Speech Recognition System Using Human Voice In Indian Language (Review)

[1]Shaikh Ajij Amirsab, [2]Tadvi Pathan Ekrar Khan Ebrahim khan

[1]Assistant Professor, [2]Lecturer
Master of Technology (VLSI System Design)
Department Electronic &Communication Engineering
Everest Education Society's college of Engineering
Aurangabad (Maharashtra)-431005.India.

***Abstract*-Speech recognition are becoming more and more useful nowadays.This paper presents an overview of speech recognition technology. This report presents an overview of speech recognition technology, software, development and applications.After years of research and development the accuracy of automatic speech recognition (ASR) remains one of the most important research challenges e.g. speaker and language variability, vocabulary size and domain, noise. With growth in the needs for embedded computing and the demand for emerging embedded platforms, it is required that the speech recognition systems (SRS) are available on them. Today after intense research, Speech Recognition System, have made a niche for themselves and can be seen in many walks of life.The advantage of discrete speech is that word boundaries can be set exactly while with continuous speech; words will be spoken without silences.This paper presents a study of basic approaches to speech recognition and their results shows better accuracy. Speech recognition systems are the efficient alternatives for such devices where typing becomes difficult. But they are usually meant for and executed on the traditional general-purpose computers. The accuracy of Speech Recognition Systems remains one of the most important research challenges e.g. noise, speaker variability, language variability, vocabulary size and domain. The design of speech recognition system require careful attentionsto the challenges or issue such as various types of speech classes, speech representation, feature extraction techniques, database and performance evaluation.In this paper we are using a HMM (hidden Markov model) to recognize speech samples to give excellent results for isolated words. This paper also presents what research has been done around for dealing with the problem of ASR.The design of speech recognition system requires careful attentions to the challenges such as various types of Speech Classes and Speech Representation, Speech Preprocessing stages. The advantage of discrete speech is that word boundaries can be set exactly while with continuous speech; words will be spoken without silences.**

***Keywords*—Speech Recognition,acoustic model,style; Feature extraction.**

Introduction

. Speech is the primary means of communication between people. Speech recognition, generation of speech waveforms, has been under development for several decades [10]. Automatic speech Recognition is a process by which a computer takes a speech signal and Converts it into words [1]. It is the process by Which a computer recognizes what a person Said. Speech recognition is also known as automatic speech recognition or computer speech recognition which means understanding voice of the computer and performing any required task or the ability to match a voice against a provided or acquired vocabulary. The Speech is the most common & primary mode of communication among human beings. It is the most natural and efficient form of exchanging information among humans. Since ages speech has been an important mean of communication between humans. Speech Recognition is the process of converting an acoustic speech into text, and / or identification of the speaker. .A mouse on the other hand requires a good hand eye co-ordination. Physically challenged people find computer difficult to use. Partially blind people find reading from a monitor difficult. All these constraints have to be eliminated. Speech interface help us to tackle these problems. The task is to getting a computer to understand spoken language. By "understand" we mean to react appropriately and convert the input speech into another medium e.g. text. Speech recognition is therefore sometimes referred to as speech-to-text (STT). Human voice conveys much more information such as gender, emotion and identity of the speaker. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means an Algorithm. Over the years with recent advent in technology it has become an essential and integral part of our lifestyle due to the increasing communication between human and computers or automated systems. The objective is to trap human voice in a digital computer and decode it into corresponding text. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. When two people speak to one another, they both recognize the words and the meaning behind them. Computers, on the other hand, are only capable of the first thing: they can recognize individual words and phrases, but they don't really understand speech in the same way as humans do.

A speech recognition system consists of a microphone, for the person to speak into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words.

2.Classification Of Speech.

I) Isolated word: The Isolated word have sample windows.itaccepts single word or single utterances at a time.Isolatedutterance might be a better name of this work [3].
II) Connected word: The Connected word system are similar to isolated words but allow separate utterance to be "run together minimum pause between them.
III) Continuousspeech: It allows user to speak almost naturally, while the computer will examine the content. There are special methods used to determine utterance boundaries and various difficulties occurred in it.
IV) Spontaneousspeech: A System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together.
Speaker Dependent: - systems that require a user to train the system according to his or her voice.
V) Speaker Independent: - systems that do not require a user to train the system i.e. they are developed to operate for any speaker.
VI) Isolated word recognizers: - accept one word at a time. These recognition systems allow us to speak naturally continuous.
VII) Connected word systems [1] allow speaker to speak slowly and distinctly each word with a short pause i.e. planned speech. VIII) Spontaneous recognition systems [1] allow us to speak spontaneously.

