Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 2, February 2015, pg.359 – 368

RESEARCH ARTICLE

The State of the Art of Automatic Speech Recognition: An Overview

Ira Badyal[1], Ms. Divya Gupta[2]

[1]CSE Dept., Amity University, Noida, India

[2]CSE Dept., Amity University, Noida, India

[1] irabadyal.badyal@gmail.com; [2] dgupta1@amity.edu

Abstract— The objective of this paper is to present an overview of the techniques used in speech recognition systems. In this paper, we discuss the types of speech recognition showcasing the development in the field to help provide a technological perspective of the progress made in the field. Further, it highlights the fundamental principles and methods of Speech Recognition to understand the basic design required to build the technology. In addition, this paper discusses the various approaches to ASR and the classification techniques of the Speech Recognition System- HMM, DTW, MLP, along with their advantages and disadvantages. After decades of research, the efficiency of an ASR system and its accuracy remains the most crucial challenge. This paper attempts to review the basic technology of Speech Recognition, based on which we can build the most advanced systems overcoming the challenges we face currently.

Keywords- Automatic Speech Recognition (ASR), Feature Extraction, Pattern Matching, Reference Pattern, Hidden Markov Model(HMM), Dynamic Time Warping(DTW), Multilayer Perceptron (MLP).

I. INTRODUCTION

Speech has been a fascinating object of research for a very long time now. Using speech to give commands to the devices and operating them with ease has been the aim of research since the beginning. Speech is converted into texts for easy communication between a man and a machine which is facilitated by the means of natural language communication. This research is forming the basis of man-machine interface to be used in most industries and is the base for revolutionizing the way in which machines take commands.

Speech Recognition enables the machines to intercept the speech of the user and act accordingly. The interception is mostly handled by speech signal processing and pattern recognition. With improved hardware and software, computer information processing technology has become one of the most widely significant technological developments. The speech recognition is cross-disciplinary involving acoustics, phonetics, linguistics, information theory, pattern recognition, neurology, signal processing, psychology, physiology and even the study of human body language. [1]

In section II, we discuss the types of speech recognition which ranges from isolated words to spontaneous speech. The main aim is to develop machines that can comprehend the natural language spoken by humans. Section III discusses the basic principles

and methods of the Speech Recognition Technology based on which we can build the fundamental ASR system. It is based on the Pattern Recognition method, involving Feature Extraction, Pattern Matching and a Model Reference Library. Section IV discusses the different approaches to develop an Automatic Speech Recognition System, that are, Acoustic-Phonetic Approach, Pattern Recognition Approach, Dynamic Time Warping Approach and the Artificial Intelligence Approach. Section V presents the various classification techniques of the Speech Recognition System namely, HMM, DTW and MLP, along with their advantages and disadvantages to portray the various characteristics of each classifier to be used in the research and development. Finally, the paper discusses the facing challenges of the field in section VI.

II. TYPES OF SPEECH RECOGNITION

1. *Isolated Words*: In this type of recognition, a single utterance is detected by the recognizer. It is required that the utterance must have a quiet background, which means lacking an audio signal. However, it does not mean that a single word is a single utterance. The utterances need to be isolated in nature, i.e. they should have "Listen" and "Not Listen" states with a time interval amongst utterances.[3]
2. *Connected Words*: As an advancement of recognizing single utterances, connected word recognizers allow the utterances to be spoken together to form connected words. Only a minimum pause is suggested between utterances. [3]
3. *Continuous Speech*: The next advancement is the recognition of continuous speech. The main problem here is deciding the utterance boundaries in a speech input. These recognizers help in recognizing the content from a fluent speaker, without "Listen" and "Not Listen" states. [3]
4. *Spontaneous Speech*: The most natural form of speech is called Spontaneous Speech. The main agenda of building an ASR System is to comprehend the natural language spoken to a machine by the user. These might include recognizing features like stutters or the words that run together. [14]

III. BASIC PRINCIPLES AND METHODS OF SPEECH RECOGNITION TECHNOLOGY

The basic speech recognition model is based on a pattern recognition system and includes the following features: [1]

