www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 3 March 2013 Page No. 632-639

Automatic Speech Recognition - An Overview

Geetha.K[*], Dr.Chandra.E[**]

[*] Research Scholar, D.J Academy for Managerial Excellence, Coimbatore, India, geethakab@rediffmail.com
[**] Director, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore-32, India.
crcspeech@gmail.com

Abstract:- In Automatic Speech Recognition (ASR), there is consistent growth since past five decades. ASR has been developed in many spoken languages. This paper gives an overview of speech recognition system and its applications.

Keywords: ASR, Acoustic Model, Lexical Model, Language Model

1. Introduction

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone to a set of words [1]. The recognized words can be fed into the applications such as commands and control, data entry, and document preparation. They can also serve as the input the further linguistic processing in order to achieve speech understanding. Many researches are emerging in the area of Speech Recognition and Signal Processing [2].

Speech Recognition (SR) systems allow people to control a computer by speaking to it through a microphone, either entering a text or issuing commands to the computer. ASR has many applications in numerous areas such as command and control systems in which it accepts command and act according to the command [3]. E.g. “Open Google ”, “Start a new Ms Word”. Some ASRs have the ability to identify the specific user. E.g. Voice Verification/Identification. This paper describes about the issues, functionality and also the types of automatic speech recognition.

2. NOMENCLATURE – ASR

Utterance: An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

Vocabulary: Vocabularies (Dictionaries) are list of words or utterance that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn’t have to be a single word. They can be as long as sentence or two. Smaller vocabularies can have as few as one or two recognized utterances while very large vocabularies can have a hundred thousand or more. Recognition is generally more difficult when vocabularies are large or have many similar-sounding words.

Speaking Mode – Isolated word vs. Connected word: Early Systems used discrete speech, in which the user had to speak one word at a time, with a pause between words. Connected word systems recognize separate utterances to be 'run-together' with a minimal pause between them. Like isolated word speech recognition, it also requires the basic input speech utterance as a word/phrase.

Speaking Style - Continuous speech vs. Spontaneous speech: Continuous ASR, allows the user to speak in a more natural way. Spontaneous speech is much more difficult to recognize than speech read from script.

Speaker Dependence vs. Speaker Independence: Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the trained speaker, but much less accurate for non-trained speakers. They assume that the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

Text dependent vs. Text independent: In Text-dependent the utterance presented to the recognizer is known beforehand. It is based on the assumption that the speaker is cooperative. This is the one that easier to implement, and with higher accuracy. In Text-independent, the content of speech is unknown. It means no assumption about the text being spoken is made. This is the one that harder to implement, but with higher flexibility.

Enrollment: Some systems require speaker enrollment i.e. a user must provide samples of his or her speech before using them, whereas other systems are said to be speaker independent, in which no enrollment is necessary.

Language model: when speech is produced in a sequence of words, language models or artificial grammars are used to restrict the combination of words. The simplest language model can be specified as a finite-state network, where the permissible words following each word are given explicitly. More general language models approximating natural language are specified in terms of context-sensitive grammar.

Perplexity: It is defined as the geometric mean of the number of words that can legally appear next in the input. It will be a difficult task to combining the vocabulary size and the language model.

Accuracy: The ability of a recognizer can be examined by measuring its accuracy or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more. The acceptable accuracy of a system really depends on the application.

Training: Some speech recognizers have the ability to adapt to the speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy. Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

3. APPROACHES OF ASR

There are three types of approaches in speech recognition systems [4],

a) the acoustic-phonetic approach

b) the pattern recognition approach

c) The artificial intelligence approach

3.1 Acoustic Phonetic Approach

Acoustic phonetic systems use knowledge of the human body such as speech production, hearing etc. to compare speech features. For every spoken language, there exist a fixed number of distinctive phonetic units. These phonetic units are broadly characterized by a set of acoustics properties varying with respect to time in a speech signal. In this approach, acoustic properties like nasality, frication, voiced-unvoiced classification and continuous features such as formant locations, ratio of high and low frequencies can be analyzed to find the phonetic units.

