Automatic Speech Recognition: A Review

Shipra J. Arora
Research Scholar, CSE Department
GJUST, Hisar

Rishi Pal Singh
CSE Department
GJUST, Hisar

ABSTRACT
This paper attempts to describe a literature review of Automatic Speech Recognition. It discusses past years advances made so as to provide progress that has been accomplished in this area of research. One of the important challenges for researchers is ASR accuracy. The Speech recognition System focuses on difficulties with ASR, basic building blocks of speech processing, feature extraction, speech recognition and performance evaluation. The main objective of the review paper is to bring to light the progress made for ASRs of different languages and the technological viewpoint of ASR in different countries and to compare and contrast the techniques used in various stages of Speech recognition and identify research topic in this challenging field. We are not presenting exhaustive descriptions of systems or mathematical formulations but rather, we are presenting distinctive and novel features of selected systems and their relative merits and demerits.

Keywords
Automatic speech recognition, Language Model, Speech Processing, Database, Pattern Recognition, Hidden Markov Model.

1. INTRODUCTION
We ask that authors The Speech is one of the most important tools for communication between human and his environment. Therefore manufacturing of ASR is need for human being all the time. Speech recognition made it feasible for machine to understand human languages. As information technology has a bang on more and more aspects of our lives with every year, the problem of communication between human beings and information processing devices becomes increasingly significant. Up to now, communication has almost fully been through the use of keyboards and screens, but speech is the most widely used, natural and the fastest means of communication for people. In a speech recognition system, many parameters affect the accuracy of the Recognition System. These parameters are: dependence or independence from speaker, discrete or continuous word recognition, vocabulary, environment, acoustic model, language model, perplexity, transducer etc. Problems such as noisy environment, different pronouncing of one word by one person in several times, dissimilar expressing of one word by two different speakers, incompatibility between train and test conditions led to made system without complete recognition. Resolving each of these problems is a good step towards this aim.

2. SPEECH AND LANGUAGE PROCESSING

2.1 Basic Building Blocks
Figure 1 shows some building blocks that perform transformations pertinent to speech recognition. A waveform modifier Figure 1(a) takes an input speech signal and produces a modified signal. The modification might be a clipping of large values of the signal; a frequency spectrum filtering that alters the shape of signal or enhances the speech and de-emphasises noise that is present. A symbol transducer Figure 1(b) can take in one discrete symbol sequence and yields a modified sequence on its output. If the input were a sequence of words in one language and the output were an equivalent word sequence in another language, this transducer would be a language translation device. Parameter extractor Figure 1(c) takes an input speech signal and yielding parameters of speech wave. In recognizers, it is often called the pre-processor. A feature extractor Figure 1(d) can receive parameters and produce a more abstract set of important information carrying features such as determining what portions of speech are voiced, whether the sound is loud and resonant like a vowel etc. A segmenter and labeller Figure 1(e) can receive the set of features and produce a linear string of phonemes or other identified segments. The unit identifier Figure 1(f) takes input symbol sequence which may be compared to the expected reference sequences for various units to determine what linguistic units appear to be in the input. The most common unit identifier is a 'word matcher' which finds the closest matching word, based on which word's stored pronunciation string is most like the input string. With these building blocks, we have the essential prerequisites for discussing the main knowledge sources needed for machine understanding of speech.