1. Feature Extraction
2. Pattern Matching
3. Model Library for reference

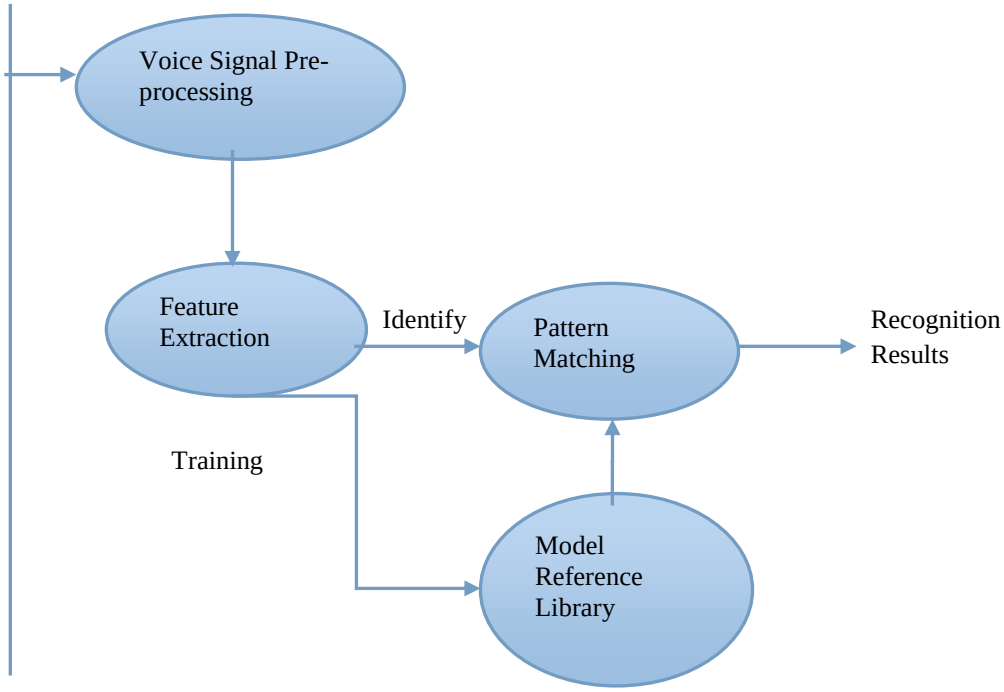


Figure 1: Basic Principles of Speech Recognition System [1]

In the English language, speech can be broken down into various fundamental structures forming a word. These fundamental units of the spoken language are called *phonemes.* The English languages comprises of 44 phonemes that can together represent all the vowels and consonants in the language.

Voice Signal Preprocessing: This is the first phase in the speech recognition technology. In this phase, an unknown voice is input to the system with the help of a microphone. This input voice signal is then transformed to an electrical signal. Then, a voice model is established by the system according to the human characteristics. The process then moves on to the next phase i.e. Feature Extraction. [1]

Feature Extraction: In simple words, all the important characteristics of the input voice that define it are extracted and a voice template is generated. Feature extraction is mainly used to compute a feature vector sequence to represent the input signal. There are three stages of feature extraction : [3]

1. *Speech Analysis or Acoustic Front End:* In this stage, spectrotemporal analysis of the input signal are performed and raw features describing the envelop of the power spectrum of short speech interval are generated.
2. *Compilation:* In this stage, a feature vector is compiled that comprises of the static and dynamic features.
3. *Transformation:* This stage, although not always present, transforms the feature vector from the second stage into more compact vectors which are then supplied to the recognizer.

The most agreed upon features in these stages are: [3]

1. An automatic system should help in distinguishing between different, though similar sounding speech.
2. Without the help of training data, an automatic creation of acoustic models should be made for these sounds.
3. Invariant statistics across speakers and speaking environments should be shown.

Pattern Matching : Computers are then used in the recognition process for matching. The stored voice template is matched with the characteristics of the input voice signal. Search and Matching Techniques help in identifying the optimal range of the input voice matching the template. [1]

Model Reference Library: The reference library consists of all the speech signal templates which are compared with the input signal during Pattern Matching, to identify and recognize the speech signal input. [1]

IV. **APPROACHES TO ASR**

There are three main approaches to speech recognition.

1. *Acoustic-Phonetic approach*: [3]

The basic principle of this approach is that the human language consists of a finite and distinct phonemes and these phonemes, though very variable, are governed by a broad number of rules which can be readily learnt by a machine. In the simplest words, sounds are recognized, labeled and referenced in the future for recognition. This factor has been exploited for the acoustic-phonetic approach which is broken down into the following steps:

- Spectral analysis of speech
- Feature Extraction, where the spectral measurements describe a set of features that describe acoustic properties of the processed speech signal.
- Segmentation and Labeling, where the input speech signal is segmented into acoustic regions and phonetic labels are given to each segment for accurate identification.
- The most important step is the last step that determines the valid word for the phonetic label. After that, certain linguistic constraints on the task are used to decode the word.

