American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Voice Recognition for English Language Pattern Recognition Approach

Asad Ullah[a]*, Lu Xuan Min[b]

[a]*School of Electronics and Information, Northwestern Polytechnical University*
ADD:127 West Youyi Road,Xi'an Shaanxi,710072,P.R.China
[b]*Associate Professor at School of Electronics and Information, Northwestern Polytechnical University*
ADD:127 West Youyi Road,Xi'an Shaanxi,710072,P.R.China
[a]*Email: Asad_uop92@yahoo.com*
[b]*Email: luxuanmin@nwpu.edu*

Abstract

Communication is the basic need of everyone for a person who want to survive in this world. Researcher are very focusing from last couple of decades in the area of speech recognition. Some peoples cannot communicate in the proper way just like physically hampered and color blind etc. For this purpose a lot of researcher and mankind people working on automatic speech recognition. ASR also sometime call computer speech recognition. In this modern world time is one of the most important factor to be saved as much as possible. Due to a lot of computer software and advancement in a science provided, now we are capable to process a lot of software which were impossible to access before few years.

In this paper we are going to discuss something about voice recognition through different feature like HMM (Hidden Markov Model), acoustic model etc. In this paper we will focused on accuracy because accuracy is the basic key factor in speech recognition. Every time environment change, speaker is not always be fixed person, there is also variation in context occurs and also how to maximize the size of vocabulary will be our aim and goal that will be discussed in this review paper.

Keywords: English Language model; Hidden Markov Model; Feature Extraction; acoustic Model Introduction.

---

* Corresponding author.

1. Introduction

From the origin of the universe everyone is not able to understand the other language and everyone is not able to speak the all other languages very well. Speech is the process that start from the mind of talker formulating words to listener by mean of voice. So we researcher always trying to developed such a mechanism through which other people can easily understand what the talker want to say and also convenient for them to be used. From last couple of year researchers are very focusing that how speech waveform are generated and how the voice can be recognize [1].

Here we will monitor the sound input and we can use different techniques and process to recognize but we are fail to know about the feelings although we have a lot of different software and process to be used, but still we are fail. If someone say “oh shit! I am fail”. The human can recognize what he or she is saying, but computer is not unable to feel his or her feeling of happiness etc. Only human have this quality. But here our concentration will be on ASR and ASR is a system through which the voice is converted into words [2].

1.0 What is voice part?

Speech recognition is the application of digital signal processing (DSP) techniques for the processing and to analysis of speech signals. Speech is made from adding or moving of multi frequency at a same time. Speech is a type of non-stationary and continuous signals and used to convey information, that are based on two parts.

- Voice part
- Non voice part

1.1 Voice part

That part is interested for us. This part contain the voice part. The frequency of that part is low.

1.2 Non voice part

This part contain the unvoiced part of the input signals. In this part we are not interested and that is the reason we always neglect it. Also this part have the high frequency.

2. Base line of the ASR System

The basic thing which we have to consider the most is to choose the relevant (probable) sequence of words (hypothesis) from the audio signal (given as the input sound) with respect to all these assumptions.

How to pick the relevant sequence of words is decided by two things statistical approach which is consist of acoustic and linguistic constraint of that specific language. In ASR system we mostly used Fourier transform, MFCC, LPC (linear predictive coding) etc. the use of software depends on our resources, knowledge, requirement and easiness. Also reliability is the key word for using a method.

As we know that the assumption is made from the signals received as an input in a form of voice. So on the base of that voice we made a sequence of words through assumption.

3. Methods of Speech Recognition

There are many approaches through which speech can be recognized, but some are very common with a lot of benefits and reliability. Some are given below.

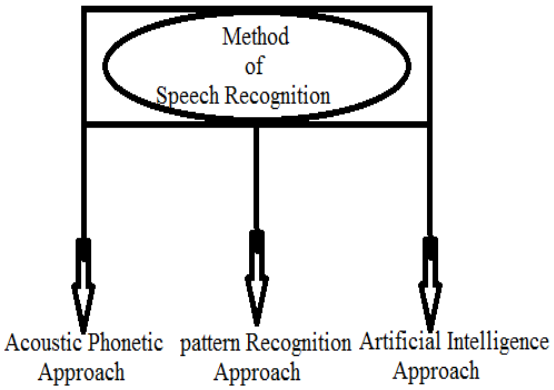

Figure 1: Methods

The above mentions approaches are very common and wisely using approaches. Here I will give a short explanation of each.