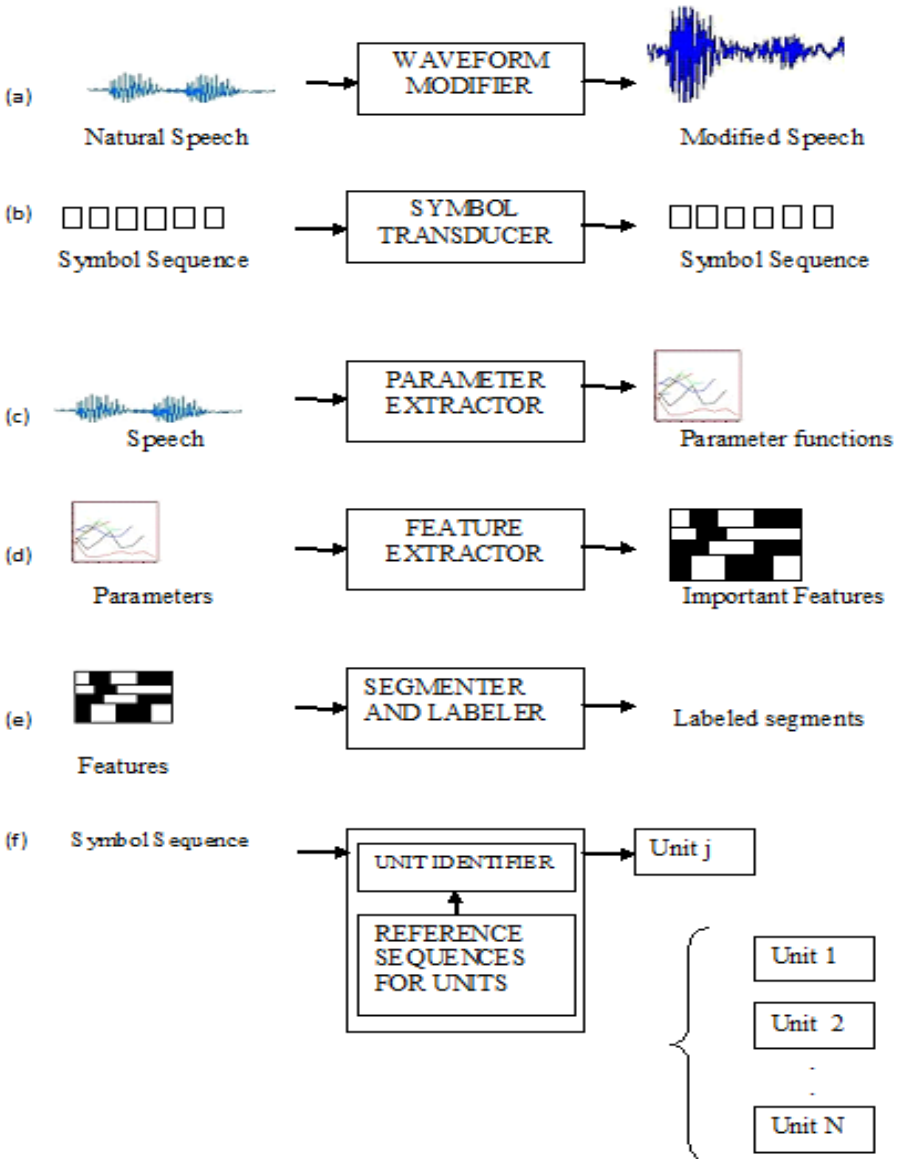


Figure 1 Basic Building Blocks for processing speech and language

2.2 Types of Speech

Speech can be classified into following categories

2.2.1 Isolated words:

Isolated word recognisers accept single word at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances. Isolated Utterance might be a better name for this class.

2.2.2 Connected words

Connected word speech recognition is the system where the words are separated by pauses. Connected word speech recognition is a class of fluent speech strings where the set of strings is derived from small-to-moderate size vocabulary such as digit strings, spelled letter sequences, combination of alphanumeric. Like isolated word speech recognition, this set too has a property that the basic speech-recognition unit is the word/phrase to much extent.

2.2.3 Continuous speech:

Continuous speech recognition deals with the speech where words are connected together instead of being separated by pauses. As a result unknown boundary information about words, co-articulation, production of surrounding phonemes and rate of speech effect the performance of continuous speech recognition systems. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

2.3 Classification of ASR system

Classification of speech recognition system is shown in Figure 2.