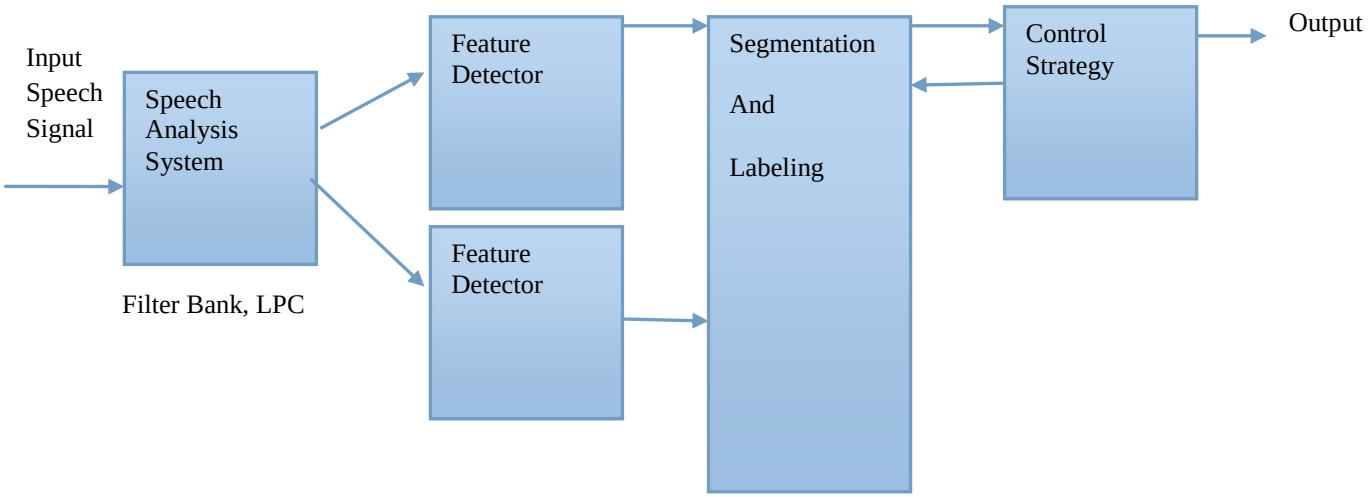


Fig 2: Acoustic phonetic speech recognition system[8]

2. *Pattern Recognition approach*: [3]

Pattern Recognition has been dominant in this domain for the last six decades. It involves two steps: Pattern Training and Pattern Comparison. During pattern training, labeled training samples via a formal training algorithm are formed for matching in the future. In pattern comparison, the unknown speech input is compared with the speech samples formed during the pattern training and accurate speech pattern representations are formed. Pattern recognition can be carried out as two types:

- Template Based Approach: A history of the candidate's dictionary forms the templates for comparison, stored as prototypes of the candidate's speech and is referenced for matching the input signal and generate appropriate results. But after sometime, the reference library size increases exponentially and it is advised to store speech frames as patterns and to compare the spectral variations of the patterns for finding a match.
- Stochastic Approach: This is used in case of uncertain information. Probabilistic models like Hidden Markov Model use the key features of the Speech Recognition technique-transition parameters and temporal variabilities, and parameters in the output distribution and spectral variabilities.

