Stochastic model: probability of present word depends on probability of word preceding it.

Finite state languages: This language modeling use finite state network to define allowed word sequence.

Context free grammar (CFG): It is used to encode allowed sequence of words in speech recognition system [11].

CFG follows a mechanism for defining languages (sets of acceptable sentences).

It is powerful as compared to Finite state grammars (FSG) due to imposing more structure on sentences. Human languages are actually close to the

- CFGs so it has application in speech recognizers. Generally, a CFG is defined by following relation,
- A finite sets of terminal symbols (Defines words in vocabulary).
- A finite set of non-terminal symbols to define concepts.
- A special non-terminal symbols, represents the CFG.

Then a finite set of production rules are applied to detect a terminal or a non terminal symbols in data sequence.

So, CFG is used to model a language in speech recognition by expanding its special non-terminal symbols after application of production rules.

D. Decoder

This stage is involved to find most likely word sequence for the given observation sequence. Generally dynamic programming algorithms are used to solve this problem. The purpose of these algorithms is to search single path through the network to have best match for the given sequence, Viterbi algorithm is mostly used for this purpose. In case of large vocabulary, a beam search method is useful for Viterbi iteration [12].

E. Speech Recognition

Speech recognition is completed in two phase: Training and testing phase. Training phase is just similar to identification of objects. It may be repeated many times for better recognition which improves the performance while testing. Testing phase includes the comparison between reference pattern scored while training, and spoken words at the time of testing. The extent of closeness in these two phases counts for improving the performance of the system. But variability encountered during recognition effects a lot for constant recognition rate [13].

III. TYPES OF SPEECH RECOGNITION

Speech recognition systems can be of various types depending on types of utterances to be recognized. These various types are classified as follows:

A. Isolated Words

Isolated word recognizers generally obtain each utterance to have quiet on both sides of sample window. These systems usually have two states Listen/Not- Listen states, where speaker has to wait between two utterances. The pauses between utterances are used for processing speech signals.

B. Connected Words

Connected words are similar to isolated words with only difference of minimal pause between them.

C. Continuous Words

Continuous speech recognition involves almost natural way of speaking. It is difficult to design continuous speech recognizers because it considers special methods to determine utterance boundaries [14].

D. Spontaneous Words

Spontaneous speech covers the mispronounced, unrehearsed non-words with false statements which are difficult to read [3]. An ASR system under this category is to handle a variety of features such as words being run together such as “ums” and “ahs” [15].

IV. METHODOLOGIES OF SPEECH RECOGNITION

ASR methodologies can be usually classified in to three categories namely, Acoustic–Phonetic approach, pattern recognition approach and Artificial intelligence approach.

A. Acoustic-Phonetic Approach

Acoustic Phonetic approach is a rule based approach of speech recognition. According to this approach there exist finite, distinctive phonetic units in spoken speech utterances. These acoustic properties are present in speech signal. Acoustic –Phonetic approach involves the spectral analysis of speech signal to extract set of features for segmentation and labeling of speech signal into stable acoustic regions. In this way a valid word from segmentation to labeling is produced [16].

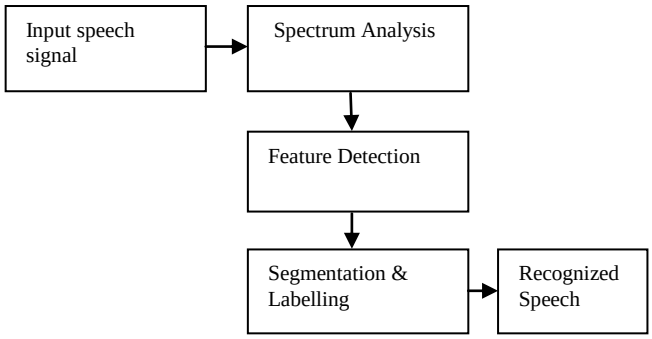

Fig. 6. Block diagram of Acoustic-Phonetic modeling approach

B. Pattern recognition Approach:

This approach is completed in two steps. First step is used to train speech recognizers with each possible pattern. The second stage provides a direct comparison between unknown speeches to the pattern learned in training stage.

