International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-3 Issue-6, November 2013

A Quantitative Study of the Automatic Speech Recognition Technique

Anchal Kaytal, Amanpreet Kaur

Abstract: ***In the last two decades, few researchers have worked for the development of Automatic Speech Recognition Systems for most of these languages in such a way that development of this technology can reach at par with the research work which has been done and is being done for the different languages in the rest of the world. Punjabi is the 10th most widely spoken language in the world for which no considerable work has been done in this area of automatic speech recognition. Being a member of Indo-Aryan languages family and a language rich in literature, Punjabi language deserves attention in this highly growing field of Automatic speech recognition. The Speech is most prominent & primary mode of Communication among of human being. Today, speech technologies are commercially available for an unlimited but interesting range of tasks. These technologies enable machines to respond correctly and reliably to human voices, and provide useful and valuable services.***

Keywords: ASR , Punjabi Speech Recognition , Recognition Techniques

I. INTRODUCTION

Spoken language is not just a means to access information, but itself information. The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech [1]. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means Algorithm implemented as a computer program. Communication among human beings is dominated by spoken language. Therefore, it is natural for people to expect speech interfaces with computers which can speak and recognize speech in native language. India has a linguistically rich area which has 18 constitutional languages, which are written in 10 different scripts [2]. Machine recognisation of speech involves generating a sequence of words best matches the given speech signal. Some of known applications include virtual reality, Multimedia searches, auto-attendants, travel Information and reservation, translators, natural language understanding and many more Applications [3].

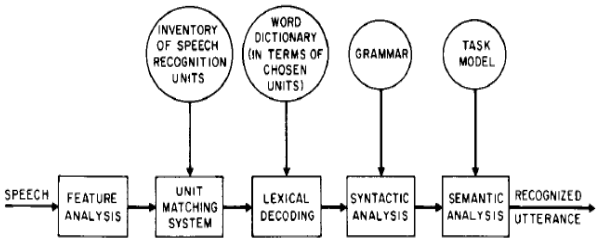

Figure 1: Block Diagram of speech recognition

Manuscript published on 30 November 2013.
*Correspondence Author(s)
Anchal Kaytal M-TECH CSE RIMT, Mandi Gobindgarh, India.
Amanpreet Kaur, Asst Proff RIMT, Mandi Gobindgarh, India.

1.1 Relevant issues of ASR design

Main issues on which recognition accuracy depends have been presented in the table [17]

Environment	Type of noise; signal/noise ratio; working conditions
Transducer	Microphone; telephone
Channel	Band amplitude; distortion; echo
Speakers	Speaker dependence/independence Sex, Age; physical and psychical state
Speech styles	Voice tone(quiet, normal, shouted); Production(isolated words or continuous speech read or spontaneous speech) Speed(slow, normal, fast)
Vocabulary	Characteristics of available training data; specific or generic vocabulary;

Table 1: Relevant issues of ASR design

II. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is the process of mapping an acoustic waveform into a text/the set of words which should be equivalent to the information being conveyed by the spoken words. This challenging field of research has almost made it possible to provide a PC which can perform as a stenographer, teach the students in their mother language and read the newspaper of reader's choice. The advent and development of ASR in the last 6 decades has resolved the issues of the requirements of certain level of literacy, typing skill, some level of proficiency in English, reading the monitor by blind or partially blind people, use of computer by physically challenged people and good hand-eye co-ordination for using mouse. In addition to this support, ASR application areas are increasing in number day by day. Research in Automatic Speech Recognition has various open issues such as Small/ Medium/ Large vocabulary, Isolated/ Connected/Continuous speech, Speaker Dependent/ Independent and Environmental robustness [9].

A. Modules of ASR

Automatic speech recognition system is comprised of modules as shown in the figure.