3.2 Pattern Recognition Approach

Pattern training and pattern matching are the two essential steps in this approach. Fig. 1 shows the outline of the pattern recognition approach. It uses a well formulated mathematical framework to develop a consistent speech pattern representation for a set of labeled speech training samples via a formal training algorithm.

Steps involved in pattern matching are

1. Parameter measurement
2. Compare the patterns
3. Decision making. .

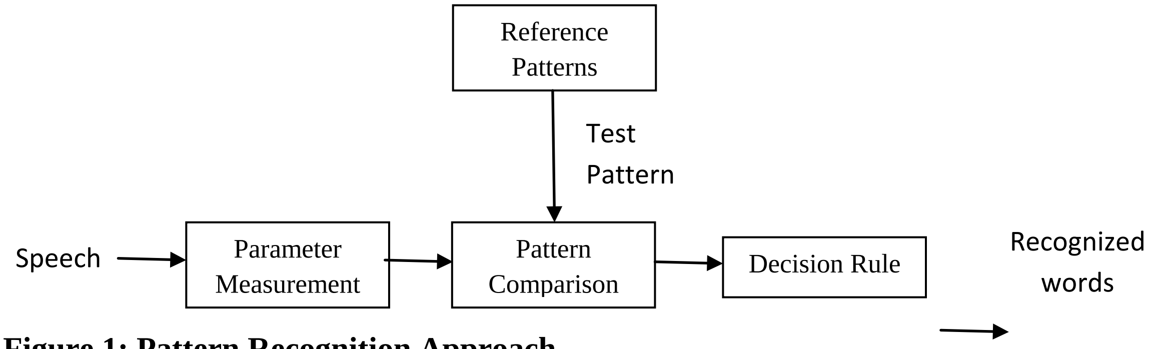

Figure 1: Pattern Recognition Approach

3.3 Artificial Intelligence Approach

Artificial intelligence approach to speech recognition is a hybrid of the acoustic-phonetic approach and the pattern recognition approach in that it exploits ideas and concepts of both methods. Both acoustic phonetic and template based approach failed at their own to explore considerable insight into human speech processing. As a result, error analysis and knowledge based system enhancement couldn't get strength. In traditional Knowledge based approach, the production rules are created heuristically from empirical linguistic knowledge or from the observations from the speech spectrogram. Knowledge helps the algorithm to perform better and also in the selection of a suitable input representation, the definition of units of speech and the design of the recognition algorithms. Most the new speech recognition systems are based on hybrid approach Hidden Markov Model/Artificial Neural Network (HMM/ANN) [5]. HMM has a great capacity to treat events in time [6], while ANN is an expert in the classification of patterns.

4. HOW ASR WORKS

There are the two phases of the general speech recognition system and they are speech training and speech testing. Figure 2 shows the major components of a typical speech recognition system. The digitized speech signal is first transformed into a set of useful measurements or features at a fixed rate typically one every 10-20 msec. These measurements are then used to search for the most likely word, making use of acoustic, lexical and language models. Throughout the process, training data are used.

Modules identified for a speech recognition system are

i. Speech Signal acquisition
ii. Feature Extraction
iii. Acoustic Modeling
iv. Language & Lexical Modeling
v. Recognition

Speech acquisition and Feature extraction modules are common to both training and testing phases of ASR.

4.1. Feature Extraction

Feature extraction is a process of extracting different features such as power, pitch, and vocal tract configuration from the speech signal. Since the performance of the ASR depends heavily on the feature extraction phase, at most care should be given in this phase. Some of the feature extraction techniques used in modern ASR are LPC, MFCC, AMFCC, RAS, DAS, ΔMFCC, Higher lag autocorrelation coefficients, PLP, MF-PLP, BFCC, RPLP. It has been found that noise robust spectral estimation is possible on the higher lag autocorrelation coefficients. Therefore, eliminating the lower lags of the noisy speech signal autocorrelation leads to removal of the main noise components.

Parameter transformation is the process of converting these features into signal parameters through process of differentiation and concatenation.