This approach have well defined mathematical framework which maintain consistent speech pattern to be recognized. The pattern recognition can have any of the form, either a speech template or a statistical modeling [16].

C. Artificial Intelligence Approach:

It is a type of hybrid approach that exploits the idea of Acoustic phonetic approach and pattern recognition. Generally pattern matching is based on dynamic time warping (DTW) and hidden markov models [17-18].

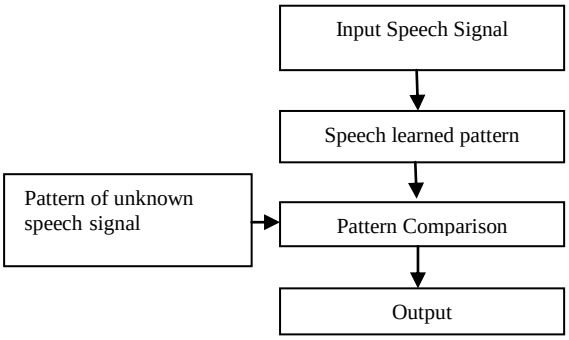

Fig. 7. Block diagram of pattern recognition approach

In DTW speech recognition works with classes. Each class can be represented by one or more templates. As the number of templates increases it improves the system modeling. In advanced system hidden markov model (HMM) is preferred over DTW, due to improved and lower memory requirements. This approach is used for complex tasks but it is not so efficient when data set are large. Phoneme recognition is basic approach of artificial neural networks. This is done by technique of intelligence, analyzing and visualizing the input speech. The network contains a large number of neurons. Each neurons counts for nonlinear weight of inputs and then send result to outgoing units. Training sets obtained, this way helps to assign values to input and output neurons.

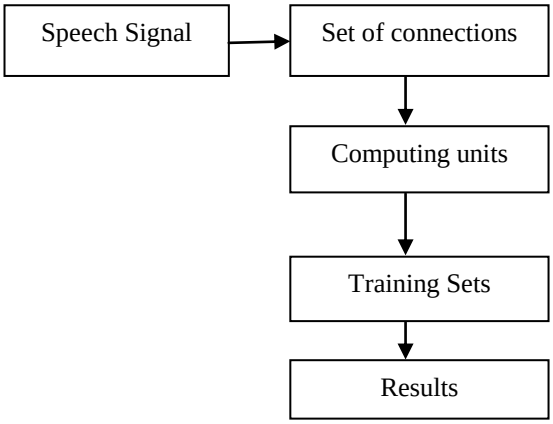

Fig. 8. Block diagram for artificial neural network approach

D. GENERATIVE LEARNING APPROACH (HMM- GMM)

Generative learning provides Gaussian-Mixture model based hidden markov models for conventional speech recognition systems. Speech signal is a short time stationary signal so it can be better represented by markov model. A GMM-HMM is parameterized vector with prior probabilities and state transition probability matrix [16-20].

It is preferred to use hidden markov model to achieve better performance in advance speech recognition systems. HMM have better control on data sequence length as per variation in word sequence and accent.

E. Discriminative Learning (HMM-ANN)

Generative learning is used to train a generative model. The neural network was used in the form of Multilayer Perceptron (MLP) in 1990's. The MLP output provides the conditional probability. The hybrid MLP-HMM can be created by feeding MLP as input to an HMM. In generative HMM general feature vectors of MLP is used in combination with traditional feature [21].

Discriminative training using neural networks provides a natural and efficient manner to estimate the probabilities of a speech segment. This type of approach have most effectiveness for the small types of speech units like isolated words and phonemes but hardly used for continuous speech recognition [22-23].