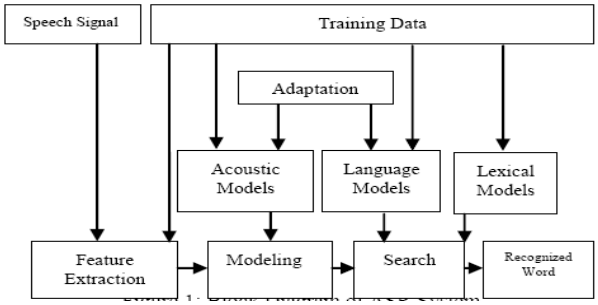

Figure 2: Block Diagram of ASR System

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

1. Speech Signal acquisition: At this stage, Analog speech signal is acquired through a high quality, noiseless, unidirectional microphone in .wav format and converted to digital speech signal.
2. Feature Extraction: Feature extraction is a very important phase of ASR development during which a parsimonious sequence of feature vectors is computed so as to provide a compact representation of the given input signal. Speech analysis of the speech signal acts as first stage of Feature extraction process where raw features describing the envelope of power spectrum are generated. An extended feature vector composed of static and dynamic features is compiled in the second stage. Finally this feature vector is transformed into more compact and robust vector. Feature extraction, using MFCC, is the famous technique used for feature extraction.

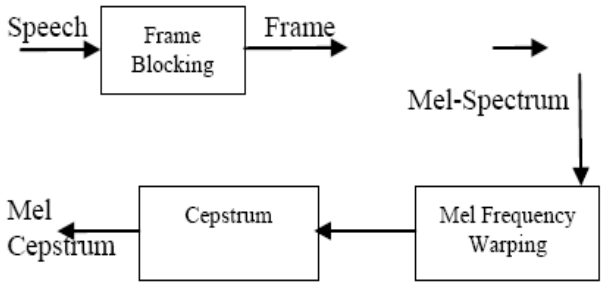

Figure 3: Block Diagram of Feature Extraction

3. Acoustic Modeling: Acoustic models are developed to link the observed features of the speech signals with the expected phonetics of the hypothesis word/sentence. For generating mapping between the basic speech units such as phones, tri-phones & syllables, a rigorous training is carried. During training, a pattern representative for the features of a class using one or more patterns corresponding to speech sounds of the same class.
4. Language & Lexical Modeling: Word ambiguity is an aspect which has to be handled carefully and acoustic model alone can't handle it. For continuous speech, word boundaries are major issue. Language model is used to resolve both these issues. Generally ASR systems use the stochastic language models. These probabilities are to be trained from a corpus. Language accepts the various competitive hypotheses of words from the acoustic models and thereby generates a probability for each sequence of words. Lexical model provides the pronunciation of the words in the specified language and contains the mapping between words and phones. Generally a canonical pronunciation available in ordinary dictionaries is used. To handle the issue of variability, multiple pronunciation variants for each word are covered in the lexicon but with care. A G2P system- Grapheme to Phoneme system is applied to better the performance the ASR system b predicting the pronunciation of words which are not found in the training data.
5. Model Adaptation: The purpose of performing adaptation is to minimize the system's performance dependence on speaker's voice, microphones, transmission channel and acoustic environment so that the generalization capability of the system can be enhanced. Language model adaptation is focused at how to select the model for specific domain. Adaptation process identifies the nature of domain and, thereby, selects the specified model.

6. Recognition: Recognition is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed. Two approaches are being used to match the patterns: First one is the Dynamic Time Warping based on the distance between the acoustic units and that of recognition. Second one is HMM based on the maximization of the occurrence probability between training and recognition units. To train the HMM and thereby to achieve good performance, a large, phonetically rich and balanced database is needed.

