IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 9, September 2016

Speech Recognition Operation with Efficient Accuracy Rate and Factors Affecting It

Pritam Padhye[1], Vijaya Chavan[2], Sagar Mane[3]

Student, Computer Technology, Navi Mumbai, India[1, 3]

Professor, Computer Technology, Navi Mumbai, India[2]

Abstract: Speech has always been an important activity in human life and also contributing a lot to the society. The interaction between human and computer or maybe any computing device became a new trend and sometimes a need of an hour. Speech recognition system actually converts the analog spoken inputs to the digital signals that a computer can understand and do the required task or provide the user with the output. This paper gives a brief introduction to this lime lighted topic and explains few of its application, in our information world and also touch topic likes accuracy, and working of the speech recognition, with its future use. This paper will also include the models of speech recognition.

Keywords: recognition pattern**,** models, accuracy, spontaneous speech, noise.

I. INTRODUCTION

Speech has always been the best way of communication between people and will always be. The problem of communication with computer led to a heavy research on speech recognition before this the communication between user and computer was simple click method which was suitable for a limited process but researchers wanted a more enhanced communication for the betterment of the people and break the small thin barrier between user and computer. This thin line was of knowledge of using a computer as a person should know what happens when we click somewhere but in case of speech recognition the user just needs to speak as he speaks with a normal individual person.

In case of speech recognition, the spoken analog signals are converted into the digital signals which a computer can understand using many algorithms which are then converted to computer programs. At this point of date speech recognition technique can analyze thousands of words from many mainstream languages (English, German, Spanish, Chinese, Hindi etc.).

II. RELATED WORK

The development in the field of speech recognition was initiated by the AT&T laboratories since many years. This speech recognition system has gone through many years of research and experienced many developments too. The researchers have done a notable work in the field of speech recognition and have also produced a few facts which can be still worked on for the enhancement in the field to provide a better service towards mankind so as to develop as fast as possible a few notable work which can be worked on is as follows: -

Atsunori Ogawa, Member, IEEE and others in their topic titled Estimating Speech Recognition Accuracy Based on Error Type Classification have specified the types of errors and the way or method to detect those errors and also to solve them [1], but they have ignored the working of the speech recognition system as a whole and also missed the application of the speech recognitions system which has been included in this paper as the main part of the paper titled as general structure and language analysis. The part of accuracy is also not missed while keeping working in the focus.

Bin Zhang and others in their IEEE paper names Study on CNN in the recognition of in emotion in audio and images have stated the use of speech and image recognition that would go hand-in-hand with the Convolution Neural Network (CNN) [2]. They have also experimented this idea and the project was successful with few flaws. But they failed to include types of input as their point which is included in this paper as a promotable point.

Jinmook Lee, Student Member, IEEE and others in their IEEE paper titled as an Energy-efficient Speech Extraction Processor for Robust User Speech Recognition in Mobile Head-mounted Display Systems have worked on the topic of disturbing voice as an added element as the noise is also a part which can hamper the accuracy of the speech recognition system [3]. This part is present in this paper too which explains the process of separating the desired speech out of noise.

Chien-Lin Huang and others in their paper titled Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis have focused on the speech recognition in local or reginal language using the analysis [4], have been included as a subtopic in this paper but they failed to explain the effect of noise on the speech recognition system also missed points like accuracy of the system in their paper which

DOI 10.17148/IJARCCE.2016.5932

have been discussed in this paper. This paper includes nearly all the topic which is essential for a person to know about the speech recognition system as a whole and aims to keep this fields developing to take the interaction between the human and computer to the next level. This system will take over the user interface within few years.

III. GENERAL STRUCTURE OF SPEECH RECOGNITION

The general structure of the speech recognition is a block of different processes clubbed into one so as to make it successful operation in the work of computation and in turn make it useful for the human – computer interaction as this is the main aim of speech recognition.