A. Type of speech there are basically two types of speech:
1. Continuous speech
2. Discrete speech.

Discrete speech consists of isolated words that are separated by silences [3]. The advantage of discrete speech is that word boundaries can be set exactly while with continuous speech; words will be spoken without silences.

B. Size of the vocabulary

The size of the vocabulary is the second typical aspect of a speech recognition technology. The vocabulary is a set of words that have to be recognized. A small vocabulary is one, which contains less than about 30 words. A 500-word vocabulary is average size. A vocabulary with more than 25000 words generally will be seen as very big, although these definitions tend to depend on the application field.

C.Speaker dependence

1. **Speaker dependent system**
2. **Speaker independent system**
3. **Speaker adaptable system**

Some speaker-dependent systems require only that the user record a subset of system vocabulary to make the entire vocabulary recognizable. A speaker-independent system does not require any recording prior to speaker-dependent system requires that the user record an example of the word, sentence, or phrase system use. A speaker independent system isdeveloped to operate for any speaker of a particular type (e.g., American English). A speaker adaptive system is developed to adapt its operation to the characteristics of new speakers.

3. Speech Recognition Techniques

The goal of speech recognition is to analyze, extract, characterize and recognize information about the speaker identity. Variety of the techniques are used for determining the speech characteristics.

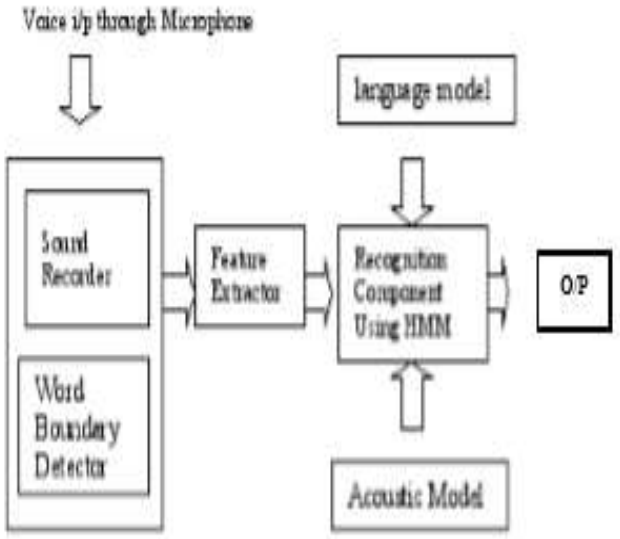
Speech analysis technique The speech data contain different type of information that shows the speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The speech analysis stage deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting [4].These are of three types.

i) Segmentation analysis In this work, speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information. This method is used to extract vocal tract information of speaker recognition.

ii) Sub segmental analysis Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmental analysis. This technique is used to mainly analyze and extract the characteristic of the excitation state. [5].

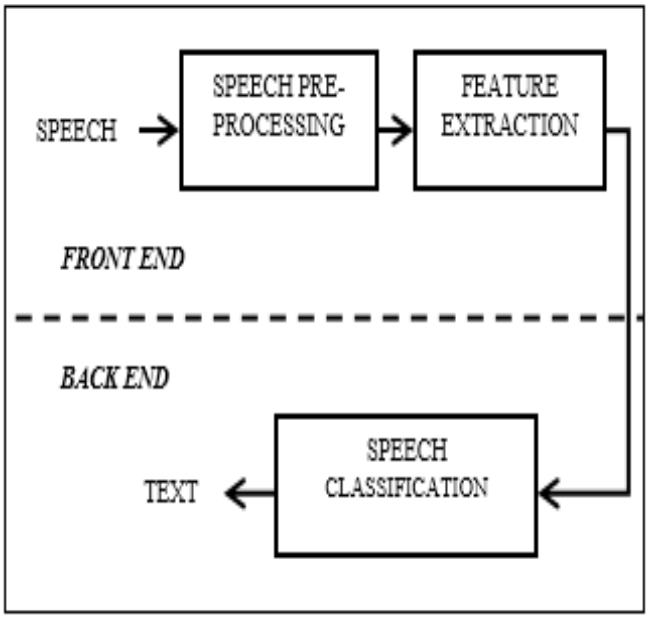
iii) Supra segmental analysisin this work, speech is also analyzed using the frame size.This technique Is mainly used to analyze and characteristic the behaviour character of the speaker.

