International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

A Review On Different Techniques for Biometrics based Speech Recognition

Kirandeep Kaur[1], Dr. Vijay Bhardwaj[2]

[12]Department of Computer Application
GURU KASHI UNIVERSITY TALWANDI SABO (BATHINDA), (INDIA)

ABSTRACT

Speech Recognition makes its place among most discussed techniques of biometrics. The major challenge in the field of speech recognition is that it can be altered by dialects, accents and mannerisms. The most natural form of human communication is speech and its processing has been one of the most exciting areas of signal processing. Because of advancements in speech recognition technology, today it has been made possible that computer understands human voice commands and human languages. The primary goal of speech recognition is to develop a system for speech input to machine. Speech recognition has also proved boon for people with disabilities who can't do typing. If a user has lost the utilization of his/her hands or for visually impaired users where it is not convenient to use Braille keyboard, the system permits personal expression through dictation and by other computer tasks.

Key Terms: *Mel, Speech, MFCC*

I. INTRODUCTION

The process of converting spoken words into text is known as speech recognition. Speech Recognition makes its place among most discussed techniques of biometrics. The major challenge in the field of speech recognition is that it can be altered by dialects, accents and mannerisms. The most natural form of human communication is speech and its processing has been one of the most exciting areas of signal processing. Because of advancements in speech recognition technology, today it has been made possible that computer understands human voice commands and human languages. The primary goal of speech recognition is to develop a system for speech input to machine. Technically stated, speech recognition is the ability of a machine or a program to identify words and phrases in spoken language and convert them into machine-readable format [5].

When we make calls in big companies, it is not a person who answers the phone, instead it is an automated voice recordings that answers and instructs people to press buttons to move through option menus. Technology has even advanced further today. There is no need to press buttons, user can just speak some words as instructed by a recording to fulfill the requirement. All this is a kind of speech recognition program and comes under automated phone system. These programs fall into two categories [6].

- Small-vocabulary/many-users

These systems are preferred for automated telephone answering. In this category, usage is restricted to small number of predetermined commands and inputs, for instance, basic menu options or numbers. There is always a possibility that users can speak with great deal of variation in accent and speech patterns and system will understand them most of the time.

- Large-vocabulary/limited-users

This system is preferred in environment with limited users. These systems are trained to perform best with small number of primary users. The percentage of accuracy in this case is above 85 percent. The vocabulary of such system s in tens of thousands. If a person other than primary user attempts to use the system, the accuracy rate can fall drastically as system is not trained for working with the voice of an outsider.

Speech recognition has also proved boon for people with disabilities who can't do typing. If a user has lost the utilization of his/her hands or for visually impaired users where it is not convenient to use Braille keyboard, the system permits personal expression through dictation and by other computer tasks.

Speech recognition systems comes under the choice between discrete and continuous speech. It has been observed that when the words are spoken separately with a pause between each one, are understood by the system more effectively as compared with the continuous speech.

A computer system has to go through several complicated steps in converting speech to on-screen text. When humans speak, we create vibrations in the air. The analog-to-digital converter (ADC) translates this analog wave into digital data. This is done by taking samples and digitizing the sound by taking accurate measurements of the wave at frequent intervals. The system filters the digitized sound in order to eradicate unwanted noise and often to separate it into different bands of frequency. It also normalizes the sound and adjusts it to a constant volume level. Sound also needs to be temporarily aligned. People do not always speak at same speed. So, there is a need to adjust the sound to match the speed of the template sound samples stored in system's memory [1, 2].Figure 1 shows the detailed functioning of speech recognition system.

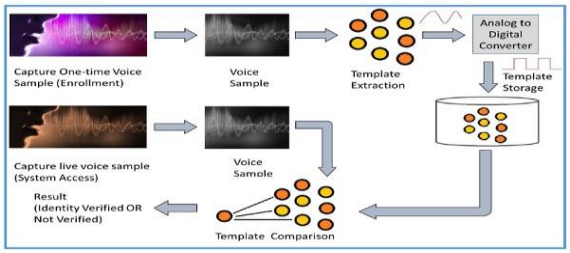

Figure 1. Working of Speech recognition system

Speech recognition deals with speech variability and account for learning relationship between corresponding word and specific utterance. There have been two popular trends in the field of speech recognition over the recent years. First one is academic approach and second is pragmatic which makes use of technology and offers simple low-level interaction with machine, replacing switches and buttons. While the former approach has always made promises for future, the second approach is in already in use [7].