F. Deep Learning (HMM-DNN)

Deep learning or unsupervised feature learning is new to the area of machine learning. It is the latest technology for speech recognition that has replaced the older one in all sense. It has generally three types of deep architectures, generative, discriminative and hybrid. The first type is used to provide high order correlation properties of data. The second type is intended to provide discriminative power for pattern classification and characterizing the posterior distributions of class labels. In third type goal is achieved through the composite output of first and second deep architectures [24-26].

V. ADVANCEMENT IN SPEECH RECOGNITION

This section provides a review of progression in speech recognition technology. There is also a brief discussion of various approaches applied to improve the recognition system. The survey of last few years in concerned technology reported a considerable development in this field.

The earliest efforts to develop Automatic speech recognition were made in 1950s when many researchers tried to explore the basic idea of acoustic-phonetic [27]. In 1952 Davis et.al (1952), devise to build a recognizer for isolated digit recognition. The proposed system was utilizing the concept of spectral resonances to show the vowel region of a digit [28]. Olson and Belar (1956), tried to detect 10 distinct syllables of a single speaker embodied in 10 monosyllabic words [29]. In 1959, Fry and Denes tried to build a phoneme recognizer to recognize vowels and consonants by analysis of spectrum and pattern matching [30].

In 1960, hardware based approach came into existence, when several Japanese laboratories had entered in this field. Suzuki and Nakata (1961), developed a hardware for vowel recognition [31]. Sakoe and Chiba in Japan proposed the importance of dynamic programming to align a pair of speech [17].

In 1970, Pattern based speech recognition for isolated words were the key focus for the researchers. Linear Predictive Coding (LPC) had been efficiently used to code low bit rate speech coding and it had also application to use by speech recognition systems by optimizing its spectral parameters [32].

In 1980, speech recognition was focused on detection of continuous word.

Hidden markov model (HMM), in addition with mixture densities ensured satisfactory results in terms of accuracy and performance [33, 34].

In 1990, a hybrid approach was used to enhance the performance of recognition by integration of neural networks and HMMs [35]. Pruthi et al. (2000), developed a real time speaker- dependent isolated word recognizer. Continuous HMM was used to recognize the isolated word of Hindi language in 2006 designed by Gupta [11].

Al- Qatab et. al (2010), implemented an Arabic speech system using HTK capable to recognize both isolated and continuous words [11].

R. K. Aggarwal and M. Dave (2011), proposed a Hindi speech recognition using Gaussian mixtures that exhibits maximum accuracy [36].

Z. Yu et. al (2014), presented the teaching experiment of speech recognition based on Hidden Markov Model(HMM) that describes HMM speech recognition principle, implementation process to realize a speech recognition system [37].

M. Agrawal and T. Raikwar (2016) discussed the role and importance of signal processing techniques in automatic speech recognition [15].

K. A. Qazi (2018), discusses a hybrid technique for speech segregation and classification using a sophisticated Deep Neural Network (DNN). The proposed method tested on different datasets shows robustness for speech segregation [38].

VI. PROBLEMS IN SPEECH RECOGNITION

The performance of speech recognition system depends on its inertness to surrounding variabilities. There are many factors which are responsible for an effective recognition rate like environment conditions, speaker dependent/ independent, rate of speech and channel variability. But for advanced speech recognition system it is required to develop adaptive algorithm to match the change occurred and auto generative modeling capabilities to fill the gaps.

VII. AREA OF APPLICATIONS

In last few years, Speech recognition has gained a wide approval. This wide acknowledgement is resulted due to the revolution in storage technology to handle big data like voice search on Google and other voice enabled interaction with mobile devices. So speech recognition has wide applications in many fields like voice user interface, domestic appliance control, voice dialing, voice enabled search, data entry, learning applications for handicapped peoples [9].

VIII. TOOLS FOR SPEECH RECOGNITION

There are many open- source toolkits available online to form a speech recognition system. Few of them are: HTK, Julius, Sphinx and Kaldi.

