Acoustic Speech Recognition

Mr. Kumar Shubham[1], Mr. Vishal Singh[2], Mr. Sonu Kumar[3],
Mr. Sumeet Raj[4], Mr. Sheshank Ran[5] , Mrs. Nilam.S. Patil[6]

[1,2,3,4,5]*BE Scholar, Department. of Computer Engineering, D Y Patil College of Engineering, Akurdi, Pune, MH, India*

[6]*Assistant professor, Department. of Computer Engineering, D Y Patil College of Engineering, Akurdi, Pune, MH, India*

ABSTRACT

Speech recognition is a process of recognition of human speech by computer producing string output of spoken sentence in written form. A model is learned from a set of audio recordings whose corresponding transcripts are created by taking audio recordings of speech and their text transcriptions, and using software to create statistical representations of the sound that make up every word. Speech based applications are getting enormous popularity by incorporating Natural Language Processing(NLP) techniques. Input to such applications is in natural language and output is obtained in natural language. In case of speech recognition system, research followers are mostly using three different methods namely Acoustic phonetic method, Pattern recognition approach and Artificial intelligence method. ASR today finds widespread application in tasks that require human machine interfaces. This paper presents a review on various existing techniques employed in building ASR models.

Keyword : *Automatic Speech Recognition system(ASR) , ASR classification, Speech Analysis, Feature extraction, Modelling techniques, Acoustic- Phonetic approach, Pattern Recognition methods, A.I approach, Language Modelling, ASR tools*

1. INTRODUCTION

Speech recognition is the process of mapping an acoustic waveform into a text (or the set of words) which should be equivalent to the information being conveyed by the spoken word.
Speech recognition in computer system domain is defined as the ability of computer systems to accept input in the form of spoken words and map into text format in order to proceed with further computations .
For this purpose, Natural Language Processing is a active area of research and development in Computer Science. NLP applications are machine translation and automatic speech recognition. For NLP, a basic unit of speech recognition is the intermediate form of speech recognition around which much of the recognition process is organized for human beings / machines.

Researchers have proposed many different types of inter-mediate units. Some of the possibilities include sub-phoneme units, phones with right or left context, biphones, diphones and variations, dyads or transemes, avents, triphones, demi-syllables, whole words and phrases.

ASR for Indian languages is still at its infancy where as western languages like English and Asian languages like Chinese are comparatively well matured. Hence ongoing decade shows growing interests in this fields and also huge scope for research work

Speech recognition involves different functionalities:

a) Speech Analysis
b) Feature Extraction
c) Acoustic Modelling
d) Language and lexical modeling recognition.

2. CLASSIFICATION OF SPEECH RECOGNITION SYSTEMS

Speech recognition system can be classify into different class based on type of speech , type of speaker model, type of channel and the type of vocabulary that it has the ability to recognize

A. Based on Speech Utterance

An utterance is the vocabulary of a word that represent a single meaning to the computer. Utterance can be a single word, a few words, a sentence, or even multiple sentence. The types of speech utterance are:
(1) Isolated Word: Isolated word recognizer usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accept single word, but does require a single utterance at a time. This is fine for situations where the user is required to give only one word response or command, but is very unnatural for multiple word inputs. It is comparatively simple and easiest to implement because word boundary are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The disadvantage of this type is choosing different boundaries affects the results.
2) Connected Words: Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.
3) Continuous Speech: Continuous Speech allows users to speak almost naturally, while the computer determines the content. Basically it's computer's dictation. It includes great deal of "co-articulation", where the adjacent words run together without pauses or any other apparent division between words.
4) Spontaneous Speech: This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features such as words being run together and even slight stutters. Spontaneous (unrehearsed) speech may include mispronunciations, false-starts, and non-words.

B. Based on Speaker Model

All speakers have their unique voices, due to their specific physical body and personality. Speech recognition system is broadly classified into two main categories based on speaker models namely speaker dependent and speaker independent.
1) Speaker dependent model: Speaker dependent models are designed for a specific speaker. They are generally more accurate for the particular speaker, but much less accurate for other speakers. These systems are easier to develop, cheap and more accurate, but not as flexible as independent models. 2)Speaker independent models: Speaker independent systems are designed for variety of speakers. It recognizes the speech patterns of a large group of people. This system is most difficult to develop, most expensive and offers less accuracy than speaker dependent systems. However, they are more flexible

C. . Based on Vocabulary

Types of Vocabulary: The size of vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. Some applications require a few words , others require large dictionaries. In ASR systems the types of vocabularies can be classified as follows:
i. Small vocabulary - tens of words
ii. Medium vocabulary - hundreds of words
iii. Large vocabulary - thousands of words
iv. Very-large vocabulary - tens of thousands of words

v. Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word

3. SPEECH RECOGNITION STEPS

A. Speech Analysis

Speech data contain different type of information that depicts a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The information about the behavior feature also embedded in signal and that can be used for speaker recognition. The speech analysis stage deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting. The speech analysis technique done with following three techniques:

1) Segmentation analysis: In this case, speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information. Studies made in used segmented analysis to extract vocal tract information of speaker recognition.

