Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Literature Survey on Automatic Speech Recognition System

**Miss Himanshu**[1]
M. Tech Student Department of Computer Science & Engineering ,Modern Institute of Engineering & Technology, Haryana, Kurukshetra University, India Haryana, India

**Sarbjit Kaur** [2]
Assistant Professor, Department of Computer Science & Engineering ,Modern Institute of Engineering & Technology, India

**Vikas Chaudhary**[3]
Assistant Professor, Department of Computer Science & Engineering, Shree Ram Mulkh Institute of Engineering & Technology, Haryana, India

***Abstract— In today's world, Speech Recognition is very important and popular. Speech recognition is the process of converting spoken words into text. One of the problems faced in speech recognition is that the spoken word can be vastly altered by accents, dialects and mannerisms. In case of speech recognition the research followers are mainly using three different approaches namely Acoustic phonetic approach, Pattern recognition approach and Artificial intelligence approach. The objective of this review paper is to summarize and compare some of the well known methods used in various stages of speech recognition system and identify research topics.***

***Keywords— Automatic Speech Recognition (ASR), Hidden Markov Model (HMM).***

## I. INTRODUCTION

Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine.

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Many speech recognition applications, such as voice dialing, simple data entry and speech-to-text are in existence today.while speech recognition sets its goals at recognizing the spoken words in speech, the aim of speaker recognition is to identity the speaker by extraction, characterization and recognition of the information contained in the speech signal.

## II. BASIC MODEL OF SPEECH RECOGNITION

Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine speech is the most natural form of human communication

The recognition process is shown below (fig .1).

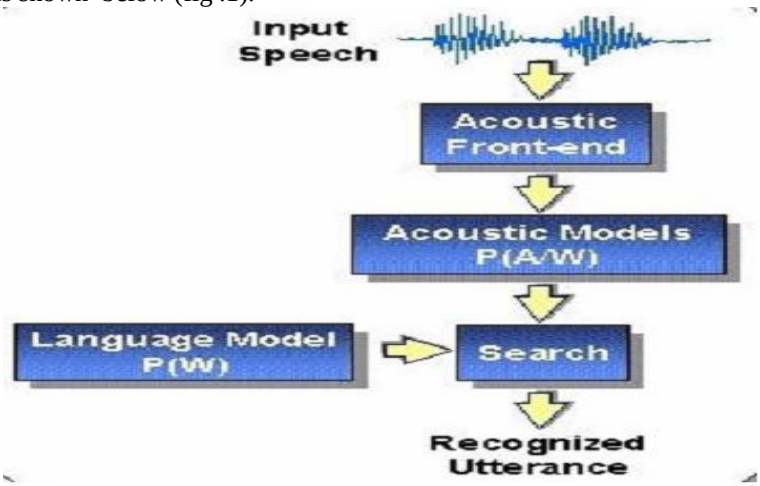


*Fig.1 Basic model of speech recognition*

 

For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years[10].

Based on major advances in statistical modeling of speech, automatic speech recognition systems today many applications in tasks that require human machine interface, like automatic call processing in telephone networks, and query based information systems that provides updated travel information, stock price quotations, weather reports, data entry, voice dictation, access to information: travel, banking, commands, avoinics, automobile portal, speech transcription, handicapped people (blind people) supermarket, railway reservations etc. Speech recognition Technology was increasingly used within telephone networks to automate as well as to enhance the operator services. fig.1 shows a representation of speech recognition system which contain front end unit,model unit, language model unit, and search unit.

## III. RELEVANT ISSUES OF ASR DESIGN

Main issues on which recognition accuracy depends have been presented in the table 1.

