International Journal of Engineering Research and General Science Volume 2, Issue 4, June-July, 2014
ISSN 2091-2730

Various Issues in Computerized Speech Recognition Systems

Shally Gujral[1], Monika Tuteja[1], Baljit Kaur[1]

[1]Electronics and Communication Department, PTU, Jalandhar, Anand College of Engineering and Management, Kapurthala

E-mail- gujralshally81@gmail.com , 09878235636

INTRODUCTION

Speech recognition is the translation of spoken words into text. It is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Speech Recognition is technology that can translate spoken words into text. Some SR systems use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription. Speech Recognition (is also known as Automatic Speech Recognition (ASR) or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.

1.1. Basic Model of Speech Recognition: Research in speech processing and communication for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing.[1] The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. This paper reviews major highlights during the last few decades in the research and development of speech recognition, so as to provide a technological perspective. Although many technological progresses have been made, still there remain many research issues that need to be tackled.

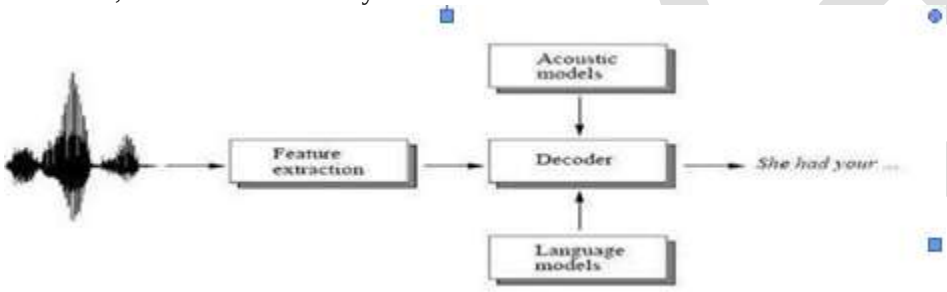

Fig 1 A Speech recognition system

TYPES OF SPEECH RECOGNITION SYSTEMS

A. Speaker dependant- A number of voice recognition systems are available on the market. The most powerful can recognize thousands of words. However, they generally require an extended training session during which the computer system becomes accustomed to a particular voice and accent. Such systems are said to be speaker dependent [2]. A speaker dependent system is developed to operate for a single speaker. These systems are usually easier to develop, cheaper to buy and more accurate, but not as flexible as speaker adaptive or speaker independent systems. Speaker–dependent software Work by learning the unique characteristics of a single person's voice, in a way similar to voice recognition. New users must first "train" the software by speaking to it, so the computer can analyze how the person talks. This often means users have to read a few pages of text to the computer before they can use the speech recognition software

B. Speaker independent - A speaker independent system is developed to operate for any speaker of a particular type (e.g. American English). These systems are the most difficult to develop, most expensive and accuracy is lower than speaker dependent systems. However, they are more flexible. Speaker–independent software is designed to recognize anyone's voice, so no training is involved. This means it is the only real option for applications such as interactive voice response systems — where businesses can't ask callers to read pages of text before using the system. The downside is that speaker–independent software is generally less accurate than speaker–dependent software.

C. Speaker adaptive - A third variation of speaker models is now emerging, called speaker adaptive. Speaker adaptive systems usually begin with a speaker independent model and adjust these models more closely to each individual during a brief training period.

3.AUTOMATIC SPEECH RECGNITION SYSTEM 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 2

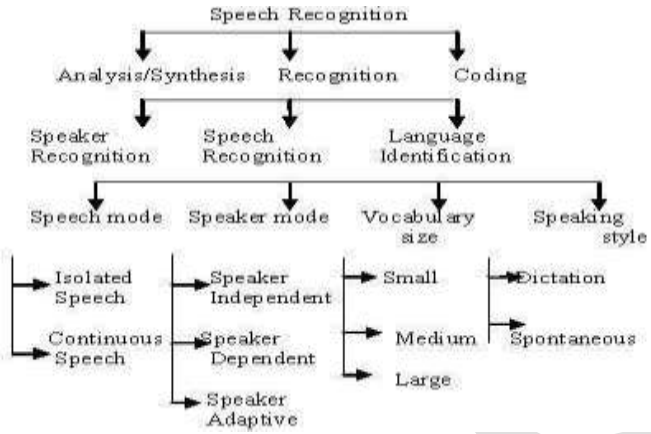

Fig. 2 Speech Processing Classification

4. RELEVENT 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
Channel	Band amplitude; distortion; echo
Speakers	Speakerdependence/independence Sex, Age; physical and psychical state
Speech styles	Voice tone(quiet, normal, shouted); (isolated words or continuous speech read or spontaneous speech) Speed
Vocabulary	Characteristics of available training data; specific or generic vocabulary;