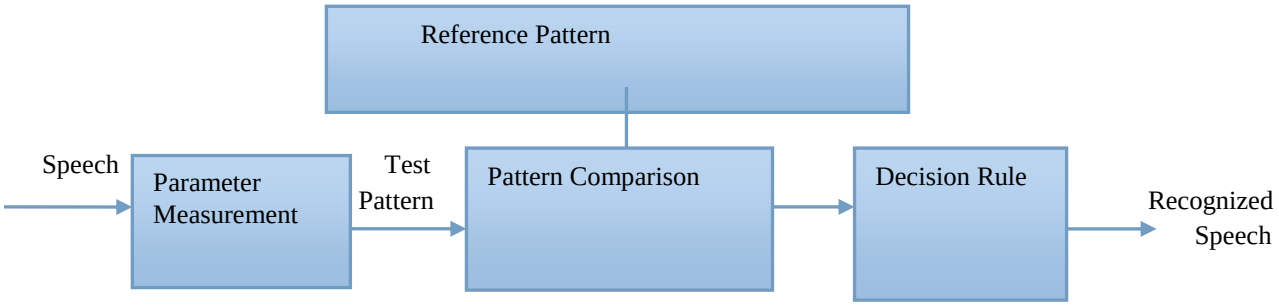


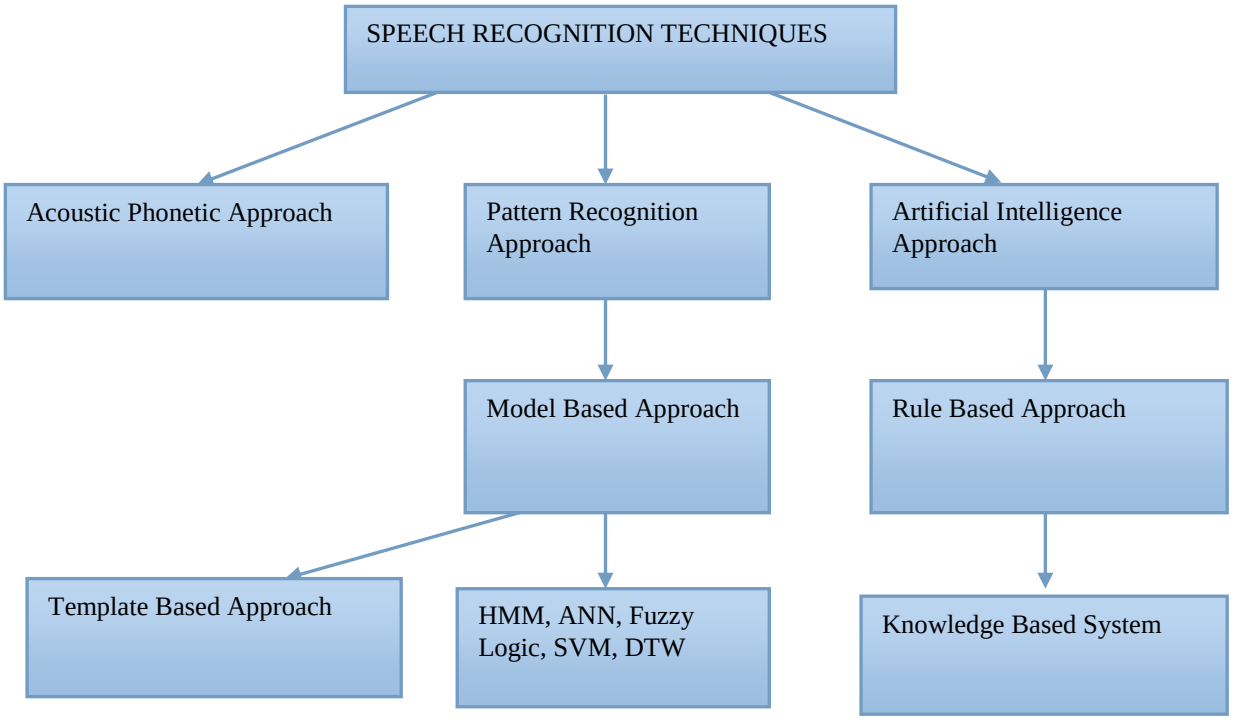
Fig 3: Pattern Recognition Approach to speech Recognition[8]

3. *Dynamic Time Warping(DTW):* This algorithm is used to measure the similarities in two speech signals that vary in time and speed. It has been applied in video and audio studying, typically where any form of data can be represented linearly. In

speech recognition, DTW helps in speeches with different speeds. In simple words, DTW helps in finding similarities in two given data keeping in mind the various constraints involved. [3]

4. *Artificial Intelligence (AI) Approach*: Artificial Intelligence has always been a fascinating branch of science that helps in understanding and mimicking the human behavior in certain situations. The main idea in using an Artificially Intelligent Approach here is to recognize the features of a speech based on the way a person applies his intelligence in analyzing, visualizing and characterizing. An expert system is applied that combines the working of the phonemic, lexical, syntactic, semantic and pragmatic knowledge. These measures features help in segmenting and labeling the speech. Artificial neural networks are also used to learn how the phonetic events are connected to each other as well as psychologically to the speaker. [3]

Taxonomy of Speech Recognition [3]



V. CLASSIFICATION TECHNIQUES OF SPEECH RECOGNITION SYSTEM

Various classification techniques of speech recognition system have been discussed below:

HMM: [6]

HMM, also known as a Hidden Markov Model, can be described as a collection of states that are connected by transitions. There are two sets of probabilities carried by each transition. First is, Transition probability. This shows the probability of taking this transition. Second is, Output Probability Density Function. If a transition is taken, this probability then provides the conditional probability of emitting each output symbol from finite alphabet. It helps in discovering an alignment path between different speech sounds and different model states, and the techniques for estimating the parameters of models from a training set of utterances of sounds being monitored.