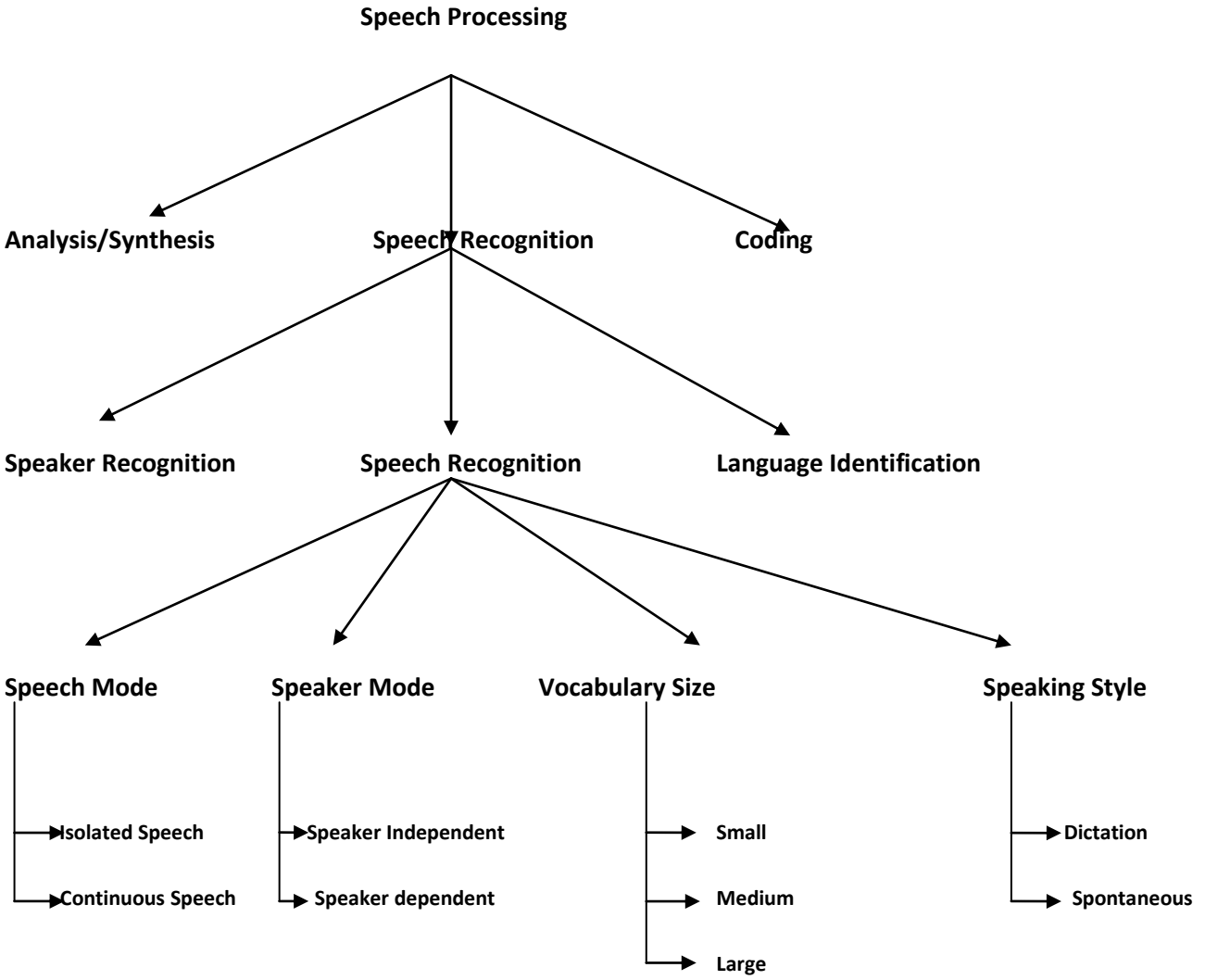


Figure 2 Speech Processing Classification

Speech recognition systems are classified as discrete or continuous systems that are speaker dependent or independent. Discrete systems maintain a separate acoustic model for each word, combination of words or phrases referred to as isolated word speech recognition (ISR). Continuous speech recognition (CSR) systems responds to a user who pronounces words, phrases or sentences that are in a series of specific order. A speaker-dependent system requires that the user record an example of the word, sentence or phrase prior to its being recognized by the system. i.e the user trains the system. A speaker-independent system does not require any recording prior to system use. It is developed to operate for any speaker of a particular type. Speaker-dependent systems are simpler to construct and are more accurate than speaker-independent systems. As a result, the focus of early voice recognition systems was primarily speaker-dependent isolated word systems that used limited vocabulary. At the time, overcoming the restrictions in the state of technology required a greater focus on human-to-computer interaction. The challenge was to identify how improved speech recognition technology could be used to support the enhancement of human interaction with machines. An important element in the creation of speech recognition system is the size of the vocabulary. Vocabulary affects the complexity and accuracy of the system. Size of the vocabulary can be small, medium or large. Obviously, it is much easier to look up one of 50 words in 50-word dictionary rather than one of the hundreds of thousands of words in a Webster's dictionary. Another important qualifier in the determination of the complexity of a speech recognition system is the type of speech that the recognition system uses: discrete or continuous. In a discrete speech system, the user must pause between each word which makes speech recognition task much easier. Continuous speech is more difficult because of several reasons. First, it is difficult to find the start and end boundary of words. Another problem is that the production of surrounding phonemes affect the production of each phoneme. Also speech rate affect the recognition of continuous speech.

2.4 Difficulties with ASR

Following are the some of the difficulties with ASR

2.4.1 *Human comprehension of speech*

Human use the knowledge about the speaker and the subject while listening more than the ears. Words are not sequences together arbitrarily but there is a grammatical structure that human use to predict words not yet uttered. In ASR, we have only speech signal. We can construct a model for grammatical structure and use some statistical model to improve prediction but there are still the problem how to model world knowledge

and the knowledge of speaker. Of course, we can not model world knowledge but the question is how much we actually require to measure up to human comprehension in the ASR.

2.4.2 Spoken language is not equal to written language

Spoken language is less critical than written language and human make much more performance error while speaking. Speech is two-way communication as compared to written communication which is one-way. The most important issue is disfluencies in speech e.g. normal speech contains repetitions, tongue slips, change of subject in the middle of an utterance etc. In the last few years, it has become clear that spoken language is different from written language. In ASR, we have to identify and address these differences.

2.4.3 Noise

Speech is uttered in an environment of sound such as ticking a clock, another speaker in the background, TV playing in another room etc. This unwanted information in the speech signal is known as noise. In ASR, we have to identify these noise and filter out it from the speech signal. Echo effect is another kind of noise in which speech signal is bounced on some surrounding object and it appears in the microphone a few milliseconds later.

2.4.4 Body Language

A human speaker not only communicates with speech but also with body gestures such as waving hands, moving eyes, postures etc. In ASR, such information is not available. This problem is addressed in Multimodality research area where studies are conducted to incorporate body language to improve human-machine communication.

2.4.5 Channel Variability

Variability is the context in which acoustic wave is uttered. Different types of microphones, noise that changes over time and anything that affects the content of acoustic wave from the speaker to the discrete representation in a computer is known as channel variability.

2.4.6 Speaker Variability

Human speak differently. The voice is not only different between speakers but also wide variation within one specific speaker. Some of the variations are given below

2.4.6.1 Speaking Style

All speakers speak differently due to their unique personality. They have a distinctive way to pronounce words. Speaking style varies in different situations. We do not speak in the same way in the public area, with our teachers or friends. Humans also express emotions while speaking i.e happy, sad, fear, surprise or anger.

2.4.6.2 Realization

If same word were pronounced again and again, the resulting speech signal would never be same. There will be some small differences in the acoustic wave produced. Speech realization changes over time.

2.4.6.3 Speaker Sex

Male and Female have different voices. Female have shorter vocal tract than male. That is why the pitch of female voice is roughly twice than male.

2.4.6.4 Dialects

Dialects are group related variations with in a language.
Regional dialect: It involves features of vocabulary, pronunciation and grammar according to geographical area the speaker belongs.
Social dialect: It involves features of vocabulary, pronunciation and grammar according to social group of speaker.
In many cases, we may be forced to consider dialects as another language because of the large differences between two dialects. We have considered some of the difficulties of speech recognition but not all. The most problematic issue is strong variability. Our goal is efficient user interface not natural verbal communication.

3 SPEECH RECOGNITION TECHNIQUES CLASSIFICATION

Basically there are three approaches to speech recognition.

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Artificial Intelligence Approach

3.1 Acoustic Phonetic Approach

Acoustic phonetic approach (Hemdal and Hughes 1967), postulates that there exist distinguishing phonetic units in spoken language and these units are characterized by a set of acoustics properties. The acoustic properties of phonetic units are highly variable both with speakers and with neighbouring sounds i.e co-articulation effect, it is understood in this approach that the rules governing the variability are straightforward and can be readily learned by a machine. The first stage in this approach is the speech spectral analysis .The next stage is a feature extraction stage that converts the spectral measurements into a set of features that describe the acoustic properties of the different phonetic units. The next stage is a segmentation and labelling stage in which the speech signal is segmented into isolated regions, followed by attaching one or more phonetic labels to each segmented region. The last stage finds a valid word from the phonetic label sequences produced by the segmentation to labelling. The acoustic phonetic approach has not been widely used .