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Speech Recognition Techniques

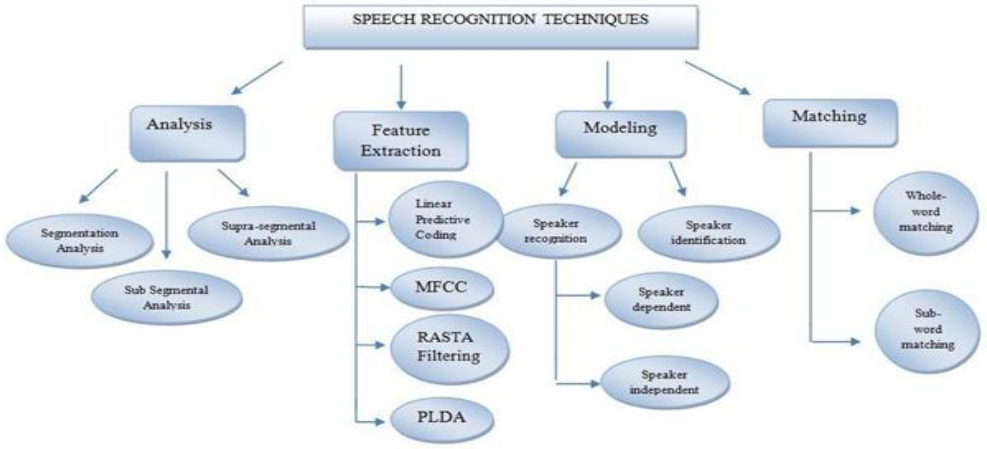

Figure 2: Speech Recognition Technique [1]

Approached involved in Speech Recognition System

There are basically two different approaches to speech recognition. These are described as under [3, 4].

1) Acoustic-Phonetic Approach

In this system tries to decode the speech signal in a sequential manner based on relations between phonetic symbols and acoustic features of the speech waveform. The steps involved in this approach are mentioned as under.

- In first step an appropriate spectral representation of the speech signal is provided and is classified as parameter measurement process.
- The next is feature detection stage where the spectral measurements are converted to a set of features describing acoustic properties of the various phonetic units.
- In last step, the recognizer makes an attempt to determine the best matching word or sequence of words.

2) *Pattern Recognition Approach*

In this speech patterns are used directly without any feature determination and segmentation. This approach works in two steps, training of speech patterns and recognition of patterns. A sequence of measurements is made on the input signal to define the test pattern. The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern and reference pattern is computed. Finally, the decision rule decides which reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase.

Advantages

- First benefits of this strategy are that degradation of the possibility of copying security passwords because there is no need of composing security passwords and the whole can be done without any worry.
- The significant advantage of this program is in contact centers where a huge number of clients are on line to enquire and a representative need to be online to be present at the call. With the help of this technological innovation calling can be joined successfully and with more efficiency.
- Person who is unable to write or see with the help of this application can perform their task such as inquiring or transaction process etc.
- Country like India has so many dialects variation with the help of this technology the dependency of human staff trained in different languages has been reduced significantly.
- This application has proved a revolution to improve customer happiness in addition to improving companies' earnings by simply achieving new customers in addition to holding onto existing customers.
- This is very beneficial for who are able to read and write up to some extent and know how to make use of cell but can't write or speak in English to type English letters as passwords.
- Speech is a very natural way to interact, and it is not necessary to sit at a keyboard or work with a remote control.
- No training required for users!

Disadvantages

- Even the best speech recognition systems sometimes make errors. If there is noise or some other sound in the room (e.g. the television or a kettle boiling), the number of errors will increase.
- Speech Recognition works best if the microphone is close to the user (e.g. in a phone, or if the user is wearing a microphone). More distant microphones (e.g. on a table or wall) will tend to increase the number of errors.
- Sometimes users encounter the denial to access account that is really their own this is known as "false negative" but this problem is very rare and has less than 3% of possibility.
- If a user registered with a certain model of mobile or phone, attempted to verify his identity with a different model, faces troubles in authentication. Development of an adaptation algorithm is going on that will adapt the voiceprint to various handsets automatically, hence will increase the quality of verification even more.
- To use voice recognition system, you have to speak loudly than your normal voice it may have the possibility of vocal cord injury but there is no scientific proof has been presented between the voice recognition and damage to the vocal cord.