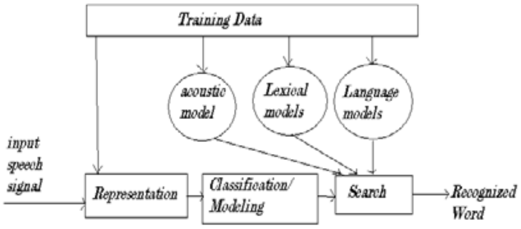

Figure 2: Framework of ASR

4.2. Acoustic Modeling

Acoustic model is the main component for an ASR and it accounts for most of the computational load and performance of the system. It is used to link the observed features of the speech signals with the expected phonetics of the hypothesis sentence. The Acoustic model is developed for detecting the spoken phoneme. Its creation involves the use of audio recordings of speech and their text scripts and then compiling them into a statistical representation of sounds which make up words.

4.3. Lexical Modeling

Lexicon is developed to provide the pronunciation of each word in a given language. Through lexical model, various combinations of phones are defined to give valid words for the recognition. Neural networks have helped to develop lexical model for non-native speech recognition.

4.4. Language Modeling

Language model is the single largest component trained on billion of words, consisting of billions of parameters and developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary. ASR systems utilize *n*-gram language models to guide the search for correct word sequence by predicting the likelihood of the n^{th} word on the basis of the $n-1$ preceding words. The probability of occurrence of a word sequence *W* is calculated as:

$$P(W)= P (w_1, w_2, .., w_{n-1}, w_n) = P(w_1).P(w_2|w_1).P(w_3|w_1w_2) ... P(w_n|w_1w_2 ...w_{n-1}). \quad (1)$$

Language models are cyclic and non-deterministic. Both these features make it complicated to compress its representations.

5 TYPES OF SPEECH RECOGNITION

Speech recognition systems can be characterized by many parameters such as speaking mode, speaking style, vocabulary, language mode, dependency of text etc. Speech recognition systems can be classified into the one of the following four classes based on speaking mode and style of the end user.

5.1 Isolated Word Recognition –IWR

Isolated word recognizers usually require the speech to be recognized as a word. It does not mean that it accepts only one utterance at a time. If the input speech signal is phrase it accept that as a single unit with no explicit knowledge of phonetic content of the word and it requires the beginning and ending of the input speech clearly defined. HMM/CD and LPC/DTW are the best models to construct isolated ASR model [4]. This type of ASR suits for command and control applications.

5.2 Connected Word Recognition –CWR

Connected word systems are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them. Like isolated word speech recognition, it also requires the basic input speech utterance as a word/phrase.

There are some problem to be resolved before implementing the connected word recognition system: a) the number of words L in the string is not known b) the boundaries of the word in the string not known except for the beginning of the first word and the end of the last word of the same string c) the word boundaries are often fuzzy or non unique d) let the V be the set of V word reference patterns and L be the number of words in the string, there are V^L possible combinations of composite matching patterns to be tested. Rabiner et al [4] analyzed three algorithms designed for connected word recognition: Two level DP approach, Level Building approach and One Pass approach. These three algorithms are found to be provided identical best matching string with the identical matching score for connected word recognition.

5.3 Continuous Speech Recognition-CSR

Continuous speech recognizers allow users to speak almost naturally and there is no separate pause between words as in connected word recognition, while the computer determines the content. Finding the starting and ending of an utterance is a challenging factor in continuous ASR. As a result unknown boundary information about words, co-articulation, production of surrounding phonemes and rate of speech effect the performance of continuous speech recognition systems. It must utilize special methods to determine utterance boundaries.

5.3.1 Probabilistic Model of the Continuous Speech Recognition:

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model [6] of speech production whereby a specified word sequence, W, produces an acoustic observation sequence A, with probability P(W,A). The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability, i.e.,

$$\widehat{W} \ni P(\widehat{W}/A) = \max_W P(W/A) \quad (2)$$

Using Baye's rule, equation (2) can be written as

$$P(W/A) = \frac{P(A/W)\,P(W)}{P(A)} \quad (3)$$

In Equation (3), Since P (A) is independent of W, the MAP decoding rule of equation (2) is

$$\widehat{W} = \arg\max_W P(A/W)P(W) \quad (4)$$

The first term in equation (4), P(A/W) is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. To compute P(A/W), it is necessary to build statistical models for sub word speech units, build up word models from these sub word speech unit models (using a lexicon to describe the composition of words), and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods..