Table I: Advantages and Disadvantages of HMM

Name of the classifier	Characteristics [6]	Advantages [8]	Disadvantages [8]
Hidden Markov Model	➢ The parameters of HMM represent the time-varying characteristics of the voice signal. ➢ The statistical characteristics of the signal are described by two inter related processes. ➢ These characteristics can be hidden or unobserved finite-state Markov chain, and an observation vector that is associated with each state of the Markov chain. [6]	➢ Gives a better compression as compared to a simple Markov Model. ➢ HMM can be extended to deal with strong tasks. ➢ HMM uses a technique called "embedded re-estimation" in which HMMs are dynamically assembled according to class sequence. To know the class sequence, the most probable paths are calculated. The traversed path corresponds to a sequence of class which gives us our final classification. Embedded re-estimation can help in including high-level domain knowledge,	➢ To use HMM to its full potential, the model has to be trained on a set of seed sequences and that requires a larger seed as compared to a simple Markov model. ➢ Different HMMs are possible for a given set of seed sequences and choosing one can be difficult. ➢ Larger models can fit the data in a better way, but smaller models can be understood better. ➢ Assumptions about the data are made. ➢ A large number of parameters need to be set for an HMM. [8]

		which is a crucial part of speech recognition. ➢ HMM is scalable. ➢ New data can be added, without affecting the previous HMMs. ➢ HMMs can also be incremental [8]	

DTW: [6]

Dynamic time warping, also referred to as DTW, is an algorithm that is used for measuring the similarities between two sequences which vary in factors like time or speed. Automatic Speech Recognition is a renowned implementation to help cope with different speaking speeds. In simple terms, DTW is a method that aids a computer to find an optimal match between two given sequences, considering certain restrictions. One example of these sequences can be a time series. Non-linear "Warping" is applied on these sequences in the time dimension to calculate a measure of their similarity which is independent of non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models.

Table II: Advantages and Disadvantages of DTW

Name of Classifier	Characteristics [6]	Advantages [8]	Disadvantages [8]
Dynamic Time Warping	➢ It is a method that allows a computer to find an optimal match between two given sequences with certain restrictions to determine a measure of their similarity independent of certain non-linear variations in the time	➢ Efficiency of DTW is very high for a small number of templates i.e. less than 20. ➢ DTW is language independent. ➢ It can be controlled by the end user and can be trained easily. ➢ DTW can be called a cost minimization technique. It has a reference	➢ DTW is speaker specific. ➢ It requires actual training samples. ➢ In some cases, a point on a time series can map to a large subsection of another time series in certain alignments. ➢ DTW may prevent the correct warping from being discovered. ➢ Sometimes, features like

	dimension. [6]	template according to which a test signal is stretched or compressed. ➢ DTW has a very simple hardware implementation, hence it is used in mobile devices. ➢ As compared to HMM, the training procedure is very fast and simple. ➢ We can impose constraints that prevent the use of sequences that are not optimal, thus reducing the computational complexity. [8]	peak, valley, inflection, plateau, point are higher or lower than their corresponding feature in another sequence by a small margin, and this prevents in finding the natural alignments in two sequences which otherwise would have been obvious. [8]

ML P: [13]

MLP, also known as Multilayer Perceptron is a form of an Artificial Neural Network model. The nature of MLP is feed forward. This model aids in mapping input data onto appropriate outputs. A MLP comprises of a directed graph with multiple layers with nodes. In this graph, each layer is connected to the next one. Except the input nodes, every node represents a neuron and has a non-linear activation function. A technique called "Back Propagation" is used by MLP, which is a supervised learning technique and is used to train the network. MLP can distinguish the data that is not separable linearly and is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