B. Data Preparation

1. Building Text Corpus: Text corpus means optimal set of textual words/sentences which will be recorded by the native speakers of a particular language. According to the domain specified for the ASR, the corresponding text is collected. Different context in which that text can be used, are also taken care of. Building a text corpus involves three steps: Text corpus collection, Grapheme to Phoneme Conversion, Optimal Text Selection.
2. Building Speech Corpus: With the help of Text Corpus, the recordings of selected words/sentences are done with the help of high quality microphones. During the development of speech corpus, information, which is generally noted down, is Personal Profile of Speakers, Technical details of microphone, Date and Time of Recording, Environmental conditions of recording. During the recording session, the parameters of the wave file to be set are: Sampling rate, Bit rate, Channel. Building Speech Corpus involves three steps: Selecting a speaker, Data Statistics, Transcription Correction.
3. Transcription File: A transcript file is required to represent what the speakers are saying in the audio file. It contains the dialogue of the speaker noted exactly in the same precise way as it has been recorded. There are two transcription files: one is meant for training the system and second one is meant for testing the system.
4. Pronunciation Dictionary: It is a language dictionary which contains mapping of each word to a sequence of sound units. The purpose of this file is to derive the sequence of sound units associated with each signal. The important point, which is to be taken care of while preparing this dictionary, the sound units must be contained in this dictionary, must be in ASCII.
5. Language Model: Language model is meant for providing the behaviour of the language. The language model describes the likelihood or the probability taken when a sequence or collection of words is seen. A language model is a probability distribution over the entire sentences/texts. The purpose of creating a language model is to narrow down the search space, constrain search and thereby to significantly improve recognition accuracy. Language model becomes very important when Continuous speech is considered. Speech recognizers seek the word sequence Ws which is most likely to be produced from acoustic evidence A as per the following formula:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

P (Ws | A) = maxw P (W|A) = maxw P (A|W) P (W)/P (A)
Where P (A): Probability of acoustic evidence. Language Model assigns a probability estimate P (W) to word sequences W= {W1, W2,....., Wn}. These probabilities can be trained from a corpus. Perplexity is a parameter to evaluate language model. Suppose sentences in a test sample contains 2000 words and can be coded using 10000 bits then the perplexity of language model = 2(10000/2000)=32 per word. Language model with low perplexity helps LM to perform well for the speech recognition system thereby compressing the test sample.
6. Filler Dictionary: It refers to a dictionary which contains the mapping of non-speech sounds to non-speech sound units.
e.g. <sil> SIL

C. Performance Parameters

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system which are described below:

1. Accuracy Parameters

Word Error Rate (WER): The WER is calculated by comparing the test set to the computer-generated document and then counting the number of substitutions (S), deletions (D), and insertions (I) and dividing by the total number of words in the test set

2. Speed Parameter

Real Time Factor is parameter to evaluate speed of automatic speech recognition. Formula: P RTF = ---------- ----- I where P: Time taken to process an input Duration of input I e. g. RTF= 3 when it takes 6 hours of computation time to process a recording of duration 2 hours. RTF ≤ 1 implies real time processing.

D. Performance Degradation

Automatic speech recognition suffers degradation in recognition performance due to following inevitable factors:

i. Prosodic and phonetic context
ii. Speaking behavior
iii. Accent & Dialect
iv. Transducer variability and distortions
v. Adverse speaking conditions
vi. Pronunciation
vii. Transmission channel variability and distortions
viii. Noisy acoustic environment
ix. Vocabulary Size and domain

III. AUTOMATIC SPEECH RECOGNITION CLASSIFICATION

The following tree structure emphasizes the speech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in figure [17]

