DOI:10.26524/jms.2015.9

An overview of speech recognition and its challenges

Banumathi.A.C[a] **and E. Chandra** [b]

[a] *Research Scholar,Mother Teresa Women's Unuiversity , Kodaikanal. Tamilnadu, India.*

[b]*Director, Dr. SNS Rajalakshmi College, Coimbatore – 32,TamilNadu.*

ABSTRACT: Speech Recognition means converting Speech into Text. This Emerging Technology makes all the field of use as more sophisticated one. The impact of this revolutionary Technology has shown its wide range of usage in all tasks. Almost all the Technical devices use the Speech recognition as their part of their project. Speech Recognition Technology used in fields like computers, artificial Intelligence, Medical , Healthcare, Smart Phones, etc., This paper provides a glimpse of the challenges that is faced by the speech recognition systems in many applications and the approaches taken to fulfill it.

Keywords: SR- Speech Recognition, IWR,- Isolated Word Recognition, DTW – Dynamic Time Wrapping, VQ – Vector Quantization

INTRODUCTION

Speech Recognition is a technology in which the spoken words are converted into text. Many devices uses this speech recognition as an important interaction between the user and the machine. People feel comfortable with speech rather than using input devices such as keyboard, mouse etc., Speech recognition would support many valuable application like dictation, Command control, Embedded machines, telephone directory assistance, medical application, translation into foreign languages and many more fields.

1. CLASSIFICATION OF COMPUTER

Speech Recognition can be classified into different types by describing the speech utterance, types of mode, Size of word, speaking style etc.,

2. TYPES OF SPEECH MODE

2.1 Isolated Word Recognition (Iwr):

The input here is just a single word. In the lack of an audio signal on both sides of the sample window the Isolated word recognizers usually require single utterance in a quiet atmosphere. It doesn"t mean that it accepts single words, but requires a single utterance at a time. The speaker should wait between the utterances. These systems have normally Listen/Not-Listen State.

2.2 Continuous Word Recognition (CWR):

It"s an advanced method of recognition. The continuous recognition is the most difficult task to determine because they must utilize special methods to determine utterance boundaries. Basically It"s like a Computer dictation. Here the speaker can speak fluently and naturally, while the computer determines the content.

2.3 Connected Word Recognition(CWR):

Connected word recognition are similar to isolated words, but allow separate utterances to be „run-together „ with a minimal pause between them.

2.4 Spontaneous Speech Recognition(SSR):

For the Basic level, it is a speech of natural sounding and not rehearsed. This system with spontaneous speech ability should be able to handle a variety of natural speech features with very minute variations.

2.5 Voice Verification /Identification:

Some Speech recognition system have the ability to identify a specific user by just hearing the commands given to it.

3. TYPES OF SPEAKER MODE:

Speakers of the system will have unique voices due to their physical body and personality. Based on the Speaker mode it can be classified into three categories.

3.1 Speaker Independent Mode:

It is based on Many users - to one System. This method is designed for many speakers. Any speaker of particular type (e.g. Indian English) can be made to use this mode of recognition. It recognizes the speech pattern of different people. It is more difficult and expensive method of recognition and even the accuracy is also lower than the speaker dependent mode. However they are more flexible and designed to recognize anyone‟s voice, so no training is involved.

3.2 Speaker Dependent Mode:

These systems depend on specific users. The software is trained extensively to get accustomed for a specific user. These are easier to develop, cheaper and more accurate for the trained user. This system is called speaker dependent and it learns the characteristics of the single speaker‟s voice, in a way similar to voice recognition. New user should train the software by speaking to it.

3.3 Speaker Adaptive Mode:

Speaker adaptive model usually begins with a speaker independent model and gets adjusted to each individual during a brief training period. This is a new emerging model.

