ISSN No. 0976-5697

Volume 2, No. 6, Nov-Dec 2011

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Automatic Speech Recognition: Architecture, Methodologies and Challenges - A Review

S.Karpagavalli*
Senior Lecturer,
Department of Computer science (PG),
PSGR Krishnammal College for Women, Coimbatore.
karpagam@grgsact.com

R.Deepika, P.Kokila, K.Usha Rani
Mphil Research Scholar,
Department of Computer science (PG),
PSGR Krishnammal College for Women, Coimbatore.
dpi.feb88@gmail.com saikokila87@gmail.com
dhanalakshmi27@gmail.com

Dr.E.Chandra
Research Supervisor,
DJ Academy for Managerial Excellence,
Coimbatore, India

Abstract: For more than three decades, a great amount of research was carried out on various aspects of speech signal processing and its applications. Highly successful application of speech processing is Automatic Speech Recognition (ASR). Early attempts to ASR consisted of making deterministic models of whole words in a small vocabulary and recognizing a given speech utterance as the word whose model comes closest to it. The introduction of Hidden Morkov Models (HMMs) in the early 1980 provided much more powerful tool for speech recognition. And the recognition can be done for continuous speech using large vocabulary, in a speaker independent manner. Today many products have been developed that successfully utilize ASR for communication between human and machines. Performance of speech recognition applications deteriorates in the presence of reverberation and even low levels of ambient noise. Robustness to noise, reverberation and characteristics of the transducer is still an unsolved problem that makes the research in the area of speech recognition still very active. A detailed study on ASR carried out and presented in this paper that covers the basic model of speech recognition, applications, feature analysis, various models used in speech recognition and challenges.

Keywords: Automatic Speech Recognition, feature extraction, performance evaluation, speaker independent, large vocabulary

I.INTRODUCTION

A. Definition of Speech Recognition:

Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting acoustic signal captured by microphone or telephone to a set of words.

a. Mathematical Representation:

Fundamentally, the problem of speech recognition can be stated as follows. When given with acoustic observation O = $o_1o_2...o_t$, the goal is to find out the corresponding word sequence $W = w_1w_2...w_n$ that has the maximum posterior probability $P(W|O)$ can be written as

$$\hat{W} = \arg\max_{W \in L} P(W \mid O) \quad \ldots\ldots (1)$$

Equation 1 can be expressed using Bayes rule as

$$\hat{W} = \arg\max_{W \in L} \frac{P(O \mid W)P(W)}{P(O)} \quad \ldots\ldots (2)$$

Since the P(O) is the same for each candidate sentence W, thus equation 2 can be reduced as

$$\hat{W} = \arg\max_{W \in L} P(O \mid W)P(W) \quad \ldots\ldots (3)$$

Where $P(W)$, the prior probability of word W uttered is called the language model and $P(O|W)$, the observation likelihood of acoustic observation O when word W is uttered is called the acoustic model [1].

B. Issues in Speech Recognition:

There are number of issues that need to be addressed in order to define the operating range of each speech recognizing systems that is built. Some of them are, speech unit like word, syllable, phoneme or phones used for recognition, vocabulary size like small, medium and large, task syntax like simple to complex task using N-gram language models, task perplexity, speaking mode like isolated, connected, continuous, spontaneous, speaker mode like speaker trained, adaptive, speaker independent, dependent, speaking environment as quiet room, noisy places, transducers may be high quality microphone, telephones, cell phones, array microphones, and also transmission channel [2].

II . TYPES OF SPEECH RECOGNITION

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize [3]. These various classes of ASR are,

A. Isolated Words:

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time.

B. Connected Words:

Connected word systems are similar to isolated words, but allow separate utterances to be run-together with a minimal pause between them.

C. Continuous Speech:

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

D. Spontaneous Speech:

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

III. SPEECH RECOGNITION ARCHITECTURE

The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operator services [1]. This report reviews major highlights during the last six decades in the research and development of automatic speech recognition, so as to provide a technological perspective. Although many technological progresses have been made, still there remain many research issues that need to be tackled.