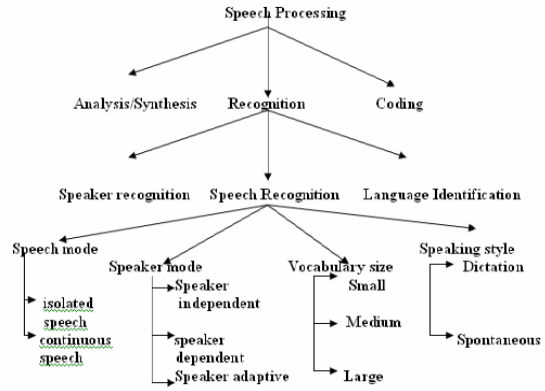

Figure 3: Speech Processing Classification

IV. UNITS OF SPEECH FOR PUNJABI LANGUAGE

The syllable comprises vowel and consonants. The presence of vowel is must in a syllable. The vowel is the nucleus, presence of consonant is optional. Vowel (V) is always the nucleus part and the left part is onset and the right part is coda that is consonant.
The seven types of syllables recognized in Punjabi language are as follows:

V, VC, CV, VCC, CVC, CCVC, CVCC

There are thirty eight consonants, ten non-nasal vowels and same number of nasal vowels in Punjabi language. Consonants can appear with vowels only. Following are the list of consonants in Punjabi language:

ਸ ਹ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ
ਟ ਠ ਡ ਢ ਣ ਤ ਥ ਦ ਧ ਨ ਪ ਫ ਬ
ਭ ਮ ਯ ਰ ਲ ਵ ੜ ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼

List of Non-Nasal Vowels:

ਈ ਇ ਏ ਐ ਅ ਆ ਔ ਊ ਉ ਓ

The number of nasal vowels is same as non-nasal ones and is represented by Bindi or Tippi over the Non-Nasal Vowels.

V	Vowel	ਅ	ਅ
VC	Vowel+ Consonant	ਇ+ਹ	ਇਹ
CV	Consonant +Vowel	ਜ+ਆ	ਜਾ
VCC	Vowel+ Consonant+ Consonant	ਪ+ ਗ+ ਗ	ਪੱਗ
CVC	Consonant +Vowel+ Consonant	ਬ+ਆ+ਤ	ਬਾਤ
CCVC	Consonant + Consonant +Vowel+ Consonant	ਹ+ਨ+ਏ+ਰ	ਹਨੇਰ
CVCC	Consonant +Vowel+ Consonant+	ਪ+ਊ+ਰ+ਬ	ਪੂਰਬ

Table 2: List of Syllables in Punjabi

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. REPRESENTATION OF SPEECH

Traditionally, the information in speech signal is represented in terms of features derived from Fourier analysis: Fourier transformation, Fast Fourier transformation, discrete fourier transformation, or Wavelets. The key difference between fourier transform and wavelets transform is that wavelet transform is a multi-resolution transform as it allows a form of time-frequency analysis. When using the fourier transform the result is a very precise analysis of its frequency contained in the signal, but no information about when certain features occurred and about the scale characteristics of the signal. Scale is similar to frequency. It is a measure of the amount of detail in the signal. Small scale means coarse details and large scale means fine details. The information in speech signals can be represented in terms of features derived from short-time Fourier analysis. The information in the short-time FT phase function can be extracted by processing the negative derivative of the FT phase, i.e., the group delay function [7].

$$H(\omega) = H1(\omega) \cdot H2(\omega) \qquad (1)$$

Group delay function $\tau h(\omega)$ can be represented as $\tau h(\omega) = -\partial(\arg(H(\omega))/\partial\omega = \tau h1(\omega) + \tau h2(\omega)$ (2)
The equation (1) shows the multiplicative property of magnitude spectra where as equation (2) is in group delay domain it becomes an addition. The group delay spectrum has been declared better due to its additive property over magnitude spectra. It was observed that in case of the magnitude spectra the peaks are clearly visible, but the peaks are not resolved in a system where the two poles are combined together [6]. The research shows the disadvantage of multiplicative property of magnitude spectra. In case of group delay spectra the peaks and valleys are better resolved when the signal is in minimum phase [6].