4. Types of Vocabulary Size:

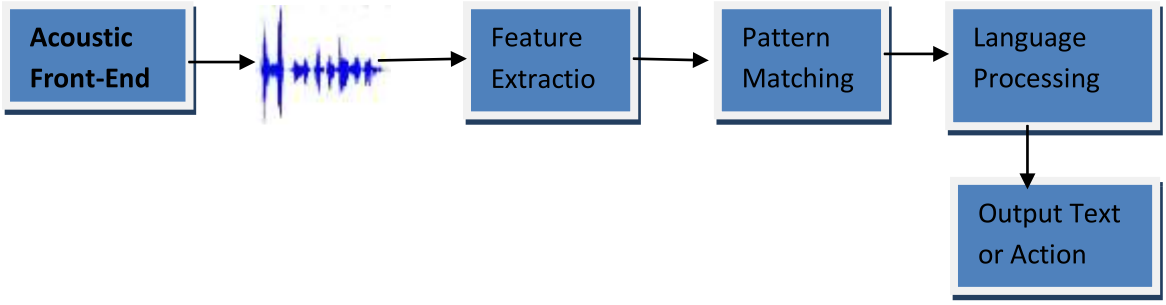
The Size of the vocabulary of a Speech Recognition system affects the complexity accuracy and the processing requirements. Some applications require small vocabularies whereas the others require large dictionaries of words. The vocabularies can be classified as follows.

- Small Vocabulary.
- Very Large Vocabular
- Medium Vocabulary.
- Out-of-Vocabulary
- Large Vocabulary.

In Addition to the other characteristics, the environment variability, channel variability, Speaking style, sex, age, speed of speech also makes the speech recognition system more complex. But an efficient system must cope up with all these barriers.

5. Basic Model of Speech Recognition:

Speech Recognition is the process of translating spoken words into text words on the computer. Speech Recognition is a process of Recognizing the Speech spoken by a Speaker and it has been in the field of Research for many years.



Voice communication is the most effective mode of communication used by humans. Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone to a set of words. The recognized words can be fed into the applications such as commands and control, data entry, and document preparation. Through a speech recognition program/application, the computer is able to process words you say and turn them into text on the screen just as if you had typed them on the keyboard.

An Acoustic model of Speech Recognition is created by taking audio recordings of speech, and their text as an input and produce as string of words. Input from the user is called as utterance. This utterance may be a single word or a sentence or a phrase or an entire sentence. This is the binary form of 1‟s and 0‟s that make up a computer programming Language.

Second is the Sound–recognition software that has acoustic models. An acoustic model takes the speech or their text transcriptions and using software it creates the statistical representation of the sounds that make up each word. [1] It is recognized by the speech engines to recognize speech. In the latest Technology the speech technology has been refined to eliminate the noise and useless information that is not needed to let the computer work. The words we speak are transformed into digital forms of the basic speech elements.(Phonemes) The Next step is the language processing which compares the digital dictionary that stored in the computer memory. Dictionary is a large collection of words usually more than 1,00,000 words. When it finds a match based on digital form it displays the words on the screen. This is basic concept of all the speech recognition systems and software. The template matching method of voice recognition is founded in the general principles of digital electronics and basic computer programming. To fully understand the challenges of efficient speaker- independent voice recognition, the fields of phonetics, linguistics, and digital signal processing should also be explored.

6. Approaches to Speech Recognition:

Basically there are three approaches to Speech Recognition. They are

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

6.1 Acoustic Phonetic Approach:

The earlier approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. The basis of acoustic Phonetic approach is that there exist finite, distinctive Phonetic unit in spoken language and these units are broadly characterized by a set of acoustic properties that are manifested in the speech signal over time. Even though the acoustic properties of phonetics units are highly variable, both with speakers and neighbouring sounds (articulation effect) it is assumed that the rules are straight forward and readily learned by a machine.

There are many steps involved in the acoustic phonetic approach. The first step is a spectral analysis of the speech combined with a feature detection that converts the spectral measurement is to set features that describe the broad acoustics property of the different phonetics units. The next step is a segmentation and labeling segmented in which the speech signal is segmented in to stable acoustic regions, followed by attaching one or more phonetic label to each segmented region, resulting in a phoneme lattice characterization of the speech . The final step is to find the word which is a valid word.