The second term in equation (4), P(W) is called the language model. It describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task.

5.3.2 Sub Word Speech Unit

There are several possible choices for sub word units that can be used to model speech include the following [4]

a) Phone like units
b) Syllable-like units
c) Dyad or demisyllable like units
d) Acoustic units

5.4 Spontaneous Speech Recognition System

In this type, the speech signal to be recognized is natural sounding and not rehearsed. An ASR system for such speech handles a variety of natural speech features. Large structured collection of speech is essential. Labeling and annotation of spontaneous speech is difficult. Some points to be noted are how to handle extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, repetitions, style shifting etc. Due to these factors the performance of spontaneous ASR gets degraded.

6. PERFORMANCE OF THE SPEECH RECOGNITION SYSTEMS

Performance of speech recognition systems are typically described in terms of Word Error Rate (WER) [7]. In Equation (5), the WER is defined as

$$WER = \frac{S+I+D}{N} \times 100 \quad (5)$$

Where N is the total number of words in the test set, and S, I and D are the total number of substitution, insertions and deletions respectively.

7. CONCLUSION

Speech recognition systems for English language have better progress than the any other languages because of the differences in morphology, grammar and other linguistic aspects. The ultimate goal of the researchers working in ASR is to enhance the natural communication between the man and machine by incorporating the fields Networks, Psychoacoustics, Linguistics, Speech Perception, Artificial Intelligence, and Acoustic-Phonetics etc. An attempt has been made through this paper to give an idea about Automatic Speech Recognition System.

AUTHOR BIOGRAPHY

Mrs.Geetha.K completed her B.Sc and M.Sc in Bharathidhasan University, Thiruchirappalli. She perceived her M.Phil in the area of Database Management System from Mother Theresa Women's University. She is a research Scholar of D.J Academy for Managerial Excellence, Coimbatore. She has attended many seminars and workshops conducted by various educational Institutions and presented her research papers. Her research interest lies in the area of Speech Recognition Systems, Data Management System and Neural Networks.

Dr.Chandra.E received her B.Sc from Bharathiar University, Coimbatore in 1992 and received M.Sc from Avinashilingam University, Coimbatore in 1994. She obtained her M.Phil in the area of Neural Networks from Bharathiar University in 1999. She obtained her PhD degree in the area of speech recognition from Alagappa University, Karaikudi in 2007. At present she is working as a Director at Department of Computer Science in SNS Rajalakshmi College of Arts and Science, Coimbatore. She has published more than 20 research papers in National, International Journals and Conferences. She has guided for more than 30 M.Phil research scholars. At present, she is guiding 8 PhD research scholars. Her research interest lies in the area of Data Mining, Artificial Intelligence, Neural Networks, Speech Recognition systems and Fuzzy Logic. She is an active member of CSI, currently management committee member of CSI, Life member of Society of Statistics and Computer Applications.

REFERENCES

[1] R. K. Aggarwal and M. Dave, Using Gaussian Mixtures for Hindi Speech Recognition System, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 4, December, 2011, pp.157 .

[2] B. H. Juang and L. R. Rabiner, Automatic Speech Recognition – A Brief History of the Technology Development, *Elsevier* Encyclopedia of Language and Linguistics, Second Edition, 2005.

[3] http://www.linuxdoc.org/HOWTO/ Speech-Recognition-HOWTO/introduction.html

[4]. Rabiner, Jung and Yegnanarayana, Fundamental of Speech recognition, Pearson Education, ©1993

[5]. E.Hocine Bourouba, Mouldi Bedda and Rafik Djemili, Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM.

[6] Saeed V. Vaseghi, Hidden Markov Models, Advanced Digital Signal Processing and Noise Reduction, Second Edition. John Wiley & Sons Ltd, ISBNs: 0-471-62692-9 (Hardback): 0-470-84162-1 (Electronic).

[7] L. R. Rabiner and B. H. Juang, Statistical Methods of Speech Recognition, Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.