Some other open-source speech recognition kits are also available that don't have the wider use like Segmental Conditional Random Field Toolkit for speech recognition (SCARF) and SHoUT speech recognition toolkit [9].

IX. PERFORMANCE EVALUATION

The performance of speech recognition system is generally measured in terms of accuracy and speed. Accuracy of the system is considered in terms of Word error rate (WER). So performance defined in Word recognition rate (WRR) is a complimentary part of Word error rate (WER). Word errors can be divided into number of insertions, substitutions and deletions while recognizing a speech. These two performance factor can be evaluated by following equations,

$$\text{Word Error Rate (WER)} = I+S+D/N$$

Where I is number of insertions,
S is number of substitutions,
D is number of deletions and
N is number of words in utterance.

$$\text{Word Recognition Rate (WRR)} = 1\text{-WER}$$

For speed, Real time factor is defined; computed by following equation,

$$\text{Real Time Factor (RTF)} = T/D$$

Where T is processing time and D is duration.

X. CONCLUSION

Speech is basic mode of communication between human beings, so a feasible interface is required to connect human with machines. Although this field has gained a wide approval to automate the services and applications but there are several parameters which affect the accuracy and efficiency of speech recognition system. The most of speech variability involves speech rate, environmental conditions, channel and context of utterance. Robustness of speech system depends on some stable parameters/ features of speech signal. To enhance the power of speech recognition system, it is required to design speech recognizers in local languages. Multilingual is new evolving field in area of speech recognition. There is a lot of development and research in the field of foreign languages but to enhance its power and utility for native people, it's essential to use this technology in native languages.

REFERENCES

1. S. K. Gaikward et.al, "A review on speech recognition technique", International journal of Computer Applications, vol. 10, no. 3, November 2010.
2. K. Samudravijaya, "Speech and Speaker recognition tutorial" TIFR Mumbai 4000005.
3. S. Naziya S. and R. R. Deshmukh, " Speech Recognition System- A Review", IOSR Journal of Computer Engineering (IOSR-JCE), vol. 18, issue 4, (Jul.-Aug. 16), pp. 01-09.
4. W. M. Campbell, D. E. Sturim et.al, "The MIT- LL/IBM speaker recognition system using high performance reduced complexity recognition" MIT Lincoln Laboratory IBM 2006.
5. K. Brady, M. Brandstein et.al, "An evaluation of audio-visual person recognition on the XM2VTS corpus using the Lausanne protocol", MIT Lincoln Laboratory, 244 Wood St., Lexington MA.
6. M. A. Anusuya and S. K. Katti, " Speech Recognition by Machine : A Review", International journal of Computer Science and Information Security, vol. 6, no. 3, 2009 , pp.181-205.
7. D. Shweta, M. Rajni, "Speech Recognition Techniques: A Review", International journal of advanced research in computer science and software engineering, vol. 4, issue 8, August 2014.
8. K. Kuldeep, , R. K Aggarwal , "A Hindi speech recognition system for connected words using HTK", Int. J. Computational systems Engineering, vol. 1, No. 1. , Haryana, India, 2012.
9. D. Mayank, Aggarwal, R. K., "Implementing a speech recognition system interface for Indian languages", proc.of the JCNLP-08 workshop on NLP", , Hyderabad, India, January 2008, pp. 105-112.