6.2 Pattern Recognition

The pattern-matching approach involves two essential steps, namely, pattern training and pattern comparison, The important feature of this approach is that is uses a well formulated mathematical framework and establishes consistent speech pattern representation for reliable pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches with each possible pattern learned in the training stage in-order to determine the identity of the unknown according to the goodness of the match of the pattern.

a. Template Based Approach:

Template based approach to speech recognition have provided a family of techniques that have advance features. 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. Usually templates for the entire words are constructed. This has the advantage that, errors due to the segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided. Each word must have its own reference template. When the vocabulary size increases beyond a few hundred words the matching becomes expensive. One key idea in template method is to derive a typical sequences of speech frames for a pattern via some averaging procedure, and to rely on some local spectral averaging procedure. Another key idea is to use some dynamic programming to temporarily align patterns to account for differences in speaking rates across the talkers.

b. Stochastic Approach

To deal with uncertain and incomplete information Stochastic model use the probabilistic models. In speech uncertainty and incompleteness arises from many sources. Eg., confusable sound, speaker variability, contextual effects and homophones words. Thus stochastic models are the particularly suitable approach to speech recognition. The most popular stochastic model today is the Hidden Markov modeling. A Hidden Markov model is characterized by finite state markov model and a set of output distributions. The two types of variabilities that are essential are the spectral variabilities and temporal variabilities.

c. Dynamic Time Wrapping (DTW)

Dynamic Time Wrapping 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 in another he is walking more quickly, if there were acceleration and declarations during the course of one observation. DTW has been applied to video, audio, and graphics and any data which can be turned into a linear representation can be analyzed with DTW. DTW is a method that allows a computer to find an optimal match between two given sequences with certain restrictions. The sequences are “wrapped” 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. One example of the restrictions imposed on the matching of the sequence is on the monotonicity of the mapping in the time dimension. Continuity is less important in DTW than in other pattern matching algorithms. DTW is an algorithm particularly suited to matching sequences with missing information, provided there are long enough segments for matching to occur.

d. Vector Quantization(VQ)

Vector Quantization is often used in efficient data reduction method. It is useful for speech coders. (data reduction). It is often applied to Automatic Speech Recognition (ASR). 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. Each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test speech is evaluated by all code books and ASR chooses the word whose codebook yields the lowest distance measure. In basic VQ, codebooks have no explicit time information since codebooks entries are not ordered and can come from any part of the training words. However, some indirect durational clues are preserved because the codebook entries are chosen to minimize average distance across all training frames, and frames corresponding to longer acoustic segments are more frequent in the training data. Such segments are thus more likely to specify code words than less frequent consonants frames, especially with small codebooks. Code words nonethless exist for constant frames because such frames would otherwise contribute large frame distances to the codebook. Often a few code words suffice t 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.

6.3 Artificial Intelligence Approach (Knowledge based approach)

It is a hybrid form of the phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of both the approach. 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 template based approaches have been very effective in the design of a variety of speech recognition systems, they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. 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 the expert systems. Another difficult problem is the integration of many levels of human knowledge phonetics, 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 modeling. 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 al 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.

a. Artificial Neural Networks.

The Artificial Intelligence approach attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analyzing 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(eg., a neural network) that integrates phonetic, lexical, syntactic , semantic, and even pragmatic knowledge for segmentation and labeling, 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. In connectionist models, knowledge or constraints are not encoded in individual units, rules, or procedures, but distributed across many simple computing units. Uncertainty is modeled not as likelihoods or probability density functions of a single unit, but by the pattern of activity in many units. The computing units are simple in nature, and knowledge is not programmed into any individual unit function; rather, it lies in the connections and interactions between linked processing elements. Because the style of computation that can be performed by networks of such units bears some resemblance to the style of computation in the nervous system. Connectionist models are also referred to as neural networks or artificial neural networks. Similarly, parallel distributed processing or massively distributed processing are terms used to describe these models.