The steps of speech recognition are as follows as explained below with the reference of the diagram

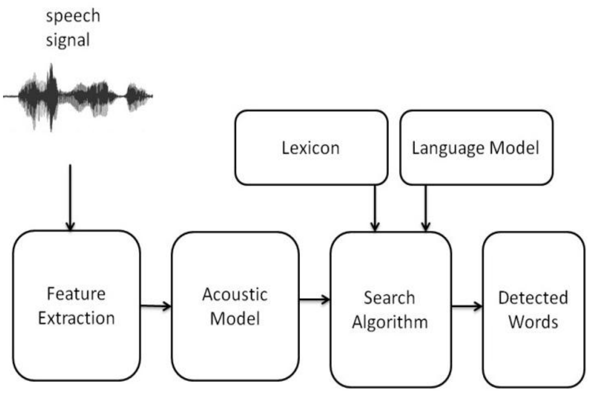

Fig 1- the block diagram for speech recognition system

Steps: -

1. The first step is to provide the input to the system so as to take the necessary action to fulfill the required task [6].
2. The second step is analysis this step includes the grammar checking and language related option which may differ according to the different language [6].
3. The third step is related to phenomes words or the pattern matching of the words that are given an as the input by the user [6].
4. Language processing is the last step to be performed by the system to generate the output for the user. [6]
5. The last step is the output process which depends on the input and the output may vary from opening an application to performing complex task depending on the system [6].

IV. TYPES OF INPUTS

One has to provide the basic input so as to get the desired output after performing the system task. The inputs are given in different formats ranging from single numerical to sentences. In case of speech recognition, the input is the users voice this voice is in form of words and syllables these words or syllables are classified as follows: -

A. Isolated Word

Isolated word recognizes attain usually require each utterance to have quiet on both side of sample windows. It accepts single words or single utterances at a time. This is having “Listen and Non Listen state”. Isolated utterance might be better name of this class. [5]

B. Connected Word

Connected word system are similar to isolated words but allow separate utterance to be “run together minimum pause between them. [5]

C. Continuous speech

Continuous speech recognizers allow user to speak almost naturally, while the computer determine the content Recognizer. With continues speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries. [5]

D. Spontaneous speech

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed .an ASR System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together. [5]

V. RECOGNISING PATTERNS

The next process in recognizing the speech for getting the desired output. Broadly speaking the approach towards the process of recognizing the speech has different option like 1. Simple pattern matching, 2. Pattern and feature analysis.

A. Simple pattern matching

The simplest kind recognizing the speech as the input is the simple pattern matching the best example of pattern matching is the automated speech during a computerized call for the banks and other automated call centers.

In this type one has to just speak a few numbers like maybe an account number or other numerical data which is far most easy to as compared to recognizing a sentence or a group of words as matching a number like ‘zero’, ‘one’ ‘four’. Is easier as it doesn’t require any added dictionary words to recognize the spoken numbers.

B. Pattern and feature analysis

Simple pattern matching is good for a selective vocabulary like numbers which is restricted to a specific domain like as explained an automated call or a bank for knowing the customer details. Most of us use a large vocabulary made from a thousands of words which are some commonly used and some are rarely used.

The problem of this type is that it is highly inefficient as a to train a computer to each and everything to imagine is very hard taking an example when a human brain thinks of a truck it is a general concept that a vehicle that can carry goods and has high power of operation but for a computer we can’t give generalized concept of anything it needs a specific truck to be looked.

DOI 10.17148/IJARCCE.2016.5932

VI. LANGUAGE ANALYSIS

A. The Recognizing Process
Basically the recognizing process is that a person speaks into any input device like a microphone and the related output is given out it maybe a process to be performed or a message to display. [5]
The process or recognizing the speech is a complex computation task for a computer this is done by the algorithms the input may range from a binary format or a numerical or maybe even a complex collection of sentences or words etc.
The speech spoken by the person is converted a spectrogram which is used to analyze the spoken words into graph like structure which tells the intensity of each word spoken which helps to recognize the word that is spoken.
To convert speech to on-screen text or a computer the computer has to go under a series of operation which need an analog to digital converter so as to convert the analog spoken data into a digital data in this process the words may sound similar as the language requires or is used then these similar sounding data is matched using a phoneme there are very limited phonemes in English but they vary according to the language used.