Table 1: Relevant issues of ASR design.

| | |
|---|---|
| Environment | Type of noise; signal/noise ratio;working conditions |
| Transducer | Microphone;telephone |
| Speakers | Speaker dependence/independence Sex,Age;physical and physical state |
| Speech Styles | Voice Tone(quiet,normal,shouted);<br>Production(isolated words or continuous speech read or spontaneous speech)<br>Speech(Slow,Normal,Fast) |
| Vocabulary | Characteristics of available training data;<br>Specific or generic vocabulary; |

## IV. APPROACHES TO SPEECH RECOGNITION

Basically there exist three approaches to speech recognition. They are

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

Generally speech recognition process deal with speech variability and account for learning the relationship between specific utterance and the corresponding word or word [1].There has been steady progress in the field of speech recognition over the recent year with two trends [2].First is academic approach and second is the pragmatic, include the technology, which provides the simple low-level interaction with machine, replacing with buttons and switches. A second approach is useful now, while the former mainly make promises for the future. There are three approaches to speech recognition [3] [4] [5]. A. Acoustic-phonetic approach [6][7][8][9] B. Artificial Intelligence approach C. Pattern recognition approach .

***1) Acoustic-Phonetic Approach*** *:* In this speech recognition algorithm, the system tries to decode the speech signal in a sequential manner based on the observed acoustic features of the speech waveform and the known relations between acoustic features and phonetic symbols. Figure 1 shows a block diagram of the acoustic- phonetic approach to speech recognition. The first step in the process is the parameter measurement process, which provides an appropriate spectral representation of the speech signal. The next step in the processing is the feature detection stage where the spectral measurements are converted to a set of features that describe the acoustic properties of the various phonetic units. Finally, the recogniser tries to determine the best matching word or sequence of words.

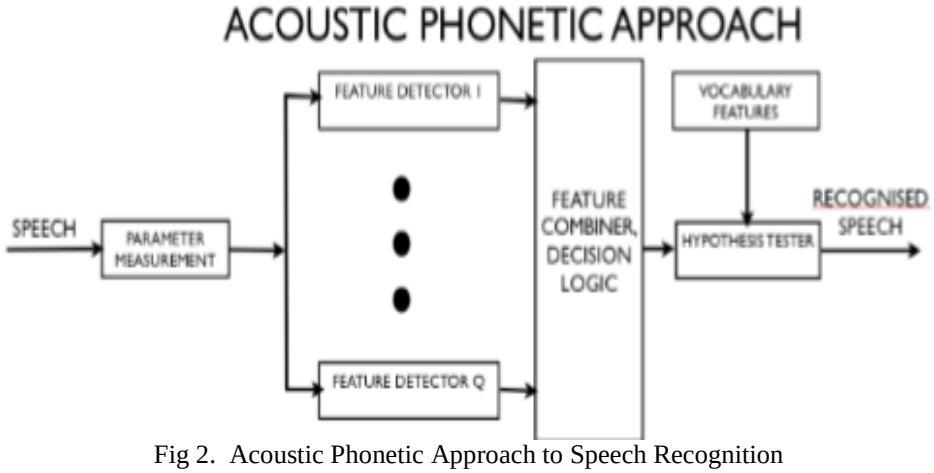


Fig 2. Acoustic Phonetic Approach to Speech Recognition

 

***2) Pattern Recognition Approach:***

In this approach, the speech patterns are used directly without explicit feature determination and segmentation. The method has two steps-namely, training of speech patterns, and recognition of patterns by way of pattern comparison. Figure 2 shows a block diagram of the pattern-recognition approach. In the parameter measurement phase, 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 and reference pattern is computed. Finally the decision rule decides which reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase.

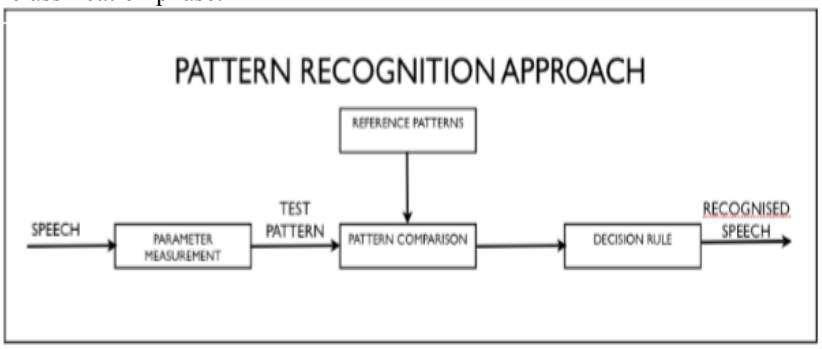


Fig 3. Pattern Recognition Approach to Speech Recognition

***a) Template Based Approach*:**

Template based approach [15] to speech recognition have provided a family of techniques that have advanced the field . 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.Each word must have its own full reference template; One key idea in template method is to derive a typical sequences of speech frames for a pattern(a word) via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another key idea is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker.