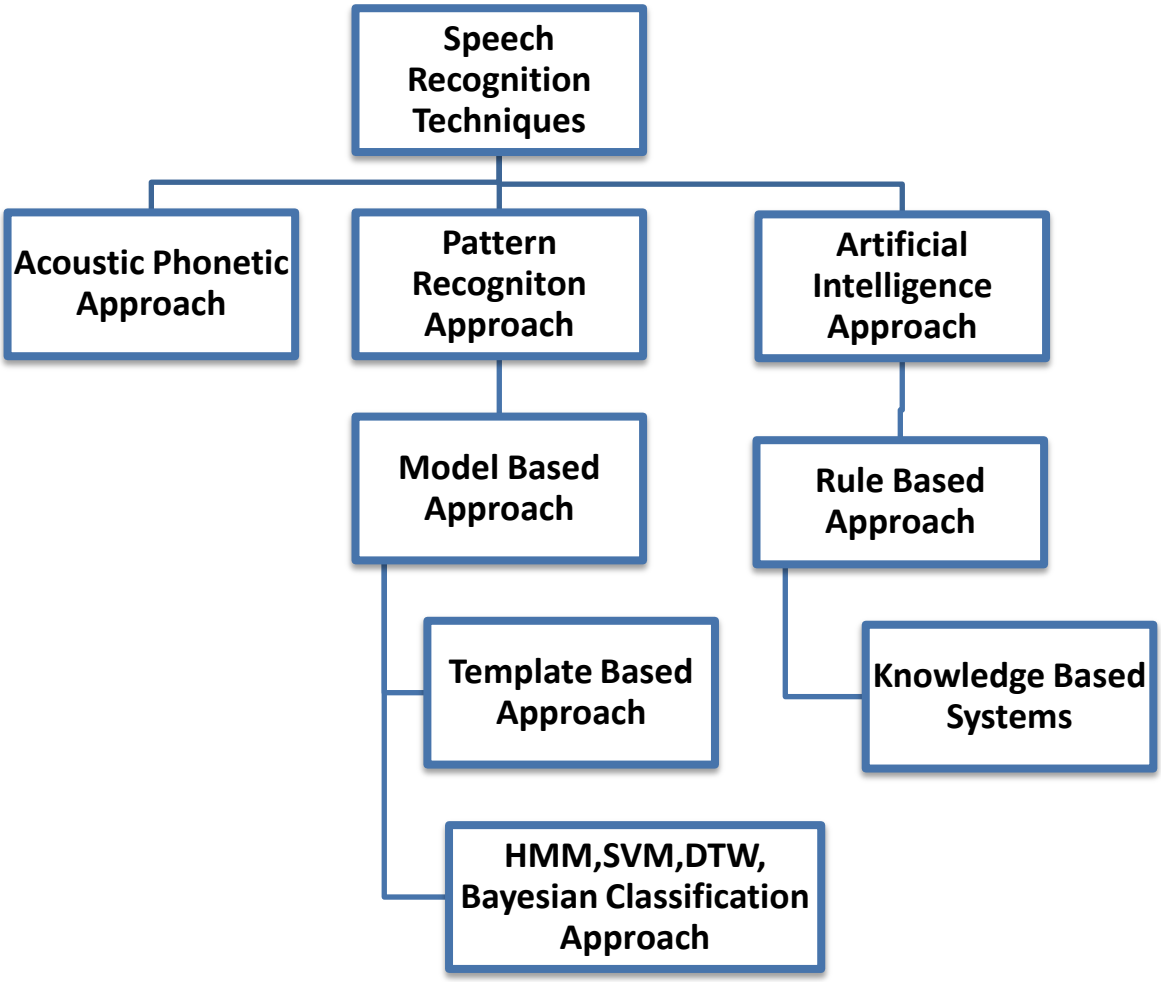


Figure 3 Speech Recognition Techniques Classification

3.2 Pattern Recognition Approach

The pattern-matching approach (Rabiner and Juang 1993) shown in Figure 4 is used to extract patterns based on certain conditions and to separate one class from the others. .It has four stages i.e. feature measurement, pattern training , pattern classification and decision logic. In order to define a test pattern, a sequence of measurements is made on the input signal. Reference pattern is created by taking one or more test patterns corresponding to speech sounds. This can be in the form of a speech template or a statistical model i.e.HMM and can be applied to a sound , a word, or a phrase. In the pattern classification stage of the approach, a direct comparison is made between the unknown test pattern and class reference pattern and a measure of distance between them is computed. Decision logic stage determines the identity of the unknown according to the integrity of match of the patterns. This approach has become the prime method for speech recognition in the last six decades . A block diagram of this approach is shown below. There exists two methods i.e. template method and stochastic model method.

3.2.1 Template Method

In this method, unknown speech is compared against a set of templates in order to find the best match. During the last six decades, template based method to speech recognition have provided a family of techniques. A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate's words. Recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. This method provides good recognition performance for a variety of practical applications. But it has the disadvantage that variations in speech can be modelled by using many templates per word, which eventually becomes unfeasible.