Not unlike stochastic models, connectionist models rely _critically on the availability of good training or learning strategies. Connectionist learning seeks to optimize or organize a Network of processing elements. However, connectionist models need not make assumptions about the underlying probability distributions. Multilayer neural networks can be trained to generate rather complex nonlinear classifier or mapping function. The simplicity and uniformity of the underlying processing element makes connectionist models attractive for hardware implementation, which enables the operation of a net to be simulated efficiently. On the other hand, training often requires much iteration over large amounts of training data, and can, in some cases, be prohibitively expensive. While connectionist appears to hold great promise as plausible model of cognition, may question relating to the concrete realization of practical connectionist recognition techniques, still remain to be resolved.

b. Support Vector Machine.(SVM)

One of the powerful tools for pattern recognition that uses a discriminative approach is a SVM. 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 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 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 dimension of its model. This method is independent of dimensionality and can utilize spaces of very large dimensions spaces, which permit a construction of very large number of non-linear features and then performing adaptive feature selection during training. By shifting all non-linearity to the features, SVM can use linear model for which VC dimensions is known. For example, a support vector machine can be used as a regularized radial basis function classifier.

7. Challenges it faces in the current Technology

There are some significant features that can be taken into consideration for the better performance of the Speech Recognition. The top most priorities are given to Improve their Training, Flexibility of the Equipment, Accuracy, throughput, Latency, User Responsive , Performance, Reliability, and Fault Tolerance.

a. Training Voice recognition Software

In order to have a useful voice recognition system, users must first train the dictation software to recognize their voice. This is often the point at which many new users become frustrated. The most common complaint about voice recognition software is that the training was a waste of time. Nevertheless, the most recent generation of voice recognition software has proven exceptionally accurate, once the user can establish good, effective voice files. The following suggestions for enhanced accuracy will help avoid the frustration often associated with the training of voice recognition software.

b. The Equipment

i. A Good Signal

Successful voice recognition requires a clear recording of the user's voice into the computer. There are two components to clear recording: microphone quality and sound card quality. Both of these components seriously affect the accuracy of the voice files during training.

ii. The Microphone

The majority of voice recognition software packages are sold with an acceptable microphone, generally worn on a headset. However, the microphone often does not continue to function over an extended period of time. This creates a problem; It is beneficial to use the same microphone for dictation that was used for training. Many students have found success in training the software using a higher-quality microphone, purchased separately.

iii. The Sound Card

New computers typically have an adequate sound card, installed by the manufacturer. At times, however, the sound card signal is not sufficient for good voice recognition. A quick solution to this problem is to purchase a USB (universal serial bus) sound system or USB microphone. A USB sound connection bypasses the sound card, bringing the signal directly to the dictation software.

c. Using the Equipment

i. Fluent Readers

Given a clear signal from the microphone and sound card, a fluent reader should be able to train voice recognition software simply by following the directions provided in the set-up information. Generally, expect to read at least three of the stories provided before the software will be sufficiently accurate. Although the dictation software may be fairly accurate after one passage is read, investing more time in training will greatly improve the overall results.

ii. Students With Difficulty In Decoding

Often, the students who most need voice recognition software are also the students for whom the training can be the most frustrating. For accurate speech recognition, the passages should be read fluently, in well-articulated phrases. If the person training the software struggles over words and makes frequent reading mistakes, the software will make mistakes when dictating. For students with a low reading level, training dictation software requires several one-on-one sessions with an experienced instructor.

8. Instructor Initializes Training

The instructor completes the initial training; the student can refine it later. Usually, training programs require the user to read several longer stories, before being allowed to read the shorter ones. In order to get through the longer stories quickly, and with some fluency, these may be read by the instructor. It is important that the instructor have vocal characteristics similar to those of the intended user. For example, if the software will be used by 14-year old female with a history of social anxiety, it would not be suitable to have an aggressive football coach initiate the training.