***b) Stochastic Approach*:**

Stochastic modeling [11] entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition.

The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions.A template based model is simply a continuous density HMM, with identity covariance matrices and a slope constrained topology. Although templates can be trained on fewer instances,the lack the probabilistic formulation of full HMMs and typically underperforms HMMs. Compared to knowledge based approaches; HMMs [12] [13] [14] [15] [16] [17] enable easy integration of knowledge sources into a compiled architecture. A negative side effect of this is that HMMs do not provide much insight on the recognition process. As a result,it is often difficult to analyze the errors of an HMM system in an attempt to improve its performance. Nevertheless,prudent incorporation of knowledge has significantly improved HMM based systems.

***3) Artificial Intelligence Approach (Knowledge Based Approach):***

The Artificial Intelligence approach [11] is a hybrid of the acoustic phonetic approach and pattern recognition approach. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand.

## V. LITERATURE REVIEW

***Related work***

Most conventional speaker recognition systems use Gaussian mixture models (GMMs) to capture frame-level characteristics of a person's voice, where the speech frames are assumed to be independent of one another. Because of this independence assumption, GMMs often fail to capture certain types of speaker-specific information that evolve over time scales of more than one frame. For example, since words usually span many frames, GMMs [46] tend to be poorly suited for modeling differences in word usage (idiolect) between speakers. In recent times, automatic speaker recognition research has expanded from utilizing only the acoustic content of speech to examining the use of higher levels of speech information, commonly referred to as "high-level features." A promising direction in high-level feature research has been the use of n-gram based models to capture speaker specific patterns in the phonetic and lexical content of speech.

 

In,Doddington performed an important initial study about using the lexical content of speech for speaker recognition, and introduced an n-gram based technique for modeling a speaker's idiolect. This direction in research was continued by Andrews, Kohler, and Campbell among others, who used similar n gram, based models to capture speaker pronunciation idiosyncrasies through analysis of automatically recognized phonetic events. This line of research is generally referred to as "Phonetic Speaker Recognition." The research of Andrews et al. and Doddington showed word and phone n-gram based models to be quite promising for speaker recognition. There have been myriad attempts, especially since the Johns Hopkins 2002 Workshop to harness the power of all kinds of high-level features.

The current "state-of-the-art" in phonetic speaker recognition uses relative frequencies of phone n-grams as features for training speaker models and for scoring test-target pairs [48]. Typically, these relative frequencies are computed from a simple 1-best phone decoding of the input speech. This line of phonetic speaker recognition research work has been extended in various ways by introducing different modeling strategies and different methods of utilizing the source information such as described inNavratil proposed a method involving binary-tree-structured statistical models for extending the phonetic context beyond that of standard n-gram (particularly bigrams) by exploiting statistical dependencies within a longer sequence window without exponentially increasing the model complexity, as is the case with n-grams. The described approach confirms the relevance of long phonetic context in phonetic speaker recognition and represents an intermediate stage between short phone context and word-level modeling without the need for any lexical knowledge. Binary-tree models represent a step towards flexible context structuring and extension in phonetic speaker recognition, consistently outperforming standard smoothed bigrams as well as trigrams.

Klusacek, proposed a conditional pronunciation modeling method. It uses time-aligned streams of phones and phonemes to model a speaker's specific pronunciation. The system uses phonemes drawn from a lexicon of pronunciations of words recognized by an automatic speech recognition system to generate the phoneme stream and an open-loop phone recognizer to generate a phone stream. The phoneme and phone streams are aligned at the frame level and conditional probabilities of a phone, given a phoneme, are estimated using co-occurrence counts. A likelihood detector is then applied to these probabilities for the speaker detection task. This approach achieves a relatively high accuracy in comparison with other phonetic methods in the Super SID project at the Johns Hopkins 2002 Workshop.