4. Design Of The Speech Recognition System.



A. Sound Recording and Word detection component:-The component is responsible for taking input from microphone and identifying the presence of words. Word detection is done using energy and zero crossing rate of the signal. The output of this component can be a wave file or a direct feed for the feature extractor.
B. Feature Extraction component: - The component generated feature vectors for the sound signals given to it. It generates Mel Frequency Cestrum Coefficients and Normalized energy as the features that should be used to uniquely identify the given sound signal.
C. Recognition component: This is a Continuous, Multi-dimensional Hidden Markov Model based component. It is the most important component of the system and is responsible for finding the best match in the knowledge base, for the incoming feature vectors.
D. Knowledge Model: The components consist of Word based Acoustic. Acoustic Model has a representation of how a word sounds. Recognition system makes use of this model while recognizing the sound signal.

Once the training is done, the basic flow can be summarized as the sound input is taken from the sound recorder and is feed to the feature extraction module. The feature extraction module generates feature vectors out of it which are then forwarded to the recognition component. The recognition component with the help of the knowledge model and comes up with the result.



Front-End Analysis

Front-End of the speech recognition system comprises of Speech Preprocessing and Feature Extraction Block. Noise and differences in Amplitude of the signal can hardly influence the integrity of a word while timing variations can cause a large difference amongst samples of the same word. These issues are dealt with in the Signal Preprocessing part. Preprocessing generally involves End Point Detection, Reemphasis Filtering, Noise Filtering, Framing, Windowing, Echo Cancelling, etc.

Signal Preprocessing

Block Diagram for Signal Preprocessing stage is shown in Figure below. Feature Extraction is a process extracting specific features of the preprocessed speech signal.

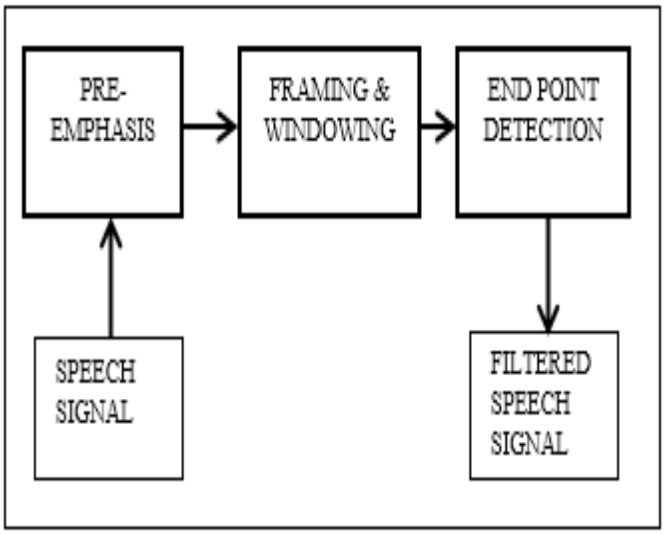


Figure 2. Signal Preprocessing

Back-End Analysis

Back-End consists of Speech Classification block. Speech Classification process is for classifying the extracted features and relates the input sound to the best fitting sound from a database and represents them as an output.

5.Modeling Technique In Speech Recognition System.

The aim of modeling technique is to use the specific feature of the speaker for creating speaker models. The speaker modeling technique is basically classified as speaker recognition and speaker identification. The speaker identification technique defines who is speaking on basis of individual information obtained from speech signal. The speaker recognition is further divided into two parts i.e. speaker dependent and speaker independent.