3.1 Acoustic phonetic Approach

In this approach, the spectral analysis of the speech combined with the feature detected that convert the spectral measurement to the set of feature which described the broad acoustic properties of the different acoustic unit. This is a model that represent the property of the specific language to be processed but along this, another property which can handle or solve any problem that are the variable characteristics of that specific language. Which will be under observation on that spot. We can also explain the acoustic model as that basically acoustic model is the unit that represent sequence associated with elementary sound with the specific language. This approach is not widely used commercially nowadays [3].

3.2 Pattern Recognition Approach

We will used this approach in our system because this approach is very dominant on all other methods due to their unique approach and from last couple of decades this method is using very frequently. This approach work in statistical model or speech template [4]. E.g. HMM. In this approach we can able to recognize a word or a phrase. Even for smaller than word, this approach is also applicable.

During this approach we first train the system and then pattern is generated from the input speech signals after that we compare this pattern to that pattern that we saved during training of the system. And in last best match is

picked and is consider as a spoken word or phrase. So it mean that introduced sound during training and newly introduced sound pattern are compared and output is gained. Pattern comparison is the main stage of such approach. Pattern recognition give us high accuracy but for very small vocabulary with speaker dependent system.

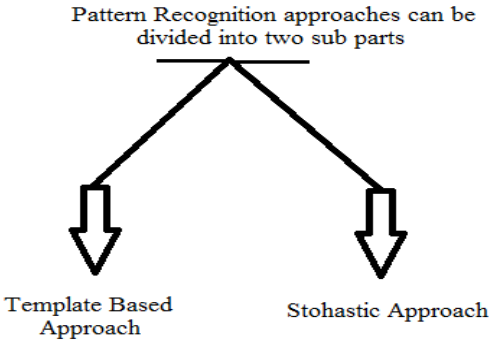

Figure 2: Approaches

3.2.1 Template Based Approach

This is a sub branch of pattern recognition approach [5]. This approach is a bit advance from other approaches. In this approach the speaker dictionary is store before and from that dictionary the pattern is match of speech utterance.

3.2.2 Stochastic Approach

This approach is perfect for all that situation in which we are facing uncertainty or we have incomplete information. Basically we are trying our best to deliver a perfect and full information as input but there are a lot of factors that affect our experiments just like confusable sound, distortion etc. so this approach is best for all those probabilistic situation. Hidden Markov Modeling is very popular now a days in stochastic approach.

In speech recognition two types of variability are the essence of the system.

- Temporal variability during transition parameters.
- Spectral variability is other parameter which is the output parameter of HMM.

As compare to other approaches this approach has a solid mathematical base. More and easy integration of knowledge source is also one of the key factor of this approach. The main drawback of HMM is that HMM have highly accurate but need a large amount of training data.

There are some other approaches that are not very common among us

- Artificial Intelligence approach
- Vector Quantization

- Taxonomy of Speech Recognition
- Dynamic Time warping
- Support Vector Machine
- Artificial Neural Networks

4. Classification of ASR

There are a lot of method through which we can divide the ASR, but the main issue that which utterance can be recognize very well by a specific class. We also keep one thing in our mind that which process or class will be easier or reliable for us.

5. Implementation Area

- Handicapped people
- Automobile
- Eye sight effected people
- Traveling
- Banking
- Offices
- Public areas
- During war
- Defense department

ASR system is best to use for all those users who are not good in using English (not convenient to use English) and they preferred to user their native language like Arabic, Hindi, German, French, etc.

The system is also ideal for the offices and other people in which we have to do a lot of work and we have less strength. In other words for our easiness and fast accomplish we are using ASR [6]. For different kind of making report we can also use ASR [7].

6. Priority of our system

- Robustness of the system
- Level of accuracy
- Decrease the human machine performance gap
- Increase the level of dictionary

7. Accuracy dependent factors

- Environment
- Transducer

- Channel
- Speech style
- Vocabulary
- Training time

If amplitude decrease, the error will increase so if the example or sample for testing should be taken in such areas where ambient noise is extreme or low definitely the output or response will be changed.

8. Procedure

8.1.1 Input

In the very first part of the system we can check the availability of the sound. Here we can used two techniques for the presence of voice, one is led blinking through small sound sensor and the other way to check is zero crossing on the analyzer machine. If there are sound we can proceed further otherwise we should check this part of the system for further assistance.

In the next part which we will consider as second part of the system, we are using the device or sub system that can be able to recognize the sound and further proceed it. For this purpose we use a device that can extract the feature of the sound and as a result to generate Mel frequency Cepstrum Coefficient.

8.1.2 Feature extractor

Feature extractor are the area from where we can get a set of parameters as a converted signals from the voice signal. Feature extraction can improve the accuracy of the system by creating better feature.