Table 2 Speech Recognition Techniques

Techniques	Representation	Recognition Function
Acoustic Phonetic Approach	Spectral analysis with feature detection Phonemes/ segmentation and labelling	Probabilistic lexical access procedure
Pattern Recognition approach _ Template _ DTW _ VQ	Speech, samples, pixels and curves Set of sequence of spectral vectors Set of spectral vectors Features	Correlation distance Measure Dynamic warping Optimal algorithm Clustering function
Neural Network	Speech features/ perceptrons/ Rules/ Units/Procedures	Network function
Support Vector Machine	Kernel based features	Maximal margin hyperplane,Radial basis
Artificial intelligence approach	Knowledge based	

5. APPROACHES TO SPEECH RECOGNITION: Basically there exist three approaches to speech recognition[3]. They are: Acoustic Phonetic Approach B. Pattern Recognition Approach C. Artificial Intelligence Approach .

A. ACOUSTIC PHONETIC APPROACH:

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, 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. Even though, the acoustic properties of phonetic units are highly variable, both with speakers and with neighbouring sounds, it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units[4]. The next step is a segmentation and labelling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labelling. In the validation process, linguistic constraints on the task (i.e., the vocabulary, the syntax, and other semantic rules) are invoked in order to access the lexicon for word decoding based on the phoneme lattice. The acoustic phonetic approach has not been widely used in most commercial applications [5].The following table 2 broadly gives the different speech recognition techniques.

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. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern

comparison, from a set of labeled training samples via a formal training algorithm. 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) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns. The pattern-matching approach has become the predominant method for speech recognition in the last six decades [6]. In this, there exists four methods discussed below:

1. Template Based Approach:

Template based approach to speech recognition have provided a family of techniques that have advanced the field considerably during the last decades. 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 references templates and selecting the category of the best matching pattern. Each word must have its own full reference template; template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. One key idea in template method is to derive 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.

2. Stochastic Approach*:*

Stochastic modelling [7] 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 s, 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. The transition parameters in the Markov chain models, temporal variabilities, while the parameters in the output distribution model, spectral variabilities. These two types of variabilites are the essence of speech recognition.

3. Dynamic Time Warping (DTW):

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed, any data which can be turned into a linear representation can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models.

4. Vector Quantization (VQ):

Vector Quantization (VQ) [8] is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. Since transmission rate is not a major issue for ASR, the utility of VQ here lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure. In basic VQ, codebooks have no explicit time information, since codebook entries are not ordered and can come from any part of the training words. However, some indirect durational cues are preserved because the codebook entries are chosen to minimize average distance across all training frames, and frames, corresponding to longer acoustic segments (e.g., vowels) are more frequent in the training data. Such segments are thus more likely to specify code words than less frequent consonant frames, especially with small codebooks. Code words nonetheless exist for constant frames because such frames would otherwise contribute large frame distances to the codebook. Often a few code words suffice to represent many frames during relatively steady sections of vowels, thus allowing more codeword to represent short, dynamic portions of the words. This relative emphasis that VQ puts on speech transients can be an advantage over other ASR comparison methods for vocabularies of similar words.

C. Artificial Intelligence Approach (Knowledge Based Approach):

The Artificial Intelligence approach [9] is a hybrid of the acoustic phonetic approach and pattern recognition approach. 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. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. On the other hand, a large body of linguistic and

phonetic literature provided insights and understanding to human speech processing. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert's speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modelling. This form of knowledge application makes an important distinction between knowledge and algorithms. Algorithms enable us to solve problems. Knowledge enables the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

D. Connectionist Approaches (Artificial Neural Networks):

The artificial intelligence approach [10], Lesser et al. 1975; Lippmann 1987) attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analysing, and characterizing speech based on a set of measured acoustic features. Among the techniques used within this class of methods are uses of an expert system (e.g., a neural network) that integrates phonemic, lexical, syntactic, semantic, and even pragmatic knowledge for segmentation and labelling, and uses tools such as artificial NEURAL NETWORKS for learning the relationships among phonetic events. The focus in this approach has been mostly in the representation of knowledge and integration of knowledge sources. This method has not been widely used in commercial systems. Connectionist modelling of speech is the youngest development in speech recognition and still the subject of much controversy.