II. LITERATURE SURVEY

Joost Van Doremalen et. al(2016), stated that Computer-Assisted Language Learning (CALL) applications for improving the oral skills of low-proficient learners have to cope with non-native speech that is particularly challenging. Since unconstrained non-native ASR is still problematic, a possible solution is to elicit constrained responses from the learners. In this paper, we describe experiments aimed at selecting utterances from lists of responses. The first experiment on utterance selection indicates that the decoding process can be improved by optimizing the language model and the acoustic models, thus reducing the utterance error rate from 29–26% to 10–8%. Since giving feedback on incorrectly recognized utterances is confusing, we verify the correctness of the utterance before providing feedback. The results of the second experiment on utterance verification indicate that combining duration-related features with a likelihood ratio (LR) yield an equal error rate (EER) of 10.3%, which is significantly better than the EER for the other measures in isolation

Youssef Zouhir et. al.(2015) explained a feature extraction method for robust speech recognition in noisy environments is proposed. The proposed method is motivated by a biologically inspired auditory model which simulates the outer/middle ear filtering by a low-pass filter and the spectral behavior of the cochlea by the Gammachirp auditory filterbank (GcFB). The speech recognition performance is tested on speech signals corrupted by real-world noises. The evaluation results show that the proposed method gives better recognition rates compared to the classic techniques such as Perceptual Linear Prediction (PLP), Linear Predictive Coding (LPC), Linear Prediction Cepstral coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC). The used recognition system is based on the Hidden Markov Models with continuous Gaussian Mixture densities (HMM-GM).

Woon S. Gan et. al., (2015), stated future audio, speech, and music processing applications need innovative intelligent algorithms that allow interactive human/environmental interfaces with surrounding devices/systems in real-world settings to control, process, render, and playback/project sound signals for different platforms under a diverse range of listening environments. These intelligent audio, speech, and music processing applications create an environment that is sensitive, adaptive, and responsive to the presence of users. Three areas of research are considered in this special issue: analysis, communication, and interaction. Analysis covered both preprocessing of sound signals and extraction of information from the environment. Communication covered the transmission path/network, coding techniques, and conversion between spatial audio formats. The final area involved intelligent interaction with the audio/speech/music environment based on the users' location, signal information, and acoustical environment.

Miss Himanshu et. al.(2015), stated speech Recognition is very important and popular. Speech recognition is the process of converting spoken words into text. One of the problems faced in speech recognition is that the spoken word can be vastly altered by accents, dialects and mannerisms. In case of speech recognition, the research followers are mainly using three different approaches namely Acoustic phonetic approach, Pattern recognition approach and Artificial intelligence approach. The objective of this review paper is to summarize

and compare some of the well-known methods used in various stages of speech recognition system and identify research topics.

Preeti Saini,Parneet Kaur, (2014) stated that after years of research and development the accuracy of automatic speech recognition (ASR) remains one of the most important research challenges e.g. speaker and language variability, vocabulary size and domain, noise. The design of speech recognition system requires careful attentions to the challenges or issue such as various types of speech classes, speech representation, feature extraction techniques, database and performance evaluation. This paper presents a study of basic approaches to speech recognition and their results shows better accuracy. This paper also presents what research has been done around for dealing with the problem of ASR.

III. COMPARITIVE ANALYSIS

a. Linear Predictive coding: LPC is a tool which is used for speech processing. LPC is based on an assumption: In a series of speech samples, we can make a prediction of the n^{th} sample which can be represented by summing up the target signal's previous samples (k). The production of an inverse filter should be done so that is corresponds to the formant regions of the speech samples. Thus the application of these filters into the samples is the LPC process.[7].

b. Mel-frequency cepstrum (MFFCs): Mel Frequency Cepstral Coefficients are based on the known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. The main purpose of the MFCC processor is to copy the behaviour of human ears.

c. RASTA filtering: RASTA is short for RelAtive SpecTrAl. It is a technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the representations of the speech signals are band pass filtered in RASTA. Initially, it was just used to lessen the impact of noise in speech signal but know it is also used to directly enhance the signal [13].

d. **Modeling Techniques**: The goal of the modeling techniques is to produce speaker models by making use of the features extracted (feature vector). As shown in the figure the modeling techniques are further categorized into speaker recognition & identification. Speaker recognition can be further classified into speaker dependent and speaker independent. Speaker identification is a process in which the system is able to identify who the speaker is on the basis of the extracted information from the speech signal. In speech recognition process we can use the following modeling approaches:

Acoustic-Phonetic approach: The basic principle that this approach follows is identifying the speech signals and then providing these speech signals with apt labels to these signals. Thus the acoustic phonetic approach postulates that there exists finite number of phonemes of a language which can be commonly described by acoustic properties.

Pattern recognition approach: It involves two steps: Pattern Comparison and Pattern Training. It is further classified into Template Based and Stochastic approach. This approach makes use of robust mathematical formulas and develops speech pattern representations.

Dynamic Time Warping (DTW): DTW is an algorithm which measures whether two of the sequences are similar that vary in time or even in speed. A good ASR system should be able to handle the different speeds of different speakers and the DTW algorithm helps with that. It helps in finding similarities in two given data keeping in mind the various constraints involved.