2) Sub segmental analysis: Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmental analysis.

3) Supra segmental analysis: In this case, speech is an-alyzed by using the frame size and shift of 100-300 ms to extract speaker information mainly due to behavioral tract and here speech is analyzed using the frame size. This technique is used mainly to analyze and characteristic due to behavior character of the speaker. These include word duration, speaker rate, accent etc.

B. Feature Extraction Techniques

a) Linear predictive analysis (LPC
b) Linear predictive cepstral coefficients (LPCC).
c) Perceptual linear predictive coefficients (PLP).
d) Mel-frequency cepstral coefficients. (MFCC).
e) Power spectral analysis (FFT).
f) Mel scale cepstral analysis (MEL).
g) Relative spectra filtering of log domain coefficients (RASTA).
h) First order derivative (DELTA).
i) Zero crossing with peak amplitude (ZCPA).

C. Language and lexical modelling

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 nth word on the basis of the n1 preceding words.

Language modelling is broadly categorized into two types:

1. Deterministic (Grammar-based) [2].
2. Stochastic (Statistical)

4. DIFFERENT APPROACHES TO AUTOMATIC SPEECH RECOGNITION

A. Acoustic-Phonetic approach [6]

This method is indeed viable and has been studied in great depth for more than 40 years [7]. This approach is based upon theory of acoustic phonetics and postulates. The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach [8] (Hemdal and Hughes 1967). Which postulates that there exist finite, distinctive phonetic units (phonemes) 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? it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine.There are three techniques that have been applied to the language identification. Problem phone recognition, Gaussian mixture modeling [9], and support vector machine [10] classification. The acoustic phonetic approach has not been widely used in most commercial applications.Pattern Recognition approach [11].

B. Pattern Recognition Approach

The pattern-matching approach (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) involves two essential steps namely, pattern training and pattern comparison. A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM [12]) and can be applied to a sound (smaller than a word), a word, or a phrase.

Different pattern recognition approaches:

i. Template based approach

ii. Dynamic time warping

iii. Knowledge based approach

iv. Statistical based approach

v. Stochastic approach

C. Artificial Inteligence Approach [13] [14].

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach used in speech recognition system . In this it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram

5. TOOLS FOR ASR

Following are the various tools used for ASR

PRAAT: It is free software with latest version 6.0 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 2.0 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

6. CONCLUSIONS

Speech based applications are getting enormous popularity as they prove to be essential for both the civilized and weakly educated. These days lot of research is being carried out and further lot of work needs to be done in the context of ASR for Indian languages. In this paper we dicussed various ASR techniques and have put forth some of the essential information and requisites for the same by surveying a small part of the mammoth work yet to be done in this field. We have briefly discussed the Speech Recognition System and various approaches used in ASR devoloped in various languages. Hidden Markov Model and Hidden Markov Model Toolkit (HTK) has been used widely.

7. REFERENCES

[1] B. A. Al-Qatab and R. N. Ainon, "Arabic speech recognition using hidden markov model toolkit (htk)," in Information Technology (ITSim), 2010 International Symposium in, vol. 2, pp. 557–562, IEEE, 2010.

[2] J. H. Martin and D. Jurafsky, "Speech and language processing," International Edition, 2000.

[3] M. Shrivastava, N. Agrawal, B. Mohapatra, S. Singh, and P. Bhat-tacharya, "Morphology based natural language processing tools for indian languages," in Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer Science, IIT, Kanpur, India, April, Citeseer, 2005.

[4] M. Bapat, H. Gune, and P. Bhattacharyya, "A paradigm-based finite state morphological analyzer for marathi," in Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pp. 26–34, 2010

[5] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 24, no. 3, pp. 201–212, 1976.

[6] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," Speech Communi-cation, vol. 35, no. 1, pp. 31–51, 2001.

[7] J. B. Allen, "From lord rayleigh to shannon: How do humans decode speech?," in International Conference on Acoustics, Speech and Signal Processing, 2002.

[8] A. Juneja and C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," in Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, vol. 2, pp. 726–730, IEEE, 2002.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," Speech and audio processing, ieee transactions on, vol. 2, no. 2, pp. 291–298, 1994.

[10] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," Signal Processing, IEEE Transactions on, vol. 52, no. 8, pp. 2348–2355, 2004.

[11] D. R. Reddy, "Approach to computer speech recognition by direct analysis of the speech wave," The Journal of the Acoustical Society of America, vol. 40, no. 5, pp. 1273–1273, 1966.

[12] S. R. Eddy, "Hidden markov models," Current opinion in structural biology, vol. 6, no. 3R. K. Moore, "Twenty things we still dont know about speech," in

Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research an Technology, 1994.

[13] L. R. Bahl, F. Jelinek, and R. L. Mercer, “A maximum likelihood ap-proach to continuous speech recognition,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 2, pp. 179–190, 1983.

[14] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 34, no. 4, pp. 744–754, 1986.

[15] S. Chu, S. Narayanan, and C. J. Kuo, “Environmental sound recognition with time–frequency audio features,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 6, pp. 1142–1158, 2009.

IJARIIE