Speaker recognition can also be divided into two methods, text- dependent and text independent methods. In text dependent method the speaker speakes key words or sentences having the same text for both training and testing trials whereas text independent does not rely on a specific texts being spoken [8]. Following are the methods used in speech recognition process are as follows:

i).Pattern Recognition approach A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. A pattern recognition has been developed over two decades and received much attention and applied widely in many practical problem .It involves two essential steps namely pattern training and pattern comparison. The essential feature of this approach is that it uses a well-defined mathematical framework and then creates speech pattern representations. The pattern-matching approach has become the predominant method for speech recognition in the last six decades

ii). The acoustic-phonetic approach This method has been studied and used for more than 40 years. This approach is based upon theory of acoustic phonetics and postulates [10]. The work done before to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach which postulates that there exist finite, distinctive phonetic units in spoken language and these units are broadly characterized by a set of acoustics properties that are changed in the speech signal over time. There are three methods that have been applied to the language identification i.e. Problem phone recognition, Gaussian mixture modeling, and support vector machine classification.

iii).Learning based approachesto overcome the disadvantage of the HMMs machine learning methods which was introduced in neural networks and genetic algorithm programming learning based approaches has been taken. In learning based approaches ,they can be learned automatically through emulations or evolutionary process.

iv) Knowledge based approaches The guidance should be taken from an expert knowledge about variations in speech is hand coded into a system.This approach gives the advantage of explicit modeling but this situation is difficult to obtain and cannot used successfully. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram.The test speech is considered by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure [11]. Vector Quantization (VQ) [12] is often

applied to ASR. It is useful for speech coders, i.e., efficient data reduction.

v)Artificial intelligence approach The artificial intelligence approach coordinate the recognition procedure according to the person who applies the intelligence of a person such as visualizing, analyzing etc. are used for making a decision on the measured acoustic features. The Artificial Intelligence approach [13] is a hybrid of the acoustic phonetic approach and pattern recognition approach. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert's speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Knowledge enables the algorithms to work better. This form of knowledge based system increases the contribution and hence successful designs and strategies has been reported.

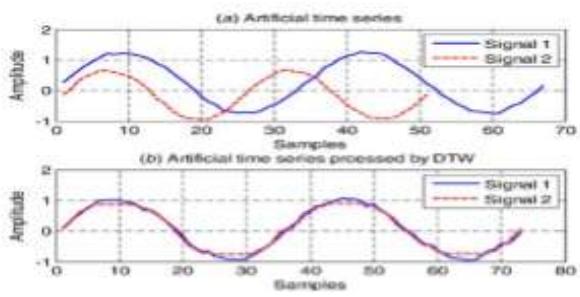


Fig.2.Dynamic Time Wrapping of two speech signal.

DTW: Dynamic time warping is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. DTW is a method that calculates an optimal match between two given sequences. DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data which can be turned into a linear sequence can be analyzed with DTW. Applications include speaker recognition and online signature recognition.

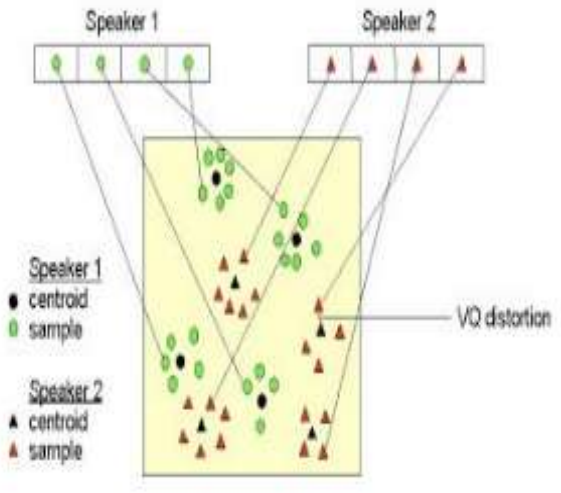


Fig.3 Vector Quantization of two speech signals.

VQ: Vector quantization (VQ) is a classical quantization technique from signal processing. It was originally used for data compression[14]. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms. The density matching property of vector quantization is very powerful for large and high-dimensioned data. Hence VQ is suitable for lossy data compression. It can also be used for lossy data correction and density estimation.

LBG: Linde-Buzo-Gray (LBG) Algorithm:This is an algorithm developed in the community of vector quantization for the purpose of data compression[15].One speaker can be discriminated from another based on the location of centroids codebook for this speaker using those training vectors for clustering a set of L training vectors into a set of M codebook vectors.