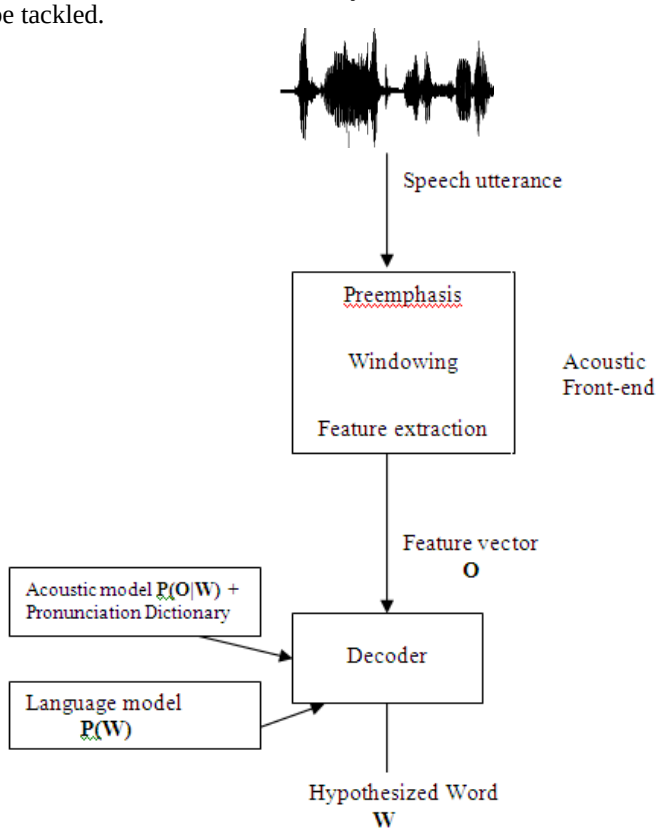

Figure 1. Speech Recognition Architecture.

A. Sampling:

In speech recognition, Common sampling rates are 8 KHz to 16 KHz, to accurately measure a wave it is necessary to have at least two samples in each cycle: one measuring the positive part of the wave and one measuring the negative part. More than two samples per cycle increases the amplitude accuracy, but less than two samples will cause the frequency of the wave to be completely missed. Thus the maximum frequency wave that can be measured in one whose frequency is half the sample rate. Most information in human speech is in frequencies below 10 KHz; thus a 20 KHz sampling rate would be necessary for complete accuracy. But the switching network filters telephone speech and only frequencies less than 4 KHz are transmitted by telephones. Thus an 8 KHz sampling rate is sufficient for telephone/ mobile speech corpus. For other applications commonly 16 KHz sampling rate is used. Tools like praat, wave surfer, sound forge, audacity, sonic visualizer, tartini can be used to record speech data.

B. Pre-Emphasis:

The spectrum for voiced segments has more energy at lower frequencies than higher frequencies. Pre-emphasis is boosting the energy in the high frequencies. Boosting high-frequency energy gives more information to Acoustic Model and improves the recognition performance.

C. Windowing and Framing:

The time for which the signal is considered for processing is called a window and the data acquired in a window is called as a frame. Typically features are extracted once every 10ms, which is called as frame rate. The window duration is typically 25ms. Thus two consecutive frames have overlapping areas. There are different types of windows like Rectangular window, Bartlett window, and Hamming window. Out of these the most widely used window is hamming window as it introduces the least amount of distortion.

D. Feature Extraction:

The feature extraction is the first stage to extract feature vectors. Some of the feature extraction methods are as follows,

a. Principle Component Analysis (PCA) method
b. Linear Discriminant Analysis (LDA) method
c. Independent Component Analysis (ICA) method
d. Linear Predictive Coding (LPC) method
e. Cepstral Analysis method.
f. Mel-Frequency Scale Analysis method.
g. Filter-Bank Analysis method.
h. Mel-Frequency Cepstral Coefficients (MFCC) method.
i. Kernal Based Feature Extraction Method.
j. Dynamic Feature Extraction.
k. Wavelet.
l. Spectral Subtraction.
m. Cepstral Mean subtraction.

Most speech recognition systems use the so–called Mel frequency cepstral coefficients (MFCC) and its first and sometimes second derivative in time to better reflect dynamic changes.

E. Mel Frequency Cepstral Coefficients:

These are coefficients based on the Mel scale that represent sound. The word cepstral comes from the word cepstrum, which is a logarithmic scale of the spectrum (and

reverses the first four letters in the word spectrum). First, the speech data are divided into 25 ms windows (frames).