Campbell performed phonetic speaker recognition with support vector machines (SVM) [51]. By computing frequencies of phones in conversations, speaker characterization was performed. A new kernel was introduced based on the standard method of log likelihood ratio scoring. The resulting SVM method reduced error rates dramatically over standard techniques. Hatch compared 1-best phone decoding vs. lattice phone decoding for the purposes of performing phonetic speaker recognition. The results indicate that lattice decoding provide a much richer sampling of phonetic patterns than 1-best decoding. All the state-of-the-art speaker recognition approaches try to model phonetic dependencies along the time scale, or in time dimension. In the following sections, we will present our contributions in the speaker recognition research. We introduce a speaker recognition approach that aims at modeling the statistical pronunciation patterns based on the information from two "orthogonal" dimensions: time dimension and cross-stream dimension. It will be shown that comparable or better results are achieved by the proposed approach.

## VI. CONCLUSION & FUTURE WORK

There has been much progress in the field of automatic speech recognition since it's humble beginnings in the 1950s. Current speech recognition systems are generally based on hidden Markov models as these models have lead to the best results in speech recognition systems thus far. Although HMMs have been very successful, there are a few limitations of the models. We are a long way from achieving perfect speech recognition and there is much research still to be done in the field of automatic speech recognition.

## REFERENCES

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

[2] Abdul Kadir K, (2010), "Recognition of Human Speech using q-Bernstein Polynominals," International Journal of Computer Application, Vol.2 – No.5, pp.22-28.

[3] Reddy, R. (1976), "Speech Recognition by Machine: A Review," in proceedings of IEEE transaction, Vol. 64, No. 4, pp. 501-531.

[4] Gaikwad, Gawali and Yannawar (2010),"A Review on Speech Recognition Technique," International Journal of Computer Application, Vol.10, No.3, pp.16-24.

[5] Rohini B Shinde and V P Pawar (2012), "A Review on Acoustic Phonetic Approach for Marathi Speech," Recognition. International Journal of Computer Applications 59(2): 40-44.

[6] Friesen, L. M., Shannon, R. V., Bas¸kent, D., and Wang, X. (2001), "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants,"J. Acoust. Soc. Am. 110(2), 1150–1163.

[7] A. Mohamed, G. Dahl, and G. Hinton (2012), "Acoustic modeling using deep belief networks," IEEE Transactions on Audio, Speech, and Language Processing,, vol. 20, no. 1, pp. 14–22.

[8] L. Deng (2003), "Switching dynamic system models for speech articulation and acoustics," in Mathematical Foundations of Speech and Language Processing, pp. 115–134. Springer-Verlag, New York

[9] Deng and D. Yu (2007), "Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition," in Proc.ICASSP, pp. 445–448.

 

[10] Dat Tat Tran, Fuzzy Approaches to Speech and Speaker Recognition , A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

[11] 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".

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, andT. Kitamura (1999), "Simultaneous modeling of spectrum, pitch,and duration in HMM-based speech synthesis," Proc. Of EUROSPEECH, pp.2347–2350.

[13] H. Zen and N. Braunschweile (2009), "Context-dependent additive log F0 model for HMM-based speech synthesis,"Proc. Interspeech, pp. 2091–2094

[14] K. Oura, H. Zen, Y. Nankaku, A, Lee, and K. Tokuda (2010),"A Covariance-Tying Technique for HMM-based Speech Synthesis," IEICE, vol. E93–D, no.3, pp.595–601.

[15] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi (1999),"Hidden Markov Models Based on Multi-SpaceProbability Distribution for Pitch Pattern Modeling,"Proc. ICASSP, pp. 229–232.

[16] G Heigold, R Schlter, and H Ney (2007), "On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields,"in Interspeech, 2007, pp. 1721–1724

[17] J. Kaiser, B. Horvat, and Z. Kacic (2000), "A novel loss function for the overall risk criterion based discriminative training of HMM models,"in Proc. ICSL.

 