iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** X **Month of publication:** October 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Survey on Speech Recognition Technique

Mr. Shetty Bhavesh.R[1], Mr.Kurapati Shyam.S.[2,] Mr.Kurapati Ram.S.[3], Prof.Joshi Parag.S.[4]
Department Of Computer Engineering RMCET Ambav, Devrukh. Mumbai University

Abstract: This paper provides a survey on feature extraction for speech recognition and discusses the techniques and system that make it possible for computers to accept speech as input. This paper shows the major developments in the field of speech recognition. This paper draw special attention to the speech recognition techniques and provides a brief description about the four stages where the speech recognition techniques are classified. In addition, this paper gives a detailed information of four feature extraction techniques: Linear Predictive Coding (LPC), Mel-frequency cepstrum coefficient (MFCCs), RASTA filtering and Probabilistic Linear Discriminate Analysis (PLDA). The objective of this paper is to summarize the feature extractions techniques used in speech recognition system.
Keywords: Automatic Speech Recognition (ASR), Feature Extraction, LPC, MFCCs, RASTA filtering, PLDA

I. INTRODUCTION

Automatic Speech Recognition (ASR) also known as computer speech recognition is a process in which speech signal is converted into a sequence of words, other linguistic units by making use of an algorithm which is implemented as a computer program. The major objective with which ASR works is the development of the techniques and a system that enables the computers to recognize speech as input. In a speech recognition system we convert speech into text in which the text is the output of the speech recognition system which is equivalent to the recognized speech. Speech recognition applications have evolved over the past few years. These applications include voice search, call routing command and control, appliance control by voice, voice dialling, computer aided language learning, robotics and many more. The modern speech recognition systems are based on the HMMs that are the Hidden Markov Models. The main reason why HMM is widely used is that HMM has parameters that can be automatically learned or trained and the techniques that are used for learning are easy and are computationally feasible to use. Many advances have been made is the automatic recognition of speech by machine but we are still unable to develop a machine that understands the human speech by numerous speakers in any kind of environment.

II. SPEECH RECOGNITION TECHIQUES

The main objective of a speech recognition system is to have capacity to listen, understand and then after act on the spoken information. A speech recognition system includes four main stages.

A. Analysis

The first stage of speech recognition is analysis. When the speaker speaks, the speech includes different types of information that help to identify a speaker. The information is different because of the vocal tract, the source of excitation in addition to behavior feature. As shown in the figure above, the speech analysis stage can be further classified into three analysis: a. Segmentation Analysis: In segmentation analysis, the testing to extort the information of speaker is done by utilizing the frame size in addition to shift which is in between 10 to 30 milliseconds (ms) [Range]. b. Sub-segmental Analysis: In this analysis technique, the testing to extract the information of speaker is done by utilizing the frame size in addition to shift which is in between 3 to 5 milliseconds (ms) [Range]. The features of excitation state are examined and extracted by using this technique. c. Supra-segmental Analysis: In Supra-segmental Analysis, the analysis to extract the behavior features of the speaker is done by utilizing the frame size as well as the shift size that varies in between 50 to 200 milliseconds.

B. Feature Extraction Technique

Feature extraction is the main process of the speech recognition system. It is considered as the heart of the system. The work of feature extraction is to extract those features from the input speech (signal) that help the system in identifying the speaker. Feature extraction squeezes the magnitude of the input signal (vector) without causing any harm to the power of speech signal.