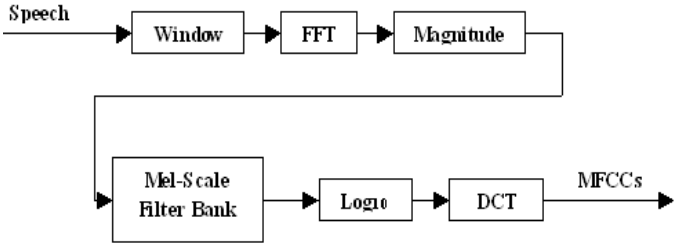

Figure 2. Block Diagram of MFCC

A new frame is started every 10 ms making this the sampling period and causing the windows to overlap each other. Next, the fast Fourier transform is performed on each frame of speech data and the magnitude is found. The next step involves filtering the signal with a frequency-warped set of log filter banks called Mel-scale filter banks [4] [5]. The log filter banks are arranged along the frequency axis according to the Mel scale, a logarithmic scale that is a measure of perceived pitch or frequency of a tone [6], thus simulating the human hearing scale. The Mel scale yields a compression of the upper frequencies where the human ear is less sensitive. Next, the logarithm is taken of the log filter bank amplitudes. Finally, the MFCCs are calculated using the discrete cosine transform (DCT). To further enhance speech recognition performance, an extra set of delta and acceleration coefficient features are sometimes calculated with MFCCs. These features are the first and second time derivatives of the original coefficients, respectively.

F. Acoustic Model:

Acoustic modeling is an important component in modern state-of-the-art automatic speech recognition (ASR) systems. The goal of acoustic modeling is to build robust statistical models for the acoustic properties of speech signals. Despite many years of research and significant advances, the acoustic models in ASR systems are far from the performance of their biological counterparts.

Traditionally, the parameters of these probabilistic acoustic models are learnt from large speech corpora with the technique of maximum likelihood (ML) estimation. The ML estimation technique finds parameters that maximize the joint likelihood over training data, namely, acoustic features extracted from speech signals and their labels (words or phonemes). For many probabilistic models, especially for those most commonly used in ASR, the joint likelihood can be maximized with simple and efficient update procedures such as the Expectation-Maximization (EM) algorithm. A weakness often cited for the ML estimation framework is that the joint maximum likelihood criterion does not directly optimize word or phoneme recognition error rates, which are more relevant metrics for ASR. Other techniques, including conditional maximum likelihood (CML) estimation, minimum classification error (MCE) and maximum mutual information (MMI) estimation also studied by researchers that optimize discriminative criteria that more closely track actual error rates, as opposed to the ML estimation. These techniques do not enjoy the simplicity and relatively fast convergence of the EM algorithm, but if carefully and skillfully implemented, they lead to lower error rates. Presently, the discriminative learning techniques have achieved significantly better results than baseline systems that are typically trained by the maximum likelihood estimation. SVMs have also been used as acoustic models and integrated with hidden Markov models for tasks in ASR [4].

G. Language Model:

In speech recognition, acoustic model and lexicon produce the optimal sequence of the words that compose the systems final output. Rules are introduced at this stage to describe the linguistic restrictions present in the language and to allow reduction of possible invalid phoneme sequences. This is accomplished through the use of language models that estimate the probability of sequences of words. Common language models are bigram and trigram models. These models contain computed probabilities of groupings of two or three particular words in a sequence, respectively. There are tools for language modeling like CMU Statistical Language Modeling (SLM) Toolkit, Stanford Research Institute Language Modeling Toolkit.

IV. SPEECH PROCESSING CLASSIFICATION

The following tree structure emphasizes the speech processing applications. In speech processing, speech recognition, speaker recognition and language identification comes under recognition task. Speech recognition can be further classified according to the size of the vocabulary/ speaker mode/ speech mode/ speaking style as small, medium and large vocabulary continuous/ isolated and speaker independent/ speaker dependent/ speaker adaptive speech recognition systems [3]. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in figure 3.