Table III: Advantages and Disadvantages of MLP

Name of the classifier	Characteristics[13]	Advantages[13]	Disadvantages[13]
MLP	➢ All the neurons in a Multilayer perceptron (MLP) have a linear activation function. ➢ It can be	➢ Adaptive learning: MLP can learn how to do the tasks given, based on the experience or training data.	➢ The dependency of MLP is entirely on the algorithms used to create it. ➢ It does not

	easily proved that the standard two-layer input-output model can be obtained by reducing any number of layers since it maps the weighted inputs to the output of each neuron. [13]	➢ MLP is well suited for gesture recognition. ➢ In comparison to other probability based models, MLP or Neural networks do not make assumptions regarding the probabilistic information about the pattern classes under study. ➢ The required decision function is produced directly with the help of training. [13]	scale well. ➢ Once MLP has been trained, it cannot be updated without retraining. ➢ MLPs previous knowledge cannot be preserved after retraining. [13]

VI. THE FACING PROBLEMS

The research in the field of Speech Recognition has been slow and there has not been a major breakthrough since the inception of the idea. The main drawback of the concept is its dependency on the environment. The environment in which the samples are created are the most ideal for recognizing future speech inputs. Also, the systems do not recognize error in the input. The main need for the speech recognition system is to gain efficiency in the noisy environments, although it has not yet been achieved. Another problem is the variation in the features of speech with each speaker, like voice, speech rate, pronunciation, pitch, volume. A new approach for signal analysis and processing must be formulated for improved efficiency. The detailed understanding of the brain control mechanism is still not very clear to us, and until then, speech recognition is going to be a little difficult. [1]

VII. CONCLUSION

Amongst humans, speech has always been the most convenient form of communication. Speech Recognition technology has helped automate the machines' work with ease. Researchers have collectively been working towards a common goal, i.e. to enable natural conversation between man and the machine. The entire research has formed stepping milestones in the journey with the aim of reducing the gap between the recognition capability of machines and that of humans to a maximum degree. Although there are still many prevalent challenges in this field, the use of Speech Recognition Technology in our day to day lives is going to be extensive and is likely to spread to other devices that we use in our day to day lives. It will not be surprising if we start giving commands to our coffee makers or our printers, with the help of this revolutionary technology. Research in

speech recognition has grown by leaps and bounds over the fast five decades. It has created a strong impact on society and has helped to develop the area of human-machine interaction. We hope that the comprehensive review of various approaches in this paper has helped demonstrate a better perspective towards the technological advancements in the domain.

REFERENCES

[1] Jianliang Meng, Junwei Zhang,Haoquan Zhao, *Overview of the Speech Recognition Technology*, School of Control and Computer Engineering, North China Electric Power University, Baoding, China.

[2] Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatan, and Walter Kellermann, NTT Communication Science Laboratories, Kyoto, Japan University of Erlangen-Nuremberg, Germany.

[3] Sanjivani S. Bhabad Gajanan K. Kharate, *An Overview of Technical Progress in Speech Recognition,* Department of E & TC, Pune university, India.

[4] Wiqas Ghai- Khalsa College (ASR) of Technology & Business Studies, Mohali, Punjab and Navdeep Singh -Mata Gujri College, Fatehgarh Sahib, Punjab, *Literature Review on Automatic Speech Recognition.*

[5] Titus Felix FURTUNĂ, Academy of Economic Studies, Bucharest, *Dynamic Programming Algorithms in Speech Recognition.*

[6] Lawrence Rabnier, Biing-Hwang Juang, B. Yegnanarayana, *Fundamentals of Speech Recognition.*

[7] Leung, H.C., Chigier, B., *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference.*

[8] M.A.Anusuya, S.K.Katti, *Classification Techniques used in Speech Recognition Applications: A Review.*

[9] Sonia Suuny, David Peter, K. Poulose, *Performance of Different Classifiers in Speech Recognition.*

[10] Mark Gales and Steve Young, *The Application of Hidden Markov Models in Speech Recognition.*

[11] Chotirat Ann Ratanamahatana Eamonn Keogh, *Everything you know about Dynamic Time Warping is Wrong.*

[12] Preeti Saini, Parneet Kaur, *Automatic Speech Recognition: A Review.*

[13] Bruce R. Maxim, *Learning and Perceptrons.*

[14] Parwinder Pal Singh- Computer science &Engg. IGCE, PTU Kapurthala, Er. Bhupinder singh ,Computer science &Engg. IGCE, PTU Kapurthala, *Speech Recognition as Emerging Revolutionary Technology.*

[15] Giuseppe Riccardi, Senior Member, IEEE, and Dilek Hakkani-Tür, Member, IEEE, *Active Learning: Theory and Applications to Automatic Speech Recognition.*