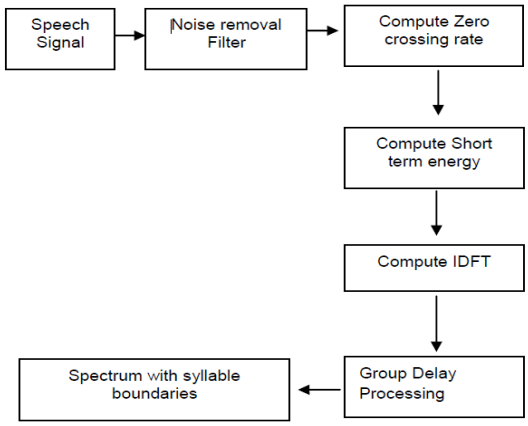

Figure 2: Steps involved in finding syllable boundaries

VI. CONCLUSION

In this review, we have discussed the fundamentals of speech recognition and its recent progress is investigated. Speech recognition has been in development for more than 50 years, and has been entertained as an alternative access method for individuals with disabilities for almost as long. Punjabi language deserves attention in this highly growing field of Automatic speech recognition. In this paper, the efforts made by various researchers to develop automatic speech recognition systems for most of the Indo-Aryan languages, have been analysed.

REFRENCES

1. Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.
2. M. Chandrasekar, M. Ponnavaikko, "Tamil speech recognition: a complete model", Electronic Journal «Technical Acoustics» 2008, 20.
3. W. M. Campbell, D. E. Sturim W. Shen D. A. Reynolds and J. Navratily, "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition", MIT Lincoln Laboratory IBM 2006.
4. Bhupinder Singh, Parminder Singh, "Voice Based user Machine Interface for Punjabi using Hidden Markov Model," JCST Vol. 2, Issue 3, September 2011 I S S N : 2 2 2 9 - 4 3 3 3 (P r i n t) | I S S N : 0 9 7 6 - 8 4 9 1.
5. N. Mikael, E. Marcus, "Speech Recognition using Hidden Markov Model, Performance evaluation in noisy environment", Degree of master of science in Electrical Engineering, Department of telecommunications and engineering, Blekinge Institute of Technology, March 2002.
6. T. Nagarajan and H. A. Murthy, "Subband-Based Group Delay Segmentation of Spontaneous Speech into Syllable-Like Units," in Eurasip Journal on Applied Signal Processing , Hindawi Publishing Corporation 2004:17, pp. 2614–2625.
7. A.Hema, and B.Yegnanarayan, "Group delay functions and its applications in speech technology," in Sadhana, Vol. 36, Part 5, October 2011, pp. 745–782.
8. Anupriya Sharma, Amanpreet Kaur, "A Survey on Punjabi Speech Segmentation into Syllable-Like Units Using Group Delay", Volume 3, Issue 6, June 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
9. Wiqas Ghai, Navdeep Singh, "Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi A Case Study", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.
10. Chetana Prakash, Suryakanth V. Gangashetty, "Fourier-Bessel Cepstral Coefficients for Robust Speech Recognition", 978-1-4673-2014 6/12/$31/00, 2012 IEEE.
11. Eliathamby Ambikairajah ," Emerging Features for Speaker Recognition", 1-4244-0983-7/07/$25.00 ©2007 IEEE ICICS 2007.
12. Dr. Joseph Picone, "FUNDAMETALS OF SPEECH RECOGNITION: A Short Course", Institute for Signal And Information Processing.
13. 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 ISSN (Online): 1694-0814.
14. Richard P. Lippmann, "Speech recognition by machines and humans", 0167-6393r97r$17.00 q 1997 Elsevier Science B.V. All rights reserved. II S0167-6393_ 97. 00021-6.
15. Kuo-Hau Wu, Chia-Ping Chen and Bing-Feng Yeh, "Noise-robust speech feature processing with empirical mode decomposition", EURASIP Journal on Audio, Speech, and Music Processing 2011, 2011:9.
16. Adam L. Buchsbaum, Raffaele Giancarlo, "Algorithmic Aspects in Speech Recognition: An Introduction".
17. M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)