Automatic Speech Recognition: A Review

Preeti Saini[#1], Parneet Kaur[*2]

[#1,2]*CSE Department, Kurukshetra University*
ACE, Haryana, India

Abstract— **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. The design of speech recognition system require careful attentions to the challenges or issue such as various types of speech classes, speech representation, feature extraction techniques, database and performance evaluation. This paper presents a study of basic approaches to speech recognition and their results shows better accuracy. This paper also presents what research has been done around for dealing with the problem of ASR.**

***Keywords*- Automatic speech recognition, hidden markov model, acoustic model, MFCC.**

I. INTRODUCTION

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 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). 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.

1.1 Mathematical representation of ASR

In statistical based ASR systems an utterance is represented by some sequence of acoustic feature observations O, derived from the sequence of words W. The recognition system needs to find the most likely word sequence, and given the observed acoustic signal is formulated by:

$W = \text{argmax}_W P(W|O)$ (i)

In "equation (i)" [4], the argument P(W|O) i.e. the word sequence W is found which shows maximum probability, given the observation vector O. Using Baye's rule it can be written as:

$W = \text{argmax}_W P(W|O). P(W)/P(O)$ (ii)

In "equation (ii)" [4], P(O) is the probability of observation sequence and is not considered as it is a constant w.r.t. W. Hence, $W = \text{argmax}_W P(W|O) P(W)$ (iii)

In "equation (iii)" [4], P(W) is determined by a language model like grammar based model and P(O|W) is the observation likelihood and is evaluated based on an acoustic model. Among the various models, Hidden Markov Model (HMM) is so far the most widely used technique due to its efficient algorithm for training and recognition.

II. TYPOLOGY OF SPEECH RECOGNITION SYSTEMS

- Speaker Dependent: - systems that require a user to train the system according to his or her voice.
- Speaker Independent: - systems that do not require a user to train the system i.e. they are developed to operate for any speaker.
- Isolated word recognizers: - accept one word at a time. These recognition systems allow us to speak naturally continuous.
- Connected word systems [1] allow speaker to speak slowly and distinctly each word with a short pause i.e. planned speech.
- Spontaneous recognition systems [1] allow us to speak spontaneously.

III. OVERVIEW OF SPEECH RECOGNITION

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified. A speech recognition system consists of five blocks: - Feature extraction, Acoustic modeling, Pronunciation modeling, Decoder. The process of speech recognition begins with a speaker creating an utterance which consists of the sound-waves. These sound waves are then captured by a microphone and converted into electrical signals. These electrical signals are then converted into digital form to make them understandable by the speech-system. Speech signal is then converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). Finally recognition component finds the best match in the knowledge base, for the incoming feature vectors. Sometimes, however the information conveyed by these feature vectors may be correlated and less discriminative which may slow down the further processing. Feature extraction methods like Mel frequency cepstral coefficient (MFCC) provides some way to get uncorrelated vectors by means of discrete cosine transforms (DCT).

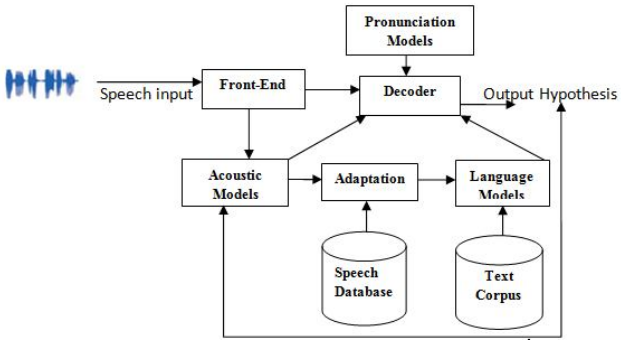

Fig. 1 Outline of speech recognition system [3]

Figure 1 shows the block diagram representing speech recognition process.

3.1 Feature Extraction

First of all, recording of various speech samples of each word of the vocabulary is done by different speakers. After the speech samples are collected, they are converted from analog to digital form by sampling at a frequency of 16 kHz. Sampling means recording the speech signals at a regular interval. The collected data is now quantized if required to eliminate noise in speech samples. The collected speech samples are then passed through the feature extraction, feature training & feature testing stages. Feature extraction transforms the incoming sound into an internal representation such that it is possible to reconstruct the original signal from it. There are various techniques to extract features like MFCC, PLP, RAST, LPCC, but mostly used is MFCC.

3.1.1 Mel Frequency Cepstral Coefficients

MFCCs [2] are used because it is designed using the knowledge of human auditory system and is used in every state of speech recognition system or art speech. MFCC is a standard method for feature extraction in speech recognition tasks. MFCC include certain steps applied on an input speech signal. These computational steps of MFCC include: - Framing, Windowing, DFTH, Mel filter bank algorithm, computing the inverse of DFT.

3.2 Decoding

It is the most important step in the speech recognition process. Decoding [3] is performed for finding the best match for the incoming feature vectors using the knowledge base. A decoder performs the actual decision about recognition of a speech utterance by combining and optimizing the information conveyed by the acoustic and language models.