Initially we applied the sound signal to this component, the input signals are induced to this component directly or first we can record it, after that the recorded sound may be injected to it, the input sound will be given to the system through micro phone. Better quality micro phone can enhance the accuracy. The main goal of this component is to generate Mel Frequency Cepstrum Co-efficient and to make normalized energy. This feature will be further proceed to complete voice recognition system. It will be crystal like clear that the generated feature vectors should be uniquely identify for the respective input sound given to it [8].

The input is given to feature extractor, we can also name it as a transducer, because the definition of transducer is that, that change the energy from one form to another form [9].

8.1.3 Recognition Component

This component is further divided into three sub components.

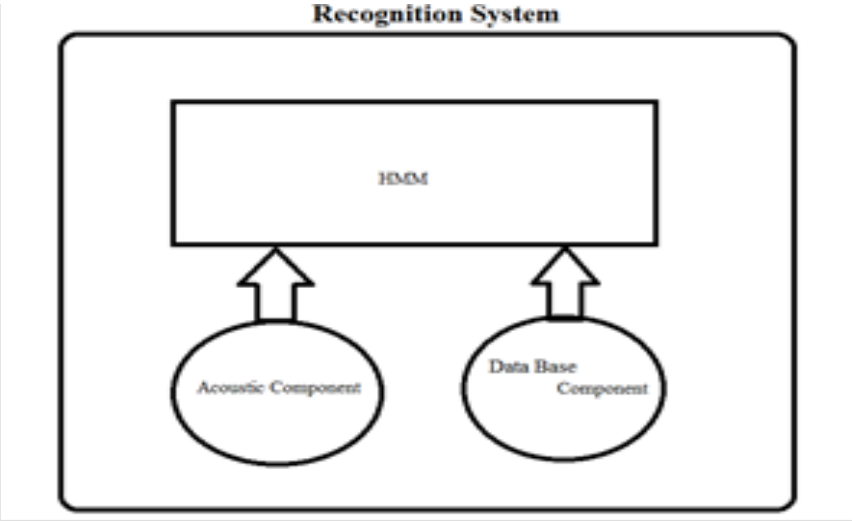

Figure 3: Approaches

The input for this sub system is the output of feature extractor. The feature which were generated before is match with the feature of acoustic model. This acoustic model know the feature of words that are placed in language model. So the language components contain all the words to be recognize and acoustic component have their feature. When the recognition system receive the signal then it is recognize through Hidden Markov Modelling.

8.1.4 HMM

There is statistical method which is widely used for characterizing spectral properties known as Hidden Markov Model. There were two scientist by the name of Baker and Jelinek from Carnegie Mellon University and at IBM respectively, they implement in 1970 for the first time in their speech recognition research [10]. HMM is pattern recognition technique that is very popular during voice recognition system [11]. Basically HMM is the statistical model to present speech pattern.

8.1.5 Language model

Language model representing the model of the language, here we can find two types of sequences of the language, one that are occur in very frequently but the other are used in very rare case. But keep in mind that this language mode are responsible for the grammatical and semantic constraints of that language. The other name of this component is Database Component because this device work as a database for the whole system.

8.1.5.1 Acoustic Component

Acoustic component is component that know that acoustic of words, means that this components know the specific words sound. This component feature are use as reference and compared to the newly introduced sound. Newly introduced sound will be that sound which we are going to be recognized.

In initial while modifying training we boost up acoustic model through the outcome of HMM which is faded by feature extraction component [12]. The acoustic system can recognize each and every feature of the word that are placed in language model of the system.

9. Result

Following result is given by the system of voice recognition when we used HMM for the isolated English words.

9.1.1 Training

Initially we started our system to train it. We also know that our priority is robustness, accuracy etc. so we done that experiment in our laboratory, and we tried our best to reduce the noise effect so that we can get accurate and well result according to our expectation. During performing this experiment we turn off all other electronics component that were or were not concerned with our project. Because I thought that may be their magnetizing field area will affect our outcome.

After training 50 users were introduce to the system, in which 25 were known persons means that we introduced these 25 persons before during the training session of the system and the rest of 25 were newly introduced. We get the following result after our experiment.

correct	incorrect
90 % words that are introduced before once during traning	10 % words that are introduced before once during traning
70 % words that are introduced for the first time to the system	30 % words that are introduced for the first time to the system

Figure 3: Comparison

The attempt were done for 20 times, the voice which was introduced before was recognize up to 90%, which we can consider as 100%, because in 20 times only two times the system was unable to recognize the correct spoken word. But we are worried about those speaker who were introduced to the system for the first time. Their recognition percentage was below to our expectation but I think the accent changed my output so much. For newly introduced voice recognition percentage was only 70 %. While my expectation was 80%.