Artificial Intelligence Approach (AI): In this approach, the procedure of recognition is developed in the same way as a person thinks, evaluates (or analyzes) and thereafter makes a decision on the basis of uniform acoustic features. This approach is the combination of acoustic phonetic approach and pattern approach. [1]

e. **Matching Techniques**: The word that has been detected is used by the engine of speech recognizer to a word that is already known by making use of one of the following techniques:

Sub word matching: Phonemes are looked up by the search engine on which the system later performs pattern recognition. These phonemes are the sub words thus the name sub word matching. The storage that is required by this technique is in the range 5 to 20 bytes per word which is much less in comparison to whole word matching but it takes a large amount of processing.

Whole word matching: In this matching technique there exists a pre-recorded template of a particular word according to which the search engine matches the input signal. The processing that this technique takes is less in comparison to sub word matching. A disadvantage that this technique has is that we need to record each and every word that is to be recognized beforehand in order for the system to recognize it and thus it can only be used when we know the vocabulary of recognition beforehand. Also these templates need storage that ranges from 50 bytes to 512 bytes per word which very large as compared to sub word matching technique.

IV. CONCLUSION

There has been a lot of research in the field of speech recognition but still the speech recognition systems till date are not a hundred percent accurate. The systems developed so far have limitations: there are a limited number of vocabularies in the current systems and we need to work towards expanding this vocabulary, there exists a problem of overlapping speech that is the systems cannot identify speech from multiple users, the user needs to be in a place which is background noise free for an accurate recognition, there occurs a problem with the accent and the pronunciation of the user or speaker. In the future the speech recognition systems need to be free of these limitations to give hundred percent results. In this paper we firstly attempt to show the major systems developed under speech recognition over the years. We then give a brief description of speech recognition techniques. A speech recognition system should include the four stages: Analysis, Feature Extraction, Modelling and Matching techniques as described in the paper. Also, through this paper we show four techniques used in feature extraction: Linear Predictive Coding, Mel-frequency cepstrum, Relative Spectral and Probabilistic Linear Discriminate Analysis. By studying each of these techniques we conclude that they have their own advantages and disadvantages and all of them are being used for different purposes. Through research we conclude the Mel frequency cepstrum is a feature extraction technique that is used widely for many speech recognition systems as it is able to mimic the human auditory system and it gives a better performance rate.

REFERENCES

[1] Joost Van Doremalen et. al., "Optimizing Automatic speech recognition for Low-Proficient Non-Native Speakers", EURASIP Journal on audio,speechamd music processing(2016).

[2] Youssef Zouhir et. al., "A bio-inspired feature extraction for robust speech recognition", SpringerPlus, 2014, 3:651(2015).

[3] Woon S. Gan et. al., "Intelligentaudio,speech, and music processing applications", EURASIP Journal of audio, speech and music processing, (2015).

[4] Miss Himanshu et. al., "Literature survey on automatic speech recognition system", IJARCSSE, Volume 4, Issue 7, July 2014, .

[5] Preeti Saini, Parneet Kaur, "Automatic speech recognition: A review", IJETT, Volume 4, Issue 2, 2013, ISSN: 2231-5381.

[6] S. B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, (1980), pp. 357–366.

[7] R. Lawrence and B.-H. Juang ―Fundamentals of Speech Recognition‖ , Prentice-Hall, Inc., (Engelwood, NJ), (1993).

[8] M. A. Anusuya and S. K. Katti, ―Speech Recognition by Machine:A Review‖ , International Journal of Computer Science and Information Security, vol. 6, no. 3, (2009), pp. 181 -205.

[9] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, no. 1, (1978) pp.43–49.

[10] J. K. Baker, "The Dragon System-An Overview',' IEEE Trans. on Acoustics Speech Signal Processing, Vol. ASSP-23, no. 1, (1975), pp. 24-9.

[11] J. Bilmes, ― hat HMMs can do,‖ IEICE Trans. Inf. Syst., vol. E89-D,no. 3,(2006), pp. 869–891.

[12] S. Renals, N. Morgan, H. Boulard, M. Cohen and H. Franco, ―Connectionist probability estimators in HMM speech recognition‖ , IEEE Trans. Speech Audio Processing , vol. 2, no. 1, (1994), pp. 161–174.

[13] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki,M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard and M. Athineos, ―Pushing the envelope—Aside [speech recognition]‖ , IEEE Signal Process. Mag., vol.22, no. 5, (2005), pp. 81–88.

[14] H. A. Bourlard and N. Morgan, ―Connectionist Speech Recognition- A Hybrid Approach", kulwer Academic Publishers, (1994).

[15] N. Smith and M. J. F. Gales, "Using SVM's and discriminative models for speech recognition", Proc. ICASSP, vol. 1, (2002), pp.77 -80.