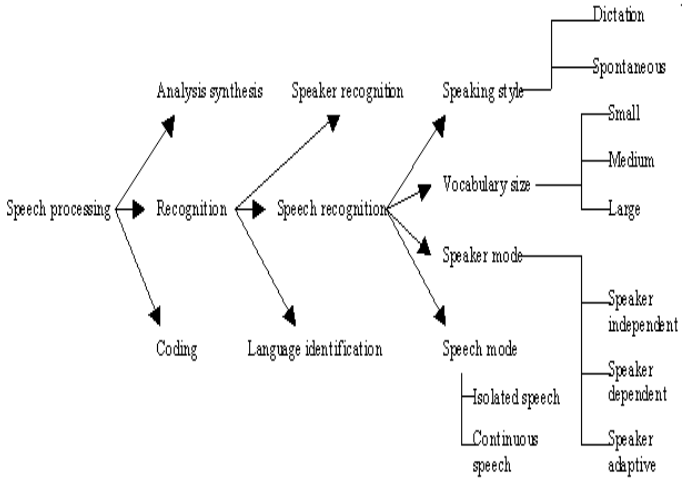

Figure 3. Speech processing classification

V. ASR METHODOLOGIES

Speech recognition research has been ongoing for more than 70 years. Over the period, there have been at least four generations of approaches and a forecast of fifth generation that is being formulated based on current research themes [2]. The five generations, and the technology themes associated with each of them are as follows:

a. ***Generation 1 (1930s to 1950s):*** Use of ad hoc methods to recognize sounds, or small vocabularies of isolated words.

b. Generation 2 (1950s to 1960s): Use of acoustic phonetic approaches to recognize phonemes, phones or digit vocabularies.

c. Generation 3 (1960s to 1980s): Use of pattern recognition approaches to speech recognition of small or medium sized vocabularies of isolated and connected word sequences, including use of linear predictive coding (LPC) as the basic methods of spectral analysis; Use of LPC distance measures for pattern similarity scores; Use of dynamic programming methods for time aligning patterns; Use of pattern recognition methods for clustering multiple patterns into consistent reference patterns; Use of Vector Quantization(VQ) code book methods for data reduction and reduced computation.

d. Generation 4 (1980s to 2000s): Use of Hidden Markov Model (HMM) statistical methods for modeling speech dynamics and statistics in a continuous speech recognition system; Use of forward-backward and segmental k-means training methods; Use of Viterbi alignment methods; Use of maximum likelihood and various other performance criteria and methods for optimizing statistical methods; Introduction of neural network methods for estimating conditional probability densities; Use of adaptation methods that modify the parameters associated with either the speech signal or the statistical model so as to enhance the compatibility between model and data for increased recognition accuracy.

e. Generation 5 (2000s to 2020s): Use of parallel processing methods to increase recognition decision reliability; Combinations of HMMs and acoustic phonetic approaches to detect and correct linguistic irregularities; increased robustness for recognition of speech in noise; Machine learning of optimal combination of models.

The main methodologies that made significant change in the speech recognition area are elaborated below.

A. Acoustic Phonetic Approach:

This is the basis of the acoustic phonetic approach, which postulates that there exist finite, distinctive phonetic units in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word from the phonetic label sequences produced by the segmentation to labeling [5].

B. *Pattern Recognition Approach:*

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns [5].

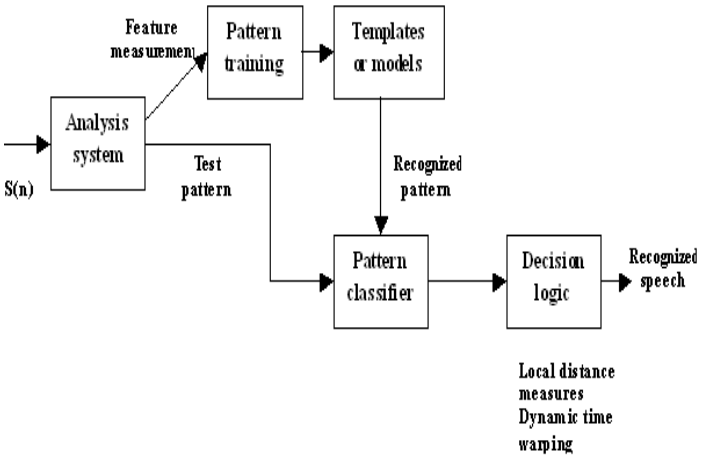

Figure 4. Block Diagram of Pattern Recognition Approach

C. Template Based Approach:

The term template is often used for two fundamentally different concepts: either for the representation of a single segment of speech with a known transcription, or for some sort of average of a number of different segments of speech. Both types of templates can be used in the DTW algorithm to compare them with a segment of input speech . It has a sequence of consecutive acoustic feature vectors, a transcription of the sounds or words it represents, knowledge of neighboring templates, a tag with meta-information. Template based approaches, in which unknown speech is compared against a set of prerecorded words (template) in order to find the best match. This has the advantage of using perfectly accurate word models; but it also has the disadvantage that the prerecorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. When considering the concrete implementation of template-based recognition, it quickly becomes apparent that the classical DTW algorithm with the Euclidean distance used as local distance metric, combined with a simple beam search will not do the job, neither from a performance nor from a computational point of view. Development of isolated word speech recognition system is based on a use of dynamic time warping (DTW) for speech pattern matching. The DTW process nonlinearly expands or contracts the time axis to match the same phoneme positions between the input speech and reference templates [7] [10].

D. Support Vector Machine (SVM):

Support Vector Machines are a comparatively new approach to the problems of classification, regression, ranking, etc [7]. While Artificial Neural Networks (ANNs) are widely used, Support Vector Machines (SVMs) are a comparatively new and efficient pattern recognition tool. SVMs are fast in training and guarantee a global optimum if the kernel satisfies Mercer's condition but require an appropriate choice of kernel function. ANNs are slow in training and can only guarantee local optima; the most successful solution seems to be using larger databases, trying to embed in the training set all the

variability of speech and speakers. In particular, some alternative approaches, most of them based on Artificial Neural Networks [8]. Most implementations of SVM algorithm require computing and storing in memory the complete kernel matrix of all the input samples.

E. Artificial Neural Network:

Recent work on neural networks raises the possibility of new approaches to the speech recognition problem. Their use of many processors operating in parallel may provide the computational power required for continuous-speech recognition. New neural net algorithms self-organize and build an internal speech model that maximizes performance. Auto Associative Neural Network (AANN) models for the task of speaker verification and speech recognition, which produce comparable performance with that of GMM based speaker verification and speech recognition. There exists a relationship between principal component analysis and weights learned by a 3-layer AANN model. AANN model has been mostly used in applications involving dimensionality reduction [8].

F. Vector Quantization:

Vector Quantization is a clustering technique that neglects the temporal information contained in a word in order to avoid the need for time alignment. The design of a vector quantization (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. Given a vector source with its statistical properties known, given a distortion measure and given the number of code vectors, find a codebook and a partition, which result in the smallest average distortion. During the recognition phase the feature vectors extracted from the test word are compared to all reference codebooks. The codebook that produces the minimum distortion determines the spoken word [9][10].

G. Hidden Morkov Model:

HMM is very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of application. The introduction of Hidden Morkov Models (HMMs) in the early 1980 provided much more powerful tool for speech recognition. The elements of HMM is characterized by following:

a. Number of state N
b. Number of distinct observation symbol per state
c. State transition probability,
d. Observation symbol probability distribution in state
e. The initial state distribution

The Three Basic Problems for HMMs are,

Problem 1: Evaluation Problem -Given the observation sequence $O = O_1\ O_2 \cdots O_T$, and model $\lambda = (A,B,\pi)$, how do we efficiently compute $P(O|\ \lambda)$, the probability of observation sequence given the model.

Problem 2: Hidden State Determination (Decoding) -Given the observation sequence $O = O_1\ O_2 \cdots O_T$, and model $\lambda = (A, B, \pi)$ how do we choose corresponding state sequence $Q = q_1\ q_2 \cdots q_T$ which is optimal in some meaningful sense.

Problem 3: Learning -How do we adjust the model parameter $\lambda = (A, B, \pi)$, to maximize $P(O|\ \lambda)$. Problem 3 is one in which we try to optimize model parameter so as to best describe as to how given observation sequence comes out.

Solution to three problems of HMM - Using Forward Algorithm for Evaluation Problem, Viterbi Algorithm for Decoding Hidden State Sequence P (Q, O| λ) and Baum-Welch Algorithm for Learning the three problem associated with HMM are solved [11]. The different types of HMM are like Context-Independent Phoneme HMM, Context-Dependent Triphone HMM Whole-Word HMM. Also many refinements are made to improve the performance of ASR systems.