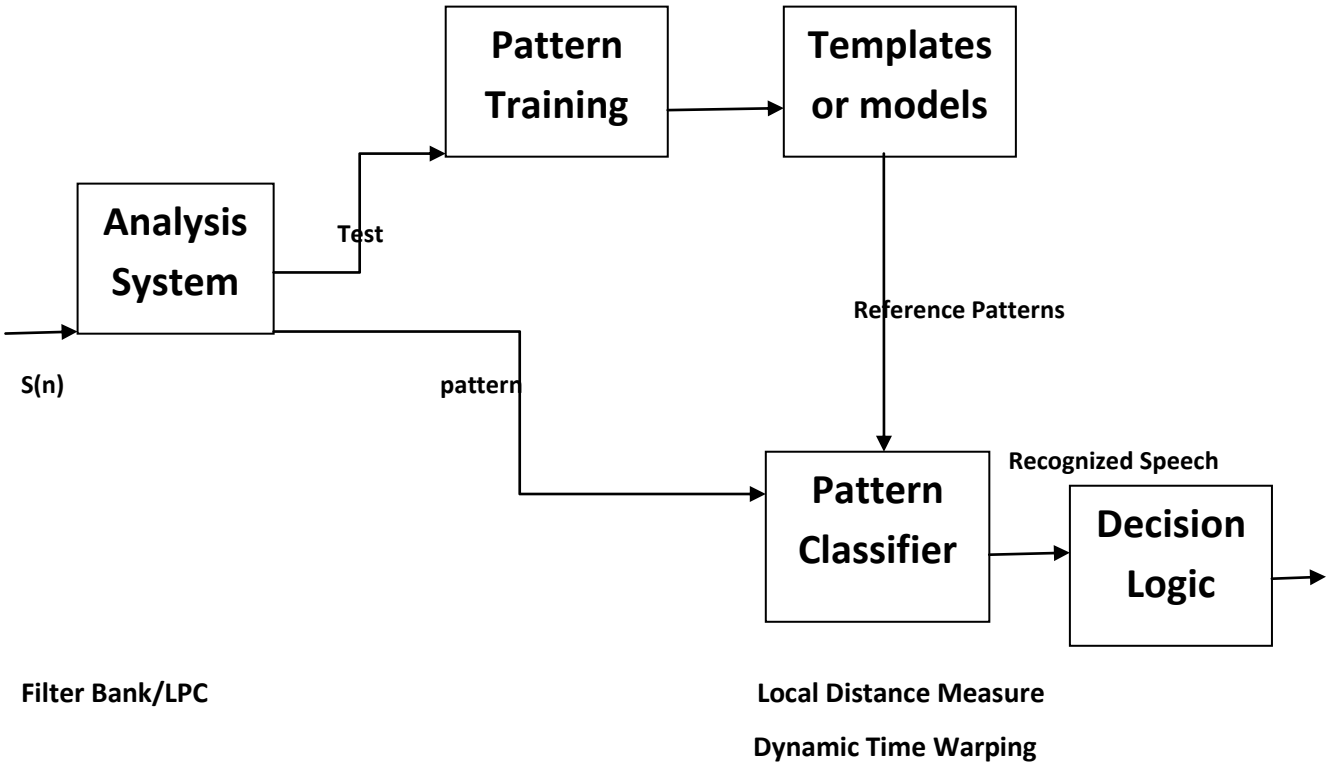
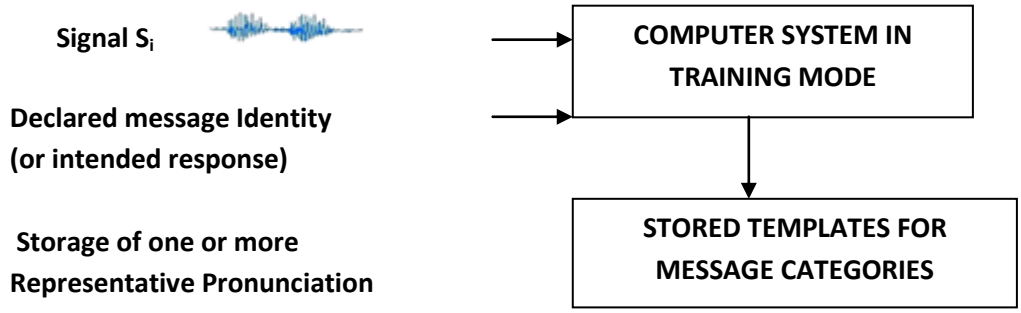


Figure 4 Pattern Recognition Approach

Training System :



Recognition System :

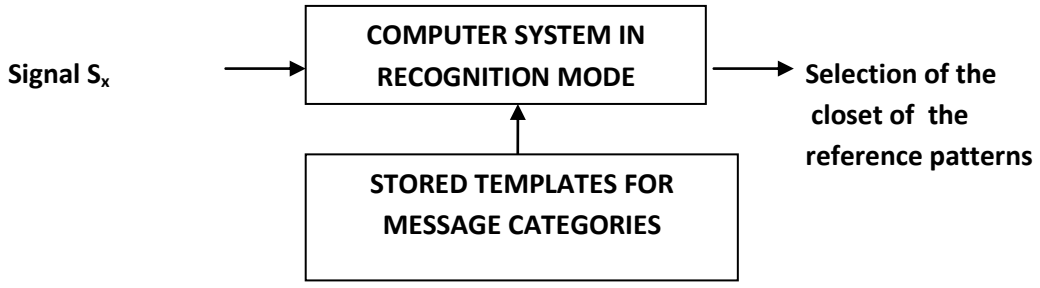


Figure 5 Template matching Method

3.2.2 Stochastic method

Stochastic methods depict the use of probabilistic models to deal with incomplete information. In speech recognition, incompleteness arises from various sources. e.g. speaker variability, contextual effect etc. The most popular stochastic approach now a day is hidden Markov modelling (HMM). A HMM is characterized by a finite state Markov model and a set of output distributions. The transition parameters in the Markov models are temporal variability's, while the parameters in the output distribution model are spectral variability's. These two types of variabilites are the core of speech recognition. It is more general as compared to template based approach. The primary problems for HMM design are a) The evaluation of the probability of a sequence of observations given a specific HMM. b) The determination of a best sequence of modal states and c) the adjustment of modal parameters so as to best account for the observed signal. Once these fundamental problems are solved, we can apply HMMs to selected problems in speech recognition.

3.3 Artificial intelligence approach

The Artificial Intelligence approach is a fusion of the acoustic phonetic approach and pattern recognition approach. Some researchers developed recognition system by using acoustic phonetic knowledge to develop classification rules for speech sounds. While template based methods provided little insight about human speech processing, but these methods have been very effective in the design of a variety of speech recognition systems. On the other hand, linguistic and phonetic literature provided insights about human speech processing. However, this approach had only partial success due to the complicatedness in quantifying expert knowledge. Another problem is the integration of levels of human knowledge i.e. phonetics, lexical access, syntax, semantics and pragmatics.