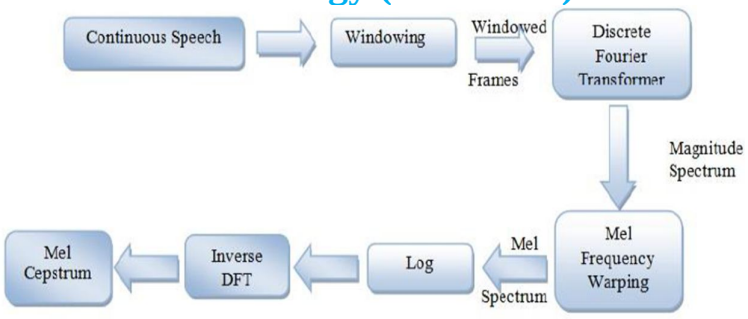


Figure 2: Feature Extraction Diagram

The figure shown above is the feature extraction diagram. In feature extraction, from one side we input the continuous speech signals which is for the process of windowing. In the process of windowing the disturbance which are present at the start and end of the frame which are minimized. After this process, the continuous speech signal is converted into windowed frames. Then windowed frames are passed into the discrete Fourier transformer which converts the windowed frames into magnitude spectrum. Now into next step, spectral analysis is done with a fixed resolution along a subjective frequency scale that is the Mel-frequency scale which makes a Mel-spectrum. This spectrum is then passed to Log and then to inverse of discrete Fourier transform which makes the final result as Mel-Cepstrum. The Mel-Cepstrum consists of the features that are required for identifying the speaker. A few feature extraction techniques include: a. Linear Predictive coding: LPC is a tool which is used for processing of speech. LPC is based on an assumption: In a series of speech samples, we can make a prediction of the nth sample which can be represented by summing up the target signal's previous samples (k). The production of an inverse filter should complete so that is corresponds to the formant regions of the speech samples. Thus the application of these filters and the samples is the LPC process.

Mel-frequency cepstrum coefficients (MFFCs): Mel Frequency Cepstral Coefficients are based onto known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. The main objective of the MFCC processor is to copy the behavior of human ears. The derivation of MFCCs is done by the following steps.

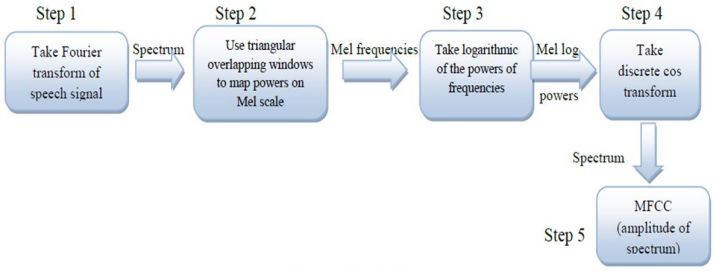


Figure 3: MFCCs Derivation

MFCC is most commonly used for feature extraction at front-ends in speech recognition systems. The technique is FFT based (Fast Fourier Transform), which means that feature vectors are extracted from the frequency spectra of the speech frames. The Mel scale, a non-linear frequency scale is used to make triangular bandpass filters and a series of such filters is called Mel frequency filter bank. The equation given below describes the mathematical relationship between the linear frequency scale and the Mel scale,

Freq Mel = 2595.0 * (Math.log (1.0 + freq / 700.0) /Math.log (10.0))

Where freq Mel is the Mel frequency in Mels and freq is the linear frequency in Hz

A. RASTA Filtering: Relative Spectral is a long form of RASTA. It is a technique which is used to enhance the speech when registered 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 diminish the impact of noise in speech signal but know it is also used to directly enhance the signal. The following figure shows the process of RASTA technique. The main thought here is to subdue the constant factors.

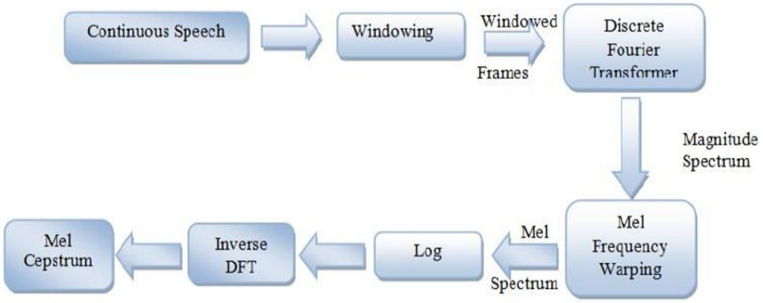


Figure 2: Feature Extraction Diagram

2) Probabilistic Linear Discriminate Analysis (PLDA): This technique is an enlargement for linear probabilistic analysis (LDA). Initially this technique was used for face recognition but now it is used for speech recognition. The following table briefly describes this technique.

C. 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:

1) 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.

2) 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.

3) 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.

4) 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.

D. 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:

1) 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.

2) 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.

III. CONCLUSION AND FUTURE SCOPE

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, Modeling 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] Santosh K.Gaikwad and Pravin Yannawar, A Review, International Journal of Computer Applications ***A*** Review on Speech Recognition Technique Volume 10– No.3, November 2010 [2] Rybach, D.; C. Gollan; G. Heigold; B. Hoffmeister; J. Lööf; R. Schlüter; H. Ney (September 2009). "The RWTH Aachen University Open Source Speech Recognition System". Interspeech-2009: 2111–2114.