B. Statistical analysis.
Speech recognition is a far complex job as it may seem easy but has many steps involved in it like firstly recognizing the data then processing it then giving the output. Every person has a different way of speaking in his/her tone in which his/her way of speaking and his pronunciation may differ this change may affect the overall performance of the software that is deployed for the speech recognition purpose [1].

For example, a word may sound similar like ‘zero’ may sound like ‘hero’ or the best example of similar sounding words is ‘two’ or ‘to’ or ‘too’ all these words maybe be sounding similar but their application is totally different thus the system has to recognize the need of grammar as well this requirement is fulfilled by the statistical analysis and language model which is different in each language.

C. Types of statistical analysis:
1) Acoustic Model (AM):
One of the most widely adopted models of speech recognition is acoustic model (AM). It has been established that acoustic models of speech recognition capture the characteristics of the basic recognition units. According to, the recognition units can be at the phoneme level, syllable level, and at the word level. Several inadequacies and constraints come into consideration with the selection of each of these units. [8]

2) Language Model (LM):
Language model is another most imminent statistical model of speech recognition. One of the major objectives of language model is to convey or transmit the behavior of the language. It is due to the fact that it intends to forecast the existence of the specific word sequences within the target speech. [9]

3) Lexicon Model:
Lexicon model provides the pronunciation of the words within the target speech, which has to be recognized. According to the perceptions the lexicon model plays an inevitable and indispensable role in automatic speech recognition. It is due to the fact that the operations of lexical model are based on two parameters, i.e., whole-word access, and decomposition of entire speech into small chunks. This process eventually results in appropriate recognition of the speech. For instance, if speech recognition models are in native language, the lexicon model has to be formulated in the native languages, in order to acquire valuable and useful results. [8]

4) Hidden Markov Models (HMM):
hidden Markov model became the most popular statistical tool, which is being used for the modeling of data. It has been analyzed that hidden Markov model has played a commendable role in reducing the issues of speech classification, which was one of the core issues, within the speech recognition approach. Hidden Markov model incorporated various issues, which used to affect the accuracy of speech recognition. In order to resolve those issues, many different algorithms based on Hidden Markov model have been incorporated. [8]

VII. ACCURACY OF SPEECH RECOGNITION

In today’s information world the performance of software and its accuracy is the main parameter to be judged for the popularity and efficient stay is the market hence this paper has added the accuracy as the prominent sub topic.

The process of error detection is classified into two types based on the type of error that has occurred in the system namely [1]

1. Word alignment network(WAN)
2. Conditional random fields(CRF)

The process of accuracy estimation is done firstly by WAN and then the second step involves the use of CRF’s to refine the accuracy of the speech recognition system. These two basic accuracy factors can be hybridized to form different algorithms for detection of accuracy and also improve the same [1].
Accuracy is the average errors a system does when an input is given it will increase if the language has large number of words and have a complex speaking pattern. Word error rate is a common unit of specifying the performance of the speech recognition system which has many parameters to judge other than the size of dictionary of that language.
First the recognized words are arranged when compared with the words that were given as the input to the system

then the words which were not recognized are sorted depending on the input similar to the first step. Then the average of this is found using a formula which is based on the power law that states the correlation between perplexity and word error rate.