9. Phrase-by-Phrase

Having completed one or two longer stories, the training program should then make available several shorter readings. At this point, the student should dictate several shorter readings, with the instructor reciting the passage quietly, over the student's shoulder. The voice recognition program is designed to recognize fluent, continuous speech; therefore, the instructor should pronounce complete phrases for the student to dictate into the microphone. If the instructor's voice is picked up by the microphone, it could interfere with the training process.

10. "Correct That"

Many voice recognition software packages also have a "correct that" feature, which allows the student to amend dictation errors and further refine the training. As the software is used, and corrected, it will learn the student's voice more accurately with each succeeding correction.

11. Back-Up Voice Files

Once accurate voice training files have been established, they should be copied onto a CD-ROM or a zip disk. By creating a back-up of the voice files, the student will not have to go through the training again if the computer ceases to function. Also, it is fairly simple to transfer the files onto the new hard drive, if the student decides to upgrade the computer.

i. Pro’s and Con’s Concerning Voice Recognition Software

- Students don't have to use a keyboard to input information. The software has to be trained to recognize the user's voice. This is accomplished by reading passages provided by the program.
- The software learns to recognize a student's unique speech patterns. Users have to speak distinctly in order for the software to work well. If the student has non-standard speech, tends to run words together, or mumble, the training process may be long. Some punctuation must be dictated.
- The software spells every word correctly. The software spells every word it recognizes correctly. Typically, it recognizes 5-20% words incorrectly. It cannot recognize homonyms.
- Students can write as quickly as they speak, 100+ words per minute. While users may talk that fast, what they produce will probably be disorganized and grammatically incorrect.
- Students can produce a large amount of writing, which they can then edit. Users have to edit.
- Students can write papers without being held back by spelling or keyboarding problems. Users can easily get words written down, but there is much more involved in writing a paper than just putting down words.
- The software will read back to students what they have written, helping with proofreading.
- Voice recognition software use is expanding rapidly. Both Windows and Macintosh operating systems have voice recognition built in. Voice recognition uses a lot of memory. The software has specific hardware requirements.

Conclusion

Speech is a most important means of communication. People find an alternative to text input by giving speech input, Which makes the usage of computer more efficient and convenient. Voice operated Intelligent machines have been an great interest in the in the current technology. Future systems need to have an efficient way of representing, storing, and retrieving knowledge required for natural conversation. For the past two decades the progress is significant in the field of speech recognition. Speech Recognition technology is one of the most interesting problem in and of itself. The Field of Speech Recognition in future is expected to undergo many changes. The computers may even talk to us back just like a human. Speech-recognition has come a long way towards making the world more accessible. It changes the way we use computers, automobiles, cell phones, and video and provides great accessibility for the hearing impaired and for foreign language speakers. Current leaders in the field are Nuance Communications and Google with their Automated Speech Recognition (ASR) technology. Speech recognition still has a long ways to go, though. There are a lot of steps between you reciting a sentence and the computer or phone writing those words out on the screen. To better explain the science and impact of automated speech recognition, Medical Transcription has created a technique called infographic that goes through the technology behind speech recognition.

.

REFERENCES

[1] Frances Alias, Xavier Servillano, Joan Claudi socoro and Xavier Gonzalvo "Towards High- Quality Next Generation Text-to-speech Synthesis:A multi domain Approach by Automatic Domain Classification",IEEE Transactions on AUDIO,SPEECH AND LANGUAGE PROCESSING, VOL16,NO,7 september 2008.

[2] Sadoki Funni, "50 Years of Progress in speech and Speaker Recognition Research" ,ECIT Trasaction on Computer Information Technology, Vol.1.No.2, November ,2005.

[3] K.H.Davis, R.Biddulph, and S.Balashek, "Automatic Recognition of Spoken Digits" , J.Acoust.Soc.Am., 24(6): 637-642. 1952.

[4] 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.

[5] JR.M.Moore, Twenty things we still don"t know about speech, Proc.CRIM/FORWISSWorkshop on Progess and prospects of speech Research and Technology, 1944

****