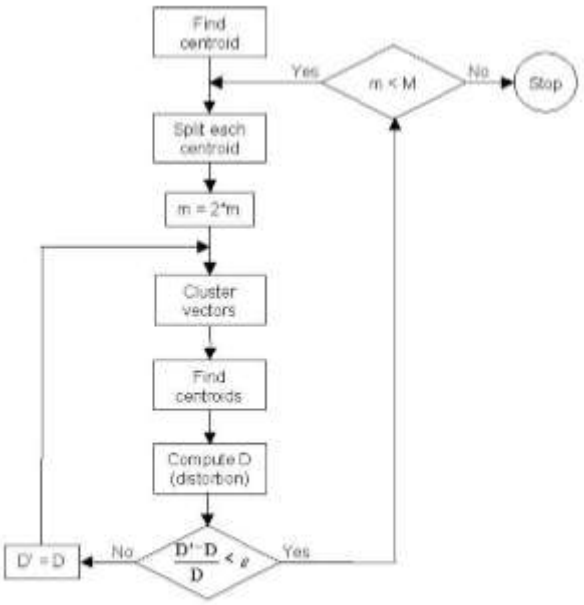


Fig.4 Linde-Buzo-Gray (LBG) Algorithm

CONCLUSION

Speech recognition is a challenging problem to deal with. We have attempted in this paper to provide a review of how much this technology has progressed in the previous years. Speech recognition is one of the most integrating areas of machine intelligence, since humans do a daily activity of speech recognition. Through this paper, we present a scheme to convert speech to text. The key factor in designing such system is the target audience. For example, physically handicapped people should be able to wear a headset and have their hands and eyes free in order to operate the system. The Preprocessing quality is giving the biggest impact on the Speech Classification performance. Signal Preprocessing consist an EPD, Filtering, Framing, Windowing, Echo Cancellation, etc. An Improvement in any individual part can improve the overall system performance. For effective working of Back-End there should be more efforts in FrontEnd processing. It has attracted scientists as an important discipline and has created a technological impact on society as well as, is expected to flourish further in area of human machine interaction.

REFERENCES

[1] Anne Johnstone Department of Artificial Intelligence Edinburgh University Hope Park Square, Meadow Lane Edinburgh EHB 9LL, (GB) Gerry Altmann "AUTOMATED SPEECH RECOGNITION: A FRAMEWORK FOR RESEARCH".

[2] Reddy, D.R. & Ermann, L.D. 1975. "Tutorial on System Organisation for Speech Understanding." In D.R. Reddy (ed) Speech Recognition, Academic Press.

[3] Rumelhart, D.E. & McClelland, J.L. 1982. "An Interactive Activation Model of Context Effects in Letter Perception: Part II. The Contextual Enhancement Effect. Some Testsand **Extensions of the Model. In Psychological** Review"

[4] Tetsuya Matsumoto, Kazuhito Hagio, and Masayuki Takeda Department of Informatics, Kyushu University, Fukuoka 819-0395, Japan {tetsuya.matsumoto, kazuhito.hagio, takeda "More Speed and More Compression: Accelerating Pattern Matching by Text Compression"

[5] C. H. Lee; F. K. Soong; K. Paliwal "An Overview of Speaker Recognition Technology", Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers 1996, Norwell, MA.

[6] "Pattern matching for large vocabulary speech recognition systems"availableat
www.freepatentsonline.com/6879954.html

[7] "Isolated-word automatic speech recognition (iwasr) design"availableat
www.dspace.fsktm.um.edu.my/xmlui/bitstream/
handle/1812/111/Chapter%205.pdf?sequence=7

[8] R. Rodman, "Computer Speech Technology". Artech House, Inc. 1999, Norwood, MA 02062

[9] M. J. Castro; J. C. Perez, "Comparison of Geometric, Connectionist and Structural Techniques on a Difficult Isolated Word Recognition Task.", Proceedings of European Conference on Speech Comm. and Tech., ESCA, Vol. 3, pp 1599-1602, Berlin, Germany, 1993

[10] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.

[11] Mohit Dua, R.K.Aggarwal, Virender Kadyan and Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK", IJCSI International Journal of Computer Science Issues, vol. 9, issue 4, no. 1, July 2012.