4 PHASES OF ASR

Automatic speech recognition system involves two phases: Training phase and recognition phase. A rigorous training procedure is followed to map the basic speech unit such as phone, syllable to the acoustic observation. In training phase, known speech is recorded, pre-processed and then enters the first stage i.e. Feature extraction. The next three stages are HMM creation, HMM training and HMM storage. The recognition phase starts with the acoustic analysis of unknown speech signal. The signal captured is converted to a series of acoustic feature vectors. Using appropriate algorithm, the input observations are processed. The speech is compared against the HMM‟s networks and the word which is pronounced is displayed. An ASR system can only recognize what it has learned during the training process. But, the system is able to recognize even those words, which are not present in the training corpus and for which sub-word units of the new word are known to the system and the new word exists in the system dictionary.

4.1 Modules of ASR

Modules for a speech recognition system are shown in Figure 6.

i. Speech Signal acquisition
ii. Feature Extraction
iii. Acoustic Modelling
iv. Language & Lexical Modelling
v. Recognition

Two of these modules Speech acquisition and Feature extraction are common to both the phases of ASR.

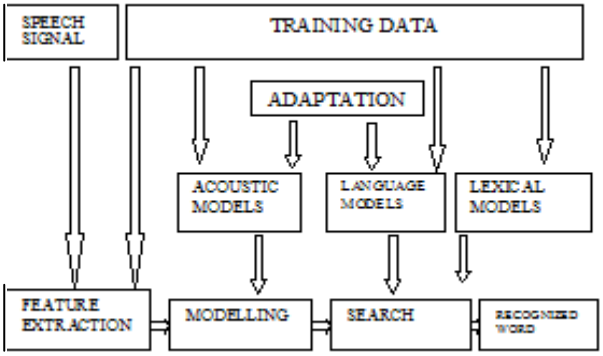


Figure 6 ASR block diagram

Model adaptation is meant for minimizing the dependencies on speakers‟ voice, acoustic environment, microphones and transmission channel, and to improve the generalization capability of the system.

4.1.1 Feature Extraction

Feature extraction requires much attention because recognition performance relies heavily on the feature extraction phase. Different techniques for feature extraction are LPC, MFCC, AMFCC, RAS, DAS, ΔMFCC, Higher lag autocorrelation coefficients, PLP, MF-PLP, BFCC, RPLP etc. It has been found that noise robust spectral estimation is possible on the higher lag autocorrelation coefficients. Therefore, eliminating the lower lags of the noisy speech signal autocorrelation leads to removal of the main noise components.

4.1.2 Acoustic Model

Acoustic model is the main component for an ASR which accounts for most of the computational load and performance of the system. The Acoustic model is developed for detecting the spoken phoneme. Its creation involves the use of audio recordings of speech and their text scripts and then compiling them into a statistical representation of sounds which make up words.

4.1.3 Lexical Models

To provide the pronunciation of each word in a given language, Lexicon is developed. Various combinations of phones are defined through lexicon model to give valid words

for the recognition. Neural networks have helped to develop

4.1.4 Language Models

Language model is the main component operated on million of words, consisting of millions of parameters and with the help of pronunciation dictionary ,developed word sequences in a sentence. ASR systems uses *n*-gram language models which are used to search for correct word sequence by predicting the likelihood of the *n*th word on the basis of the *n*−1 preceding words. The probability of occurrence of a word sequence *W* is calculated as:

$$P(W) = P(w_1, w_2, \ldots\ldots\ldots\ldots, w_{m-1}, w_m) = P(w_1).P(w_2|w_1).P(w_3|w_1w_2)\ldots\ldots\ldots\ldots\ldots P(w_m|w_1w_2w_3\ldots\ldots\ldots\ldots w_{m-1})$$

For large vocabulary speech recognizers, two problems occur during the construction of n-gram language models. For real applications, large amount of training data generally leads to large models. Second is the sparseness problem, is being faced during the training of domain specific models. Language models are non-deterministic. Both these features make it complicated.

5 LITERATURE SURVEY

Speech recognition came into existence during 1920. The first machine i.e. Radio Rex ,a toy to recognize voice was manufactured. Bell Labs developed a speech synthesis machine at the World fair in New York. But later on they discarded efforts based on an incorrect conclusion that the AI is ultimately required for success. In order to develop systems for ASR, attempts were made in 1950s where researchers studied the fundamental concepts of phonetic-acoustic. Most of the systems in 1950[1] for recognizing speech examines the vowels spectral resonancews of each utterances. At Bell Labs Davis, Biddulph and Balashek(1952) premeditated a isolated digit recognition system for a single speaker[2] using formant frequencies estimated during vowel regions of each digit. At RCA Labs, Olson and Belar (1950) built 10 syllables recognizer of a single speaker[3] and Forgie and Forgie built a speaker-independent 10-vowel recognizer [4] at MIT Lincoln Lab, by measuring spectral resonances for vowels. Fry and Denes(1959) tried to build a phoneme recognizer to recognize four vowels and nine consonants [5] at University College in England by using a spectrum analyser and a pattern matcher to make the recognition decision. Japanese labs entering recognition field in 1960-70. As computers are not fast enough, they designed special purpose H/W as a part of their system. In Tokyo, Nagata et.al described a system of the Radio Research lab, was a H/W vowel recognizer[6]. Another effort was the work of Sakai and Doshita in 1962, of Kyoto University who built a H/W phoneme recognizer[7]. In 1963, Nagata and co-workers at NEC Labs built a the digit recognizer[7]. This led to a long productive research program.

In 1970 , the key focus of research was on isolated word recognition. IBM researchers studied in large vocabulary speech recognition. At AT&T Bell Labs, researchers began speaker independent speech recognition experiments[8]. A large number of clustering algorithms were used to find the number of distinct patterns required to represent words to achieve speaker independent speech recognition. This research has been refined so that the techniques for speaker independent patterns are widely used. Carnegie Mellon University's Harphy system[9] recognize speech with vocabulary size of 1011 words with reasonable accuracy. It was the first to make use of finite state network to reduce computation and determine the closest matching strings efficiently.

In 1980, the key focus of research was on connected words speech recognition. In the beginning of 1980, Moshey J. Lasry studied speech spectrogram of letters and digits[10] and developed a feature based speech recognition. There is a change in technology in 1980 from template based approaches to statistical modelling approach specially HMM [11,12] in speech research. The most significant paradigm shift has been the introduction of statistical methods, especially stochastic processing with HMM(Baker, 1975 & Jelinek, 1976) in the early 1970's (Portiz 1988). More than 30 years later, this methodology still predominates. Despite their simplicity, N-gram language models have proved remarkably powerful. Now days, most practical speech recognition systems are based on statistical approach and their results with additional improvements have been made in 1990s. In 1980, Hidden Markov model(HMM) approach is one of the key technologies developed. IBM, Institute for Defence Analysis (IDA) and Dragon Systems understood HMM , but it was not renowned in the mid-1980s. Neural networks to speech recognition problems is the another technology that was reintroduced in the late 1980s.

In 1990, Pattern recognition approach was developed. It followed Bayes's framework traditionally but it has been altered into an optimization problem with minimization of the empirical recognition error[13]. The reason for this alteration is that the distribution functions for the speech signal could not be chosen accurately and under these conditions, Bayes' theory can not be applied. However, aim is to design recognizer with least recognition error rather than best fitting to given data. The techniques used for error minimization are Minimum Classification error (MCE) and Maximum Mutual Information(MMI). These techniques led to maximum likelihood based approach[14] to speech recognition performance. A weighted HMM algorithm is proposed to address HMM based speech recognition issues of robustness and discrimination,. In order to decrease the acoustic mismatch between given set of speech model and test utterance, a maximum likelihood stochastic matching approach[15] was proposed. A narrative approach [16] for HMM speech recognition system is based on the use of a Neural network as a vector quantizer which is remarkable innovation in training the neural network. Nam Soo Kim et.al. described a variety of methods for estimating a robust output probability distribution based on HMM [17]. An extension of the viterbi algorithm[18] made second order HMM efficient as compared to existing viterbi algorithm. In 1990s, the DARPA program continued. After that the centre of attention is Air Travel Information Service (ATIS) task and later focus on transcription of broadcast news (BN). Advances in

lexical model for non-native speech recognition.

continuous speech recognition and noisy environment speech recognition, have been explained. In the area of noisy robust speech recognition, minor work has been done.For noisy environment, for robust speech recognition, a new approach to an auditory model was proposed[26]. This approach is computationally efficient as compared with other models. A model based spectral estimation algorithm has been developed[27].

In 2000, a variational Bayesian estimation technique was developed[19]. It is based on posterior distribution of parameters. Giuseppe Richardi have developed a technique to solve the problem of adaptive learning[20] in ASR. In 2005, some improvements have been made on Large Vocabulary Continuous Speech recognition system for performance improvement[21]. A 5-year national project Corpus of Spontaneous Japanese (CSJ) [22] was conducted in Japan. It consists of 7 millions of words approximately, corresponding to speech of 700 hours. The techniques used in this project are acoustic modelling, sentence boundary detection, pronunciation modelling, acoustic as well as language model adaptation, and automatic speech summarization [23]. Utterance verification is being investigated [24] to further increase the robustness of speech recognition systems, especially for spontaneous speech,. When humans speak to each other, they use multimodal communication. It increases the rate of successful transfer of information when the communication take place in a noisy environment. In speech recognition, the use of the visual face information, specially lip movement, has been examined, and results show that using both mode of information gives better performances than using only the audio or only the visual information, specially, in noisy environment.

6 TOOLS FOR ASR

Following are the various tools used for ASR

PRAAT: It is free software with latest version 5.3.04 which can run on wide range of OS platforms and meant for recording and analysis of human speech in mono or stereo

AUDACITY: It is free, open source software available with latest version of 1.3.14(Beta) which can run on wide range of OS platforms and meant for recording and editing sounds.

CSL: Computerised Speech Lab is a highly advanced speech and signal processing workstation (software and hardware). It possesses robust hardware for data acquisition and a versatile suite of software for speech analysis.

HTK: The basic application of open source Hidden Markov Toolkit (HTK), written completely in ANSI C, is to build and manipulate hidden Markov models.

SPHINX: Sphinx 4 is a latest version of Sphinx series of speech recognizer tools, written completely in Java programming language. It provides a more flexible framework for research in speech recognition.

SCARF: It is a software toolkit designed for doing speech recognition with the help of segmental conditional random fields.

MICROPHONES: They are being used by researchers for recording speech database. Sony and I-ball has developed some microphones which are unidirectional and noiseless.

7 PERFORMANCE OF SPEECH RECOGNITION SYSTEM

The performance of speech recognition is specified in terms of accuracy and speed. Accuracy is measured in terms of performance accuracy which is known as word error rate (WER) whereas speed is measured with the real time factor.

Word Error Rate

It is a common metric of the speech recognition performance. As recognized word sequence have a different length from the reference word sequence, there is difficulty in measuring performance.

$$\text{WER} = \frac{S + D + I}{N}$$

where S is number of substitutions
D is number of deletions
I is number of insertions
and N is number of words in the reference.

Sometimes word recognition rate (WRR) is used instead of WER while describing performance of speech recognition.

WRR = 1- WER

$$= 1 - \frac{S + D + I}{N}$$

$$= \frac{N - S - D - I}{N}$$

Speed

It is measured by real time factor. If it takes time T to process an input of duration D then real time factor is defined by

$$\text{RTF} = \frac{T}{D}$$

RTF $\leq$1 implies real time processing.