The word error rate is given by the following formula: - where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of the corrects, N is the number of words in the reference (N=S+D+C). [10]

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$$

VIII. NOISE AS A BARRIER IN SYSTEM

Speech recognition has recently come into the spotlight as a User Interface in the head-mounted display (HMD) system such as a Google Glass. However, speech recognition (SR) accuracy is significantly degraded (~20%) in the multi-sound source environment, because other unrelated sound sources interfere the user's voice. In this situation, the SR accuracy can be enhanced by Speech Extraction (SE) technique, which extracts only user's voice from the noisy signal. So the speech extraction hardware is implemented for the robust speech recognition in multi-source environment. Moreover, the SE concept can be applicable to another kinds of signals, such as bio-potentials, therefore several kinds of hardware have been implemented to enhance the target signals. [3]

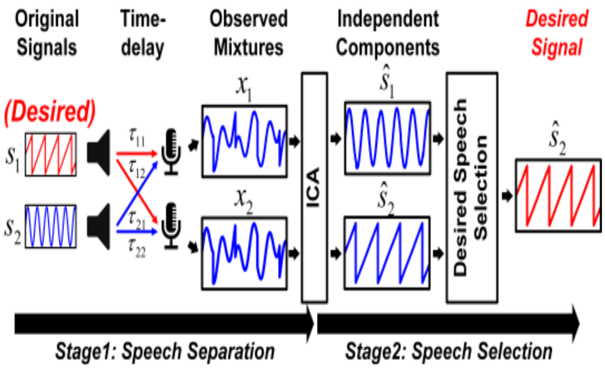

Fig 2 – Noise separation for desired speech.

Fig 2 describes the overall flow of the SE. Basically, it consists of two stages: speech separation and speech selection. In speech separation, observed mixtures (x1, x2) are separated into Independent Components that are similar with the original signals (s1, s2) before they are mixed. Generally, Independent Component Analysis (ICA) is used for this stage. However, ICA cannot distinguish which signal is the desired user speech among the separated ones due to permutation ambiguity. Therefore, the speech selection stage has to be carried out for decision of what is the user speech. [3]

IX. CONCLUSION

In this work the accuracy and factors causing distortion in the speech like noise and human traits of speaking have been discussed and also other concepts like working of speech recognition system with the speech processing have also been covered. This paper also touches point like how an error is found out or estimated with the help of algorithms like WAN and then refined by using CRF algorithms. This paper also discusses the types of input like isolate words, continues words or even spontaneous words. We have tried our level best to take this topic to a higher level for more enhancements in the same.

ACKNOWLEDGEMENT

Several people deserve thanks for the contribution in projecting this paper. Thanks to our guide professor of Bharati Vidyapeeth Institute of Technology for the efforts that she took on us. We thank the college for providing the resources and helped in all ways possible. We also thank readers of this journal for reading this topic and contributing towards the enhancement of this topic as well.

REFERENCES

1. Atsunori Ogawa, Member, IEEE, Takaaki Hori, Senior Member, IEEE, and Atsushi Nakamura, Senior Member, IEEE "Estimating Speech Recognition Accuracy Based on Error Type Classification", JOURNAL OF LATEX CLASS FILES, VOL. 13, NO. 9, SEPTEMBER 2014
2. Bin Zhang, Changqin Quan, Fuji Ren," CNN in the recognition of in emotion in audio and images", IEEE.
3. Jinmook Lee, Student Member, IEEE, Seongwook Park, Student Member, IEEE, Injoon Hong, Student Member, IEEE, and Hoi-Jun Yoo, Fellow, IEEE," An Energy-efficient Speech Extraction Processor for Robust User Speech Recognition in Mobile Head-mounted Display Systems". DOI 10.1109/TCSII.2016.2571902, IEEE
4. Chien-Lin Huang, Student Member, IEEE, and Chung-Hsien Wu, Senior Member, IEEE "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis", IEEE TRANSACTIONS ON COMPUTERS, VOL. 56, NO. 9, SEPTEMBER 2007
5. Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010
6. Parwinder pal Singh, Er. Bhupinder singh, "Speech Recognition as Emerging Revolutionary Technology", International Journal of Advanced Research in Computer Science and Software Engg 2 (10), October- 2012, pp. 410-413
7. Suma Swamy1 and K.V Ramakrishnan, "AN EFFICIENT SPEECH RECOGNITION SYSTEM", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013
8. Khaled M. Alhawiti, "Advances in Artificial Intelligence Using Speech Recognition", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:6, 2015