Here we are taking the average percentage because in our project we tried 20 peoples, so most of the time the

percentage of recognition was different. For some words the accuracy was 100 %, but some words was recognize only 60 %. Because our result are every time varies with the variation of time and also due to the variation (number variation) of speakers. We are just representing our result in average.

9.1.2 Conclusion

In conclusion we are able to say that if we increase the size of vocabulary there will be effect on the output. In this paper I discussed all the aspect that can affect our system during voice to text conversion. Here we are using isolated word base acoustic system but with little modification we can increase the size of vocabulary along with continuous word recognition.

10. Why I am using discrete speech

Here I am using discrete speech because in continuous speech recognition we have to speak without any silence or break which is very difficult and hard job but as compared to this, in discrete sound system the words are isolated by silence [16]. Which is more easy and reliable for us. So we are using discrete sound system in our project [13].

11. Implementation

The whole system of speech recognition is mainly perform two kind of tasks, the first one is to model a signals and the other is known as pattern match [14].

12. Vocabulary

Here we used 500 words and it is consider as average size vocabulary [15].

Acknowledgment

Thanks of my seniors, teachers, parents but especially to Shahbaz khan.

References

[1] Rabiner Lawrence , Juang Biig-hwang, “Fundamental of Speech Recognition”, AT and T,1993.

[2] Anne Johnstone Department of Artificial Intelligence Edinburgh University Hope Park Square Meadow Lane Edinburgh EHB 9LL, (GB) Gerry Attamann “Automated Speech Recognition: A Framework for Research Capital

[3] Dat Tat Tran, “Fuzzy Approach to Speech and Speaker Recognition”, A Thesis submitted for the degree of Doctor of Philosophy of the University of Canberra.

[4] R. K. Moore. Twenty things we still don’t know about speech, Proc. CRIM/FORWISS Workshop on Progress and Prospects of Speech Research in Technology, 1994.

[5] C. H. Lee; F. K. Soong; K. Paliwal “An Overview of Speaker Recognition Technology’, Automatic Speech and Speaker Recognition: Advance Topics. Kluwer Academic Publishers 1996, Norwell, MA.

[6] R. Rodman, "Computer Speech Technology". Artech House, Inc. 1999, Norwood, MA 02062.

[7] M. J Castro; J. C Perez, "Comparison of Geometric, Connectionist and Structural Techniques on a Difficult Isolated Word Recognition task." Proceeding of European Conference on Speech Communication and Technology. ESCA, Vol. 3 pp 1599-1602, Berlin, Germany, 1993.

[8] C. H. Lee; F. K. Soong; K. Paliwal "An Overview of Speaker Recognition Technology', Automatic Speech and Speaker Recognition: Advance Topics. Kluwer Academic Publishers 1996, Norwell, MA.

[9] AN LTCC HYBRID PRESSURE TRANSDUCER FOR HIGH TEMPERATURE APPLICATIONS. Jolymar Gonalez-Esteves (Mechanical Engineering), University of Puerto Rico, Mayaguez Campus NSF Summer Undergraduate Fellowship in Sensor Technologies (SUNFEST).

[10] Rabiner, L. and Juang, B. H. (1986), "An International to Hidden Markov Models", IEEE ASSP Magazine, Vol. 3, No.1, part 1, pp. 4-16.

[11] Atal, Bishnu S. and Rabiner, Lawrence R. (1976), "A Pattern Recognition Approach to Voiced_Unvoiced Classification with Application to Speech Recognition", in Proceedings of the IEEE international conference on Acoustic, Speech and Signal Processing (ICASSP'76), Pennsylvania, Vol. 24, No.3, pp. 201-212.

[12] "Speech Recognition for Hindi Language",. C-DAC India.

[13] Reddy D. R and Ermann, L. D_1975. "Tutorial on System Organization for Speech Understanding." In D. R Reddy (ed) Speech Recognition, Academic Press.

[14] Picone J. W.,"Signal Modelling Technique in Speech Recognition". Proc of the IEEE Vol 81, No. 9, pp. 1215-1247, 1993.

[15] Reference "Pattern matching for a large vocabulary Speech Recognition System".

[16] Rumelhart, D.E. and McClelland, J.L. 1982. "An Interactive Activation Model of Context Effects in Letter Perception: Part II. The Contextual Enhancement Effect. Some Tests and Extensions of the Model. In Psychological Review".