8 SUMMARY

Research in speech recognition has been carried out intensively for the last 60 years. The technological progress can be summarized in Table-1[28]

Table 1. SPEECH RECOGNITION SUMMARIZATION

Sr.No.	Past	Present
1	Template matching	Corpus-based statistical modelling,e.g.HMM, n-gram
2	Distance–based methods	Likelihood –based methods
3	Maximum likelihood	Discriminative approach e.g. MCE/GPD and MMI
4	Isolated word recognition	Continuous speech recognition
5	Small Vocabulary	Large Vocabulary
6	Clean speech recognition	Noisy/Telephone speech recognition
7	Single speaker recognition	Speaker-independent/adaptive
8	Read speech recognition	Spontaneous speech recognition
9	Single modality	Multimodal speech recognition

9 DISCUSSIONS AND CONCLUSIONS

Speech is the most prominent and primary mode of communication between human beings. Over the past five decades, research in the area of speech is a first step towards ordinary man-machine communication. We also have encountered some limitations. What we know about speech processing is very limited. This paper attempts to give a comprehensive survey of research in speech recognition and some year-wise progress to this date and its current status. Although significant amount of work has been done in the last two decades but there is still work to be done.

At present research is focussing on creating and developing systems that would be much more robust against variability and shift in acoustic environment, speaker characteristics, language characteristics, external noise sources etc. It has been found that HMM is the best technique in developing language model. Speech recognition is very fascinating problem. It has attracted scientists and researchers and created a technological bang on society and is expected to boom further in this area of man machine interaction.

10 ACKNOWLEDGEMENT

I would like to thank my husband and son for their unwavering support. My special thanks to Dr. S.S.Agrawal, Dr. R.K.Jain, Dr. Harsh Vardhan Kamrah and Mrs. Neelima Kamrah for their kind support and encouragement.

11 REFERENCES

[1]. Sadaoki Furui, November 2005, 50 years of Progress in speech and Speaker Recognition Research , ECTI Transactions on Computer and Information Technology,Vol.1. No.2.

[2]. K.H.Davis, R.Biddulph, and S.Balashek, 1952, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am.,24(6):637-642.

[3]. H.F.Olson and H.Belar, 1956, Phonetic Typewriter , J.Acoust.Soc.Am.,28(6):1072-1081.

[4]. J.W.Forgie and C.D.Forgie, 1959, Results obtained from a vowel recognition computer program , J.Acoust.Soc.Am., 31(11),pp.1480-1489.

[5] D.B.Fry, 1959, Theoritical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299.

[6]. K.Nagata, Y.Kato, and S.Chiba, 1963, Spoken Digit Recognizer for Japanese Language , NEC Res.Develop., No.6.

[7]. T.Sakai and S.Doshita, 1962 The phonetic typewriter, information processing 1962 , Proc.IFIP Congress.

[8]. L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, and J.G.Wilpon, August 1979, Speaker Independent Recognition of Isolated Words Using Clustering Techniques , IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27:336-349.

[9]. B.Lowrre, 1990, The HARPY speech understanding system ,Trends in Speech Recognition, W.Lea,Ed., Speech Science Pub., pp.576-586.

[10]. R.K.Moore, 1994, Twenty things we still don t know about speech , Proc. CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology.

[11]. J.Ferguson, 1980, Hidden Markov Models for Speech, IDA,Princeton, NJ.

[12]. L.R.Rabiner, February 1989, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition , Proc.IEEE,77(2):257-286.

[13]. B.H.Juang and S.Furui, 2000, Automatic speech recognition and understanding: A first step toward natural human machine communication , Proc.IEEE,88,8,pp.1142-1165.

[14]. K.P.Li and G.W.Hughes, 1974, Talker differences as they appear in correlation matrices of continuous speech spectra , J.Acoust.Soc.Am. , 55,pp.833-837.

[15] Ananth Sankar, May 1996, " A maximum likelihood approach to stochastic matching for Robust Speech recognition", IEEE Transactions on Audio, Speech and Language processing Vol.4,No.3.

[16]. Gerhard Rigoll, Jan.1994, "Maximum Mutual Information Neural Networks for Hybrid connectionist-HMM speech Recognition Systems ", IEEE Transactions on Audio, Speech and Language processing Vol.2,No.1, PartII.

[17].Nam Soo Kim et.al., July1995, On estimating Robust probability Distribution in HMM in HMM based speech recognition , IEEE Transactions on Audio, Speech and Language processing Vol.3,No.4.

[18].Jean Francois, Jan.1997, Automatic Word Recognition Based on Second Order Hidden Markov Models , IEEE Transactions on Audio, Speech and Language processing Vol.5,No.1.

[19].Mohamed Afify and Olivier Siohan, January 2004, Sequential Estimation With Optimal Forgetting for Robust Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol. 12, No. 1.

[20].Giuseppe Riccardi, July 2005, Active Learning: Theory and Applications to Automatic Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4.

[21]. Mohamed Afify, Feng Liu, Hui Jiang, July 2005, A New Verification-Based Fast-Match for Large Vocabulary Continuous Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4.

[22].S. Furui, 2005, Recent progress in corpus-based spontaneous speech recognition, IEICE Trans. Inf. & Syst., E88-D, 3, pp. 366-375.

[23].S. Furui, 2004, Speech-to-text and speech-to-speech summarization of spontaneous speech, IEEE Trans. Speech & Audio Processing, 12, 4, pp. 401-408.

[24].Eduardo Lleida et.al. March 2000, Utterance Verification In Decoding And Training Procedures , IEEE Transactions On Speech And Audio Processing, Vol. 8, No. 2.

[25] Geoff Bristow, 1986, "Electronic Speech recognition: Techniques, Technology and Applications'' , Collins .

[26] Doh-Suk Kim, 1999, " Auditory processing of Speech Signals for Robust Speech Recognition in Real World Noisy Environment", IEEE Transactions on Speech and Audio Processing Vol.7,No.1.

[27] Adoram Erell et.al. ,1993, " Filter bank energy estimation using mixture and Markov models for Recognition of Noisy Speech" IEEE Transactions on Audio, Speech and Language processing Vol.1,No.1.

[28] M.A.Anusuya, S.K.Katti, 2009, "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security, vol. 6, No. 3.