[12] Rajesh Kumar Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309– 320, 2011.

[13] Kuldeep Kumar, Ankita Jain and R.K. Aggarwal, "A Hindi speech recognition system for connected words using HTK", Int. J. Computational Systems Engineering, vol. 1, no. 1, pp. 25-32, 2012.

[14] Kuldeep Kumar R. K. Aggarwal, "Hindi speech recognition system using HTK", International Journal of Computing and Business Research, vol. 2, issue 2, May 2011.

[15] R.K. Aggarwal and M. Dave, "Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system", 01 September 2011.

[16] Anusuya, M. A., & Katti, S. K.. Front end analysis of speech recognition: A review. International Journal of Speech Technology,Springer, vol.14, pp. 99–145, 2011.

[17] N. N. Lokhande, N. S. Nehe, P. S. Vikhe, "Voice Activity Detection Algorithm for Speech Recognition Applications", ICCIA, 2011.

[18] Hui Jiang, K. Hirose and Qiang Huo, "A Minimax Search Algorithm for Robust Continuous Speech Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 8, No. 6, November 2000.

[19] J. K. Lee and C. D. Yoo, "Wavelet Speech Enhancement Based On Voiced/Unvoiced Decision", the 32nd International Congress and Exposition on Noise Control Engineering Jeju International Convention Center, Seogwipo, Korea, August 25-28, 2003.

[20] W. Gevaert, G. Tsenov and V. Mladenov, "Neural Networks used for Speech Recognition", Journal of Automatic Control, University Of Belgrade, Vol. 20:17, 2010.

[21] Amr Rashed, "Fast Algorithm for Noisy Speaker Recognition Using ANN", IJCET, Volume 5, Issue 2, February (2014), pp. 56-65.

[22] T. Lee, C. Ching and Lai-Wan Chan, "Isolated Word Recognition Using Modular Recurrent Neural Networks", Pattern Recognition, Vol. 31, No. 6, pp. 751—760, 1998.

[23] K. Dutta and K. K. Sarma, "Multiple Feature Extraction for RNN-based Assamese Speech Recognition for Speech to Text Conversion Application", International Conference on Communications, Devices and Intelligent Systems (CODIS), IEEE, 2012.

[24] K. Dutta and K. K. Sarma, "Dynamic Segmentation of Vocal Extract for Assamese Speech to Text Conversion using RNN", CISP, IEEE, 2012.

[25] A. Singh, Dr. D. K. Rajoria, V. Singh, "Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks", International Journal of Computer Applications, Volume 52, No.12, August 2012.

[26] M. Vyas, "A Gaussian Mixture Model Based Speech Recognition System Using Matlab", SIPIJ, Vol.4, No.4, August 2013.

[27] Hiroaki Sakoe, "Two-Level DP-Matching, A Dynamic Programming Based Pattern Matching Algorithm For Connected Word Recognition", IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp27, No. 6, December 1979.
IJCATM

[28] Dr. Shaila D.Apte, "Speech and Audio Processing",Wiley India Edition.

[29] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang, "Springer Handbook of Speech Processing", Springer.

[30] L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice Hall Signal Processing Series.

[31] N. Srivastava, "Speech Recognition using Artificial Neural Network", IJESIT, Volume 3, Issue 3, May 2014.

[32] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-24,No. 5, October 1976.

[33] S. Ahmadi and A. S. Spanias, "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm", IEEE Transactions on Speech And Audio Processing, Vol. 7, No. 3, May 1999.

[34] K.K. Paliwal, "Effect of Preemphasis on Vowel Recognition Performance", Elsevier Science Publishers B.V. (North-Holland), Vol. 3. No. 1. April 1984.

[35] R. Vergin, Douglas O'Shaughnessy and A. Farhat, "Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent ContinuousSpeech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 5, September 1999

[36] I. Patel, Dr. Y. Srinivas Rao, "Speech Recognition Using HMM with MFCC-AN Analysis Using Frequency Spectral Decomposition Technique", SIPIJ,Vol. 1,No. 2,December 2010.

[37] A. N. Mishra, M. Chandra, A. Biswas, S. N. Sharana, "Robust Features for Connected Hindi Digits Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 2, June, 2011.

[38] Sadaoki Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSp-34, No. 1, February 1986.