3.2.1 Acoustic Modelling

There are two kinds of acoustic models [3] i.e. word model and phoneme model. An acoustic model is implemented using different approaches such as HMM, ANNs, dynamic Bayesian networks (DBN), support vector machines (SVM). HMM is used in some form or the other in every state of the art speech and speech recognition system.

3.2.1.1 Hidden Markov Model

HMMs [3] are used for acoustic modelling. There are two stochastic processes which are inter-related which are same as Markov Chain except that the output symbol and well as the transitions are probabilistic. Each HMM state may have a set of output symbols known as output probabilities and having a finite number of states $Q= \{q_1, q_2, ..., q_n\}$. One process is related to the transitions among the states which are controlled by a set of probabilities called transition probabilities to model the temporal variability of speech. Other process is concerned with the state output observations $O= \{o_1, o_2, ..., o_n\}$ regulated by Gaussian mixture distributions $b_j(o_t)$ where $1<=j<=N$, to simulate the spectral variability of speech. Any and every sequence of states that has the same length as the symbol sequence is possible, each with a different probability. The sequence of states is sad to be "hidden" from the observer who only sees the output symbol sequence, and that is why this model is known as Hidden Markov Model. The Markov nature of the HMM i.e. the probability of being in a state is dependent only on the previous sate, admits use of the Viterbi algorithm to generate the given sequence symbols, without having to search all possible sequences. At each distinct instance of time, one process is assumed to be in some state and an observation is produced by the other process representing the current state. The underlying Markov chain then changes states according to its transition from state i to state j denoted as:

$a_{ij}= P[Q_{t+1} = j| Q_t = i]$.

3.2.2 Language Modelling

Language models [3] are used to guide the search correct word sequence by predicting the likelihood of nth word using (n-1) preceding words. Language models can be classified into:

- Uniform model: each word has equal probability of occurrence.
- Stochastic model: probability of occurrence of a word depends on the word preceding it.
- Finite state languages: languages use a finite3 state network to define the allowed word sequences.
- Context free grammar: It can be used to encode which kind of sentences is allowed.

3.3 Pronunciation Modelling

In pronunciation modelling [3], during recognition, the sequence of symbols generated by acoustic model HMM is compared with the set of words present in dictionary to produce sequence of words that is the system's final output contains information about which words are known to the system and how these words are pronounced i.e. what is their phonetic representation. Decoder is then used for recognizing

words by combining and optimizing the information of acoustic & language models.

IV. APPROACHES TO SPEECH RECOGNITION

There are three types of approaches to ASR. They are:

- Acoustic phonetic approach
- Pattern Recognition approach
- Artificial intelligence approach.

4.1 Acoustic Phonetic Approach

Acoustic phonetic approach [9] is also known as rule-based approach. This approach uses knowledge of phonetics & linguistics to guide search process. There are usually some rules which are defined expressing everything or anything that might help to decode based in "blackboard" architecture i.e. at each decision point it lays out the possibilities and apply rules to determine which sequences are permitted. It has poor performance due to difficulty to express rules, to improve the system. This approach identifies individual phonemes, words, sentence structure and/or meaning.

4.2 Pattern Recognition Approach

This method has two steps i.e. training of speech patterns and recognition of pattern by way of pattern comparison. In the parameter measurement phase (filter bank, LFC, DFT), a sequence of measurements is made on the input signal to define the "test pattern". The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern & reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase (dynamic time warping).

4.2.1 Template Matching Approach

Test pattern T, and reference pattern $\{R_1,\ldots, R_v\}$ are represented by sequences of feature measurements. Pattern similarity is determined by aligning test pattern T with reference pattern R_v with distortion $D(T,R_v)$. Decision rule chooses reference pattern R* with smallest alignment distortion D(T,R*).

$R^* = \text{argmin } D(T,R_v)$

Dynamic Time Warping (DTW) is used to compute the best possible alignment $\square_v$ between T and R_v and the associated distortion $D(T, R_v)$.

4.2.2 Stochastic based approach

It can be seen as extension of template based approach, using some powerful and statistical tools and sometimes seen as anti-linguistic approach. It collects a large corpus of transcribed speech recording and train the computer to learn the correspondences. At run time, statistical processes are applied to search for all the possible solutions & pick the best one.

4.3 Artificial Intelligence Recognition Approach

This approach is a combination of the acoustic phonetic approach & the pattern recognition approach. In the AI, an expert system implemented by neural networks is used to classify sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand.

V. ADVANCES IN SPEECH RECOGNITION

In this section, a review [1] of some reported works on speech recognition has been performed. A brief introduction to different types of speech recognition approaches has also been given. Research on automatic speech recognition by machine has fascinated much attention over the past five decades. It is due to the technological curiosity about understanding the mechanisms for mechanical realization of human speech capabilities. Desire to automate simple tasks requiring human-machine interactions also motivated the researchers to work on this appealing field.