10. H. Hermansky, "Perceptually predictive analysis of speech", Journal of Acoustic Society of America, vol. 87, 1990 , pp. 1738-1752.
11. S. Preeti, K. Parneet, "Automatic Speech Recognition: A Review", International journal of engineering trends and technology, vol. 41, Issue 2, Haryana India, 2013.
12. S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches", International journal of signal processing, image processing and pattern recognition", vol. 9, no. 4, Coimbatore, India, 2016, pp. 393-204.
13. K. Kumar and R. K. Aggarwal, "A Hindi speech recognition system for connected words using HTK", Int. J. Computational systems Engineering, vol. 1, no.1, 2012.
14. S. S. Bhabad, G. K. Kharate, "An overview of technical progress in speech recognition", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3 , issue 3, March 2013.
15. M. Agarwal and T. Raikhwar, "Speech recognition using signal processing techniques", International journal of engineering and innovative technology (IJEIT), vol. 5, Issue 8, February 2016.
16. R. Lawrence and B. H. Juang, "Fundamentals of speech recognition" Pearson Education, Inc. (AT & T), 1993.
17. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustic Speech and Signal Processing, vol. 26, no. 1, 1978, pp. 43-49,
18. J. K . Baker, "The Dragon System- An Overview", IEEE Transaction on Acoustics speech signal Processing, vol. ASSP-23, no.1, 1975, pp. 24-9,
19. P. Kannadaguli, P. Bhat, "A Comparison of Gaussian mixture modeling (GMM) and Hidden Markov modeling (HMM) based approaches for automatic phoneme recognition in Kannada", IEEE, 2015, pp. 257-260.
20. J. Blimes , "What HMM can do", IEICE Transaction Inf. , vol. E89-D, no.3, 2006, pp. 869-891.
21. N. Morgan et.al, " Pushing the envelope- Aside [Speech Recognition]", IEEE Signal Processing Mag., vol. 22, no. 5, 2005, pp. 81-88.
22. H. A. Bourlard and N. Morgan, "Connectionist Speech Recognition- A Hybrid Approach", kulwer academic publishers, 1994.
23. N. Smith and M. J. F. Gales, "Using SVM's and discriminative models for speech recognition", Proc. ICASSP, vol. 1, 2002 , pp. 77-80.
24. D. Yu and L. Deng., "Automatic Speech Recognition- A deep learning approach", Springer- Verlag London, 2015.
25. L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview", IEEE transactions on Audio, Speech and language processing, vol. 21, no. 5, , 2013, pp. 1060-1089.
26. G. Hinton et.al, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE signal process. Magazine, vol. 21, no. 6, , 2012, pp. 82-97.
27. S. Furui, " 50 years of progress in speech and Speaker Recognition Research", ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, November 2005.
28. K. H. Davis et.al, " Automatic Recognition of spoken Digits", Journal of Acoust. Soc. America, 24(6), , 1952, pp. 637-642.
29. H. F. Olson and H. Belar, " Phonetic Typewriter", Journal Acoust. Soc. Am., 28(6): 1072-1081, 1956.
30. D. B. Fry , " Theoretical Aspects of Mechanical Speech Recognition", P. Denes, " The design and operation of the Mechanical Speech Recognizers at University College London", J. British Inst. Radio Engr., 19:4, 1959, pp. 211-299.
31. J. Suzuki and K. Nakata, " Recognition of apaneseVowels-Preliminary to the Recognition of Speech", J. Radio Research Lab 37(8), 1961,pp.193-212,
32. V. M. Velichko and N. G. Zagoruyko, " Automatic Recognition of 200 words", Int J. Man –Machine Studies, 2:223, June 1970.
33. J. Ferguson, Ed., Hidden Markov models for Speech, IDA, Princeton, NJ, 1980.
34. L.R. Rabiner, " A Tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proc. IEEE, 77(2), February 1989, pp.257-286.
35. S. Katagiri, " Speech pattern recognition using neural networks", W. Chou and B. H. Juang (Eds.) Pattern recognition in Speech and language Processing, CRC Press, 2003, pp. 115-147.
36. R. K. Aggarwal and M. Dave, " Using Gaussian Mixtures for Hindi Speech Recognition System", International Journal of Signal Processing and Pattern recognition, vol.4, no. 4 , December 2011.
37. Z. Yu et.al, " The Teaching Experiment of Speech Recognition based on HMM", IEEE, , 2014, pp. 2416-2420.
38. K. A. Qazi, T. Nawaz et.al, " A hybrid technique for speech segregation and classification using a sophisticated deep neural network", PLoS ONE 13(3):e0194151, , March 20, 2018 , pp. 1-15.