E. Support Vector Machine (SVM):

One of the powerful tools for pattern recognition that uses a discriminative approach is a SVM [9]. SVMs use linear and nonlinear separating hyper-planes for data classification. However, since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. The variable length data has to be transformed to fixed length vectors before SVMs can be used. It is a generalized linear classifier with maximum-margin fitting functions. This fitting function provides regularization which helps the classifier generalized better. The classifier tends to ignore many of the features. Conventional statistical and Neural Network methods control model complexity by using a small number of features (the problem dimensionality or the number of hidden units). SVM controls the model complexity by controlling the VC dimensions of its model. This method is independent of dimensionality and can utilize spaces of very large dimensions spaces, which permits a construction of very large number of non-linear features and then performing adaptive feature selection during training.

6.CURRENT AND FUTURE USES OF SPEECH RECOGNITION SYSTEM:

Currently speech recognition is used in many fields like Voice Recognition System for the Visually Impaired [10] highlights the Mg Sys Visi system that has the capability of access to World Wide Web by browsing in the Internet, checking, sending and receiving email, searching in the Internet, and listening to the content of the search only by giving a voice command to the system. In addition, the system is built with a translator that has the functionality to convert html codes to voice; voice to Braille and then to text again. This system comprises of five modules namely: Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Search engine, Print (Text-Braille) and Translator (Text-to-Braille andBraille-to - Text) module, was originally designed and developed for the visually impaired learners, can be used for other users of specially needs like the elderly, and the physically impaired learners. Speech Recognition in Radiology Information System. The Radiology report is the fundamental means by which radiologists communicate with clinicians and patients. The traditional method of generating reports is time consuming and expensive. Recent advances in computer hardware and software technology have improved Speech Recognition systems used for radiology reporting**. [**6] Integration of Robust Voice Recognition and Navigation System on Mobile Robot [7] and there are many other fields in which speech recognition can be used.

7. CONCLUSIONS:

This paper introduces the basics of speech recognition technology and also highlights the difference between different speech recognition systems. In this paper the most common algorithms which are used to do speech recognition are also discussed along with the current and its future use.

REFERENCES:

[1] 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.

[2] R.K.Moore, Twenty things we still don t know about speech,Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology , 1994.

[3] Behrang P. Dep. of Info. Science, UKM,Selangor,Mlaysia.hani_p114@yahoo.com

[4] Choo W.O.UTAR, Kampar, Perak, Malaysia.kenny@yahoo.comVoiceRecognition System for the Visually Impaired: Virtual Cognitive Approach, IEEE2008.

[5] Xinxin Wang1,Feiran Wu1,Zhiqian Ye11College of Biomedical Engineering& Instrument Science, Zhejiang University, Hangzhou, China meaita2009@gmail.com com, yezhiqian@hzcnc, The Application of Speech Recognition in Radiology Information System,IEEE2010.

[6] Huu-Cong Nguyen, Shim-Byoung, Chang-Hak Kang, Dong-Jun Park and Sung-Hyun Han Division of Mechanical System Eng., Graduate School, Kyungnam University, Masan, Korea Integration of Robust Voice Recognition and Navigation System on Mobile Robot, ICROS-SICE International Joint Conference 2009

[7] O. Khalifa, S. Khan, M.R. Islam, M. Faizal and D. Dol, ―Text‖ Independent Automatic Speaker Recognition , 3rd International Conference on Electrical & Computer Engineering,, Dhaka, Bangladesh, 28-30 December 2004, pp. 561-564.

[8] C.R. Buchanan, ―Informatics Research Proposal – Modeling the Semantics of Sound , School of Informatics, University of Edinburgh, United Kingdom, March 2005. http://ozanmut.sitemynet.com/asr.htm, Retrieved in November 2005.

[9] D., Jurafsky, ―Speech‖ Recognition and Synthesis: Acoustic Modeling , winter 2005.

[10] M., Jackson, ―Automatic Speech Recognition: Human ‖ Computer Interface for Kinyarwanda Language . Master Thesis, Faculty ofComputing and Information Technology, Makerere University, 2005.

[11] M.R., Hasan, M., Jamil, and M.G., Saifur Rahman, ―Speaker Identification‖ Using MelFrequency Cepstral Coefficients . 3rd International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 2004, pp. 565-568.

[12] http://project.uet.itgo.com/speech.htm

[13] http://www.speech.be.philips.com/index.htm