5.1 Acoustic phonetic based speech recognition (1920-1960s)

In 1920s machine recognition came into existence. The earliest attempt to devise an acoustic-phonetic based system for speech recognition was made in 1950. At Bell laboratories, Davis et al. (1952) built a system for isolated digit recognition for a single speaker. The developed system relied on measuring the spectral resonances during the vowel region of each digit. Olson and Belar (1956), at RCA laboratories tried to recognize 10 distinct syllables of a single talker embodied in 10 monosyllabic words. At MIT Lincoln laboratories, Forgie and Forgie (1959) constructed a vowel recognizer which can recognize 10 vowels embedded in a /b/-vowel-/t/ format in a speaker independent mannerThe heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

5.2 Hardware based recognizer (1960-1970s)

In 1960-1970, several Japanese laboratories had entered in the recognition arena. Suzuki and Nakata (1961) of the Radio Research Lab in Tokyo developed a hardware vowel recognizer. An elaborate filter bank spectrum analyzer was used along with the logic that connects the outputs of each channel of the spectrum analyzer to a vowel decision circuit and majority decisions logic scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita (1962) of Kyoto University, who built a hardware phoneme recognizer. During the implementation, a hardware speech segmenter was used along with a zero crossing analysis of different regions of the spoken input to provide the recognition output. Nagata et al. (1963) designed a digit recognizer at Nippon Electric Corporation (NEC) laboratories. Reddy's research program at Carnegie Mellon University. Vintsyuk (1968) proposed the use of dynamic programming methods for time aligning a pair of speech

utterances (generally known as Dynamic Time Warping (DTW).

5.3 Pattern based speech recognition (1970-1980s)

Isolated word recognition [1] was a key focus of research in the 1970s because of the fundamental studies done by Velichko and Zagoruyko (1970) in Russia, Sakoe and Chiba (1978) in Japan and Itakura (1975) in the United States. Velichko and Zagoruyko's works helped the advance use of pattern recognition ideas in speech recognition. Itakura's study showed how the idea of Linear Predictive Coding (LPC), which had already been successfully used in low bit rate speech coding, could be extended to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters. `Sambur and Rabiner (1976) described a statistical decision approach for the recognition of connected digits for speaker dependent as well as speaker independent system.

5.4 Continuous word based speech recognition (1980-1990s)

During 1980s, the main focus of research was tuned to continuous word recognition. Robust training methodology that has advantages of both averaging and clustering techniques has been presented by Rabiner and Wilpon (1980). A continuous word recognition system [1] has capability of recognizing a fluently spoken string of words based on matching a concatenated pattern of individual words. In the mid 1980's at Bell Laboratories, the theory of HMM was extended to mixture densities (Juang, 1985; Juang et al., 1986) which have ensured satisfactory recognition accuracy for speaker independent, large vocabulary speech recognition tasks.

5.5 Hybrid statistical and connectionist (HMM/ANN) based speech recognition (1990-2000s)

W. Ma with his associates Compernolle and Katholieke (Ma et al., 1990) proposed a system which combines the good short-time classification properties of time delay neural networks with the good integration and overall recognition capabilities of HMMs. A novel approach for a hybrid connectionist HMM speech recognition system based on the use of a neural network as a vector quantizer has been proposed in (Rigoll, 1994). Bourlard and Morgan (1998) in their work have described the use of ANN as statistical estimator in automatic speech recognition process.

5.6 Variational Bayesian (VB) estimation based speech recognition (2000-2010)

Pruthi et al. (2000) have developed a speaker-dependent, real-time, isolated word recognizer for Hindi. An isolated word speech recognition system for Hindi language is designed by Gupta (2006). System uses continuous HMM and consists of word based acoustic. The work in (Al-Qatab and Ainon, 2010) discusses the development and implementation of an Arabic speech system. System is developed using HTK. System can recognize both continuous speech and isolated words.

VI. PERFORMANCE

The recognition performance evaluation of an ASR system must be measured on a corpus of data different from the training corpus. The performance of speech recognition system is usually specified in terms of accuracy and speed. Accuracy is computed by word error rate, whereas speed is measured with the real time factor. Other measures of accuracy include single word error rate and command success rate (CSR) [1]. Word error rate (WER) [1] is a common metric of the performance of a speech recognition or machine translation system. WER can then be computed as: - WER = S+D+1/ N where S is the no. of substitutions, D is the no. of deletions, I is the no. of insertions, N is the no. of words in the reference. When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) [1] is used: - WRR = 1-WER = N-S-D-1/N = H-I/N where H is N-(S+D) the no. of correctly recognized words.

VII. CONCLUSIONS

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. 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] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, Handbook of Speech Processing, Springer, 2008.
[9] Wiqas Ghai and Navdeep Singh,"Literature Review on Automatic Speech Recognition", International Journal of Computer Applications vol. 41– no.8, pp. 42-50, March 2012.
[10] R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, **v**ol. 1, no.1,pp. 38-47, 2012.
[11] Dev, A. (2009) 'Effect of retroflex sounds on the recognition of hindi voiced and unvoiced stops', Journal of AI and Soc., Springer, vol. 23, pp. 603-612.