In future with the use of parallel processing methods, Machine learning of optimal combination of models will increase recognition decision reliability; correct linguistic irregularities and improve the performance of speech recognition applications in greater level.

VI. PERFORMANCE OF SPEECH RECOGNITION SYSTEM

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy, which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

A. Word Error Rate (WER):

The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed in equation.

$$\text{Word error rate (\%)} = (100)\,\frac{\text{Insertion(I)+Substitution(S)+Deletion(D)}}{\text{No of reference words (N)}} \quad \ldots (4)$$

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead:

$$\text{WRR=1-WER=}\frac{\text{N-S-D-I}}{\text{N}} \quad \ldots\ldots\ldots \quad (5)$$

VII. APPLICATIONS IN SPEECH RECOGNITION

Various applications of speech recognition domain have been discussed below.

a. Automatic call processing in telephone networks
b. Teaching students of foreign languages to pronounce vocabulary correctly.
c. Computer and video games, gambling, precision surgery.
d. Query based information systems that provide updated travel information, stock price quotations
e. High performance fighter aircraft, battle management, telephony and other domains.
f. Medical transcriptions (digital speech to text).
g. Multi modal interacting, court reporting, physically handicapped.
h. Hands and eyes free applications
i. Data entry, voice dictation and Speech transcription

VIII. CHALLENGES IN SPEECH RECOGNITION

a. Accurately and efficiently convert a speech signal into a text message independent of the device, speaker or the environment.
b. Automatic generation of word lexicons.
c. Automatic generation of language models for new tasks.
d. Finding the theoretical limit for implementation of automatic speech recognition.
e. Optimal utterance verification-rejection algorithm.
f. Achieving or surpassing human performance on ASR tasks**.**

CONCLUSION

Speech is the primary, and the most convenient means of communication between people. Building automated systems to perform spoken language understanding as well as recognizing speech, as human being do is a complex task. The goal of automatic speech recognition research is to address the various issues relating to speech recognition. Various methodologies are identified and applied to ASR area, which led to many successful ASR applications in limited domains. Robust speech recognition, Multimodal speech recognition, Multilingual speech recognition are some of the research areas gaining momentum. In future we can expect many more ASR applications with improved performance.

REFERENCES

[1]. Daniel Jurafsky, James H. Martin, "Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", (2002) Pearson Education.

[2]. Benesty Jacob, Sondhi M.M, Huang Yiteng, "Springer Handbook of Speech Processing"(2008) Springer.

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

[4]. Kamm, Terri, Hynek Hermansky, and Andreas G. Andreou "Learning the Mel-scale and Optimal VTN Mapping", Johns Hopkins University, Center for Language and Speech Processing, workshop (WS97), (1997).

[5]. Rabiner, Lawrence and Biing-Hwang Juang "Fundamentals of Speech Recognition", Prentice-Hall, Inc., (Engelwood, NJ), (1993).

[6]. Fei Sha "Large Margin Training of Acoustic Models For Speech Recognition", A Dissertation in Computer and Information Science, (2007).

[7]. Li Deng, Helmer Strik, "Structure-Based and Template-Based Automatic Speech Recognition-Comparing parametric and non-parametric approaches", Microsoft Research, One Microsoft Way, Redmond, WA, USA, CLST, Department of Linguistics, Radboud University, Nijmegen, the Netherlands, (2007).

[8]. R. Solera-Ure˜na, J. Padrell-Sendra, D. Mart´ın-Iglesias, A. Gallardo-Antol "SVMs for Automatic Speech recognition" Avda. de la Universidad, 30, 28911-Legan´es (Madrid), SPAIN.

[9]. Richard P. Lippmann "Neural Network Classifiers for Speech Recognition" the Lincoln Laboratory Journal, Volume 1, Number 1, (1988).

[10]. Ganesh K Venayagamoorthy, Viresh Moonasar and Kumbes Sandrasegaran, "Voice Recognition Using Neural Networks", Institute for Information Sciences and Technology (IIST), Massey University, New Zealand, Electronics Engineering Department, M L Sultan Technikon, Durban, South Africa, March (2009).

[11]. Lawrence R. Rabiner "A tutorial on Hidden Markov Models and selected applications in speech recognition"Proceedings of the IEEE 77 (2): 257–286, (February 1989).