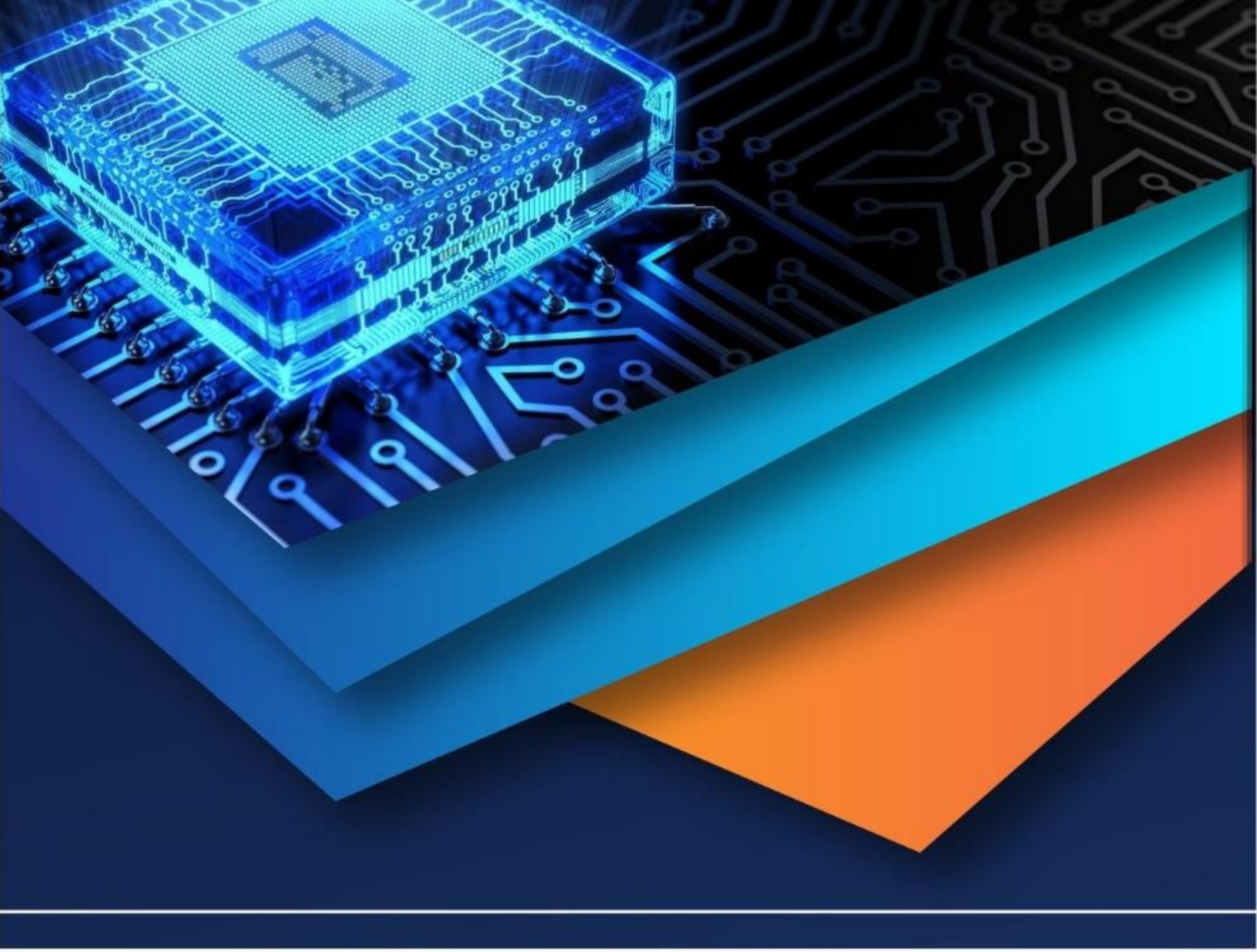
[2] Sanjivani S. Bhabad Gajanan K. Kharate International Journal of Advanced Research in Computer Science and Software Engineering **,** An Overview of Technical Progress in Speech Recognition Volume 3, Issue 3, March 2013

[3] Wiqas Ghai and Navdeep Singh International Journal of Computer Applications (0975 – 8887) a Literature Review on Automatic Speech Recognition, Volume 41– No.8, March 2012.

[4] Melanie Pinola (2011-11-02). "Speech Recognition Through the Decades: How We Ended Up With Siri". www.techhive.com.

[5] Celso Auguiar, in CCRMA - Center for Computer Research in Music and Acoustics. Stanford University on Modeling the Excitation Function to Improve Quality in LPC's Resynthesis**.**

[6] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication 54 (4): 543–565. doi:10.1016/j.specom.2011.11.004. (CSLT), on I-vectors, a Discriminative Scoring for Speaker Recognition Based, 2014



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY