INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Analysis of speech recognition techniques

Pranit Gadekar
pranit.gadekar1997@gmail.com
Ramrao Adik Institute of Technology, Mumbai, Maharashtra

Mohmmad Hilal Kaldane
hilalkaldane.kaldane937@gmail.com
Ramrao Adik Institute of Technology, Mumbai, Maharashtra

Dipesh Pawar
dipeshpawar6231@gmail.com
Ramrao Adik Institute of Technology, Mumbai, Maharashtra

Omkar Jadhav
omkar.jadhav2601@gmail.com
Ramrao Adik Institute of Technology, Mumbai, Maharashtra

Anita Patil
anita.patil@rait.ac.in
Ramrao Adik Institute of Technology, Mumbai, Maharashtra

ABSTRACT

This paper focuses on speech recognition techniques such as LPC (linear predictive coding), MFCC (Mel-frequency Cepstral coefficients) with Hidden Markov Models, LPCC (linear predictive Cepstral coding), and RASTA and will compare these techniques to find a most accurate and efficient way to recognize speech. Speech recognition is the process in which program or machine do the identification of words or phrases and convert them to machine-readable format. Additionally, this paper also focuses on NLP (natural language processing) techniques used with the speech recognition process. Once the speech signal is converted to text then NLP is used to understand and generate what has been said. NLU (natural language understanding) and NLG (natural language generation) are two important steps in NLP, through this paper, we will compare and analysis techniques to find out which we can use with speech recognition for effective results. The Objective of this paper is to find out the best technique which is currently used.

Keywords—*LPC, LPCC, MFCC, Hidden Markov model, NLP, NLU, NLG*

1. INTRODUCTION

Speech signalling and processing is now used in many applications, whether it may be Google speech recognizer, intelligent robots, call center automation etc. speech recognition can be done in two ways automatic way speech recognizing speech signal and text-speech synthesis, this process includes acquiring, manipulating, storing, transferring, and display of speech signals. The speech recognition techniques play a crucial role in output these signals, the first one is LPC which is a digitized method for encoding the analogous signal. This method uses linearity function for predicting determined value from previous gained value. For retrieving feature from the speech signal and for speech analysis and resynthesis LPC is used. LPC involves following phases such as pre-emphasis, frame block where output of the previous phase is divided into blocks, windowing to reduce the Interruption at the begin and last of every frame/block for every individual frame, Autocorrelation analysis to used find features, LPC analysis phase to convert each frame to set of parameter values, LPC parameter to find the cepstral coefficient of cepstral. The disadvantage of the LPC method is that it cannot capture inexpressible and prolong speech signals in an accurate way, the execution is poor where there is noise. The next technique for speech recognition is LPCC (linear predictive coefficient coding), this method assumes that the identity of a speech signal generated is directed by the shape of the articulate area. The goal of feature/attribute retrieval is to output speech signals through discrete measures of signals. Through LPC, extraction of the LPCC's coefficients is done, these are furthermore translated to cepstral signals. To represent speech signal parameters like pitch duration, the intensity of voice, quality and format of the signal, LPCC is used widely. The LPCC is comparatively more reliable as well as with respect to the LPC technique, it gives better performance and is simple to implement. The drawback here is that its output of the signal is influenced due to quantization noise and if improper order is employed the accuracy of this technique degrades [1].

The third technique used for Speech recognition is RASTA (Relative spectral filtering) this method was initially developed for the purpose of reducing the undesirable and incremental noise which gets further added

In the speech recognition process. RASTA method weakens the effect of noise in the vocal signal as well this method increases the value of the voice signal with respect to surrounding noisy elements. It also filters the trajectory of the potential area in case of noisy data. RASTA algorithm is basically a modulated bandpass filter (MBPF) method which is used either in cepstral's domain or in the logarithmic frequency domain. For better results RASTA technique is combined with PLP (perceptual linear prediction), PLP works similarly with the LPC expect the spectral attributes are converted here to match

the traits of human acoustic system. The PLP method is more adaptive to human hearing compared to LPC technique [2].

The Last technique we will be discussing here is MFCC, this technique is the most standard way to extract features out of the speech. This technique is based on dissimilarity of person ear's analytical level bandwidth. Here the coefficients are derived from the cepstral representation of the audio clip. The difference between Cepstrum's based and Mel-level-frequency coefficients is that in case of the MFCC, the frequency bands are equally divided on to the Mel-spectrum-scale. The main aim of MFFC is to mimic the human auditory system. The process of MFCC:

- Find out the Fast Fourier Transform (FFT) of the audio signal.
- Mapping the power of the spectrum-scale obtained o to Mel-frequency-scale using windowing.
- Calculate the log of power for relative Mel-level-frequencies.
- Find discrete cosine transform (DCT) of the listing of Mel-log -scale powers.
- Keep DCT coefficients in the range of 2-13 discard others
- The MFCC are amplitudes of the resulting signal spectrum.

MFCC techniques are therefore used music retrieval management systems such as to classify the genre and to measure similarities between audio signals. It is robust and cepstral based method that matches the human auditory system and therefore widely recognized and deployed.

Once we have selected which speech recognition technique to use, the next step is to understand what has been said, known as NLP. The purpose of NLP is to make communication between the computer and people feel exactly like interaction in between persons. There are two components of NLP first one is Natural language Understanding (NLU) and second is Natural language Generation (NLG). Natural language Understanding takes information in the textual form and tries to fetch the meaning of the textual content. Mostly today 's speech analyzing systems uses systems today are based on Hidden Markov Models (HMMs). These are based on statistic models based on mathematical calculations, this model break your speech into smallest units then compare with prerecorded phonemes and through statistics determine the most likely words you were speaking and outputs this information in form of text. Words must be understood in the form of whether it is a noun, verb their tenses called Parts-of-Speech tagger (POS). The next component is a generation of languages, NLG does the translation of computer artificial language into textual form and also goes a step further by translating that text into audible speech with text to speech then it organizes the structure of how it's going to say it. Using a lexicon system and a set of grammatical rules, an NLG system can generate complete words or lines.

2. COMPARATIVE STUDY

LPC technique works well when it comes to estimating speech parameters in a precise manner, while LPCC has good reliability and robustness due to the conversion of coefficients. LPC works well at high frequencies but the difficulty here is that it cannot differentiate words which have the same vowels sounds. The next technique is RASTA which works well in environmental noise and disturbance conditions this technique is reliable and robust when working and capturing frequencies with low modulation, but for better performance, we have combined with technique with PLP, which increases the complexity and implementation cost of this method. While in case of MFCC accuracy and performance is good and it works with the human auditory system, background noise can hamper the quality of signal but if we pass through linear filtering MFFC works best, providing better efficiency than LPC and RASTA making it more popular and useful among all.

3. METHODOLOGY

There are various methods for efficient extraction of speech parameters for voice recognition, but still, the MFCC method with enhanced recognition methods such as HMMs is most widely used. The approaches have been developed from the human auditory system, spectrogram study to simple template matching to dynamic time warping to more modern approaches such as neural networks and Hidden Markov Models. The methodology here depicts the use of MFCC for recorded speech as speech recognition parameter and HMM as a recognition parameter. The four steps are [3]:

- **Speech:** this is the initial stage which takes speech samples, in the form of waveforms and these samples are forwarded for pre-processing.
- **Pre-processing:** here the speech samples are filtered and converted to an appropriate format that can be recognized by the computer. Preprocessing deals with issues such as sampling and windowing and frame formation.
- **Feature extraction:** Eloquent features are extracted sampled and the windowed speech signal to enable classification of sounds. The first step is to represent the signal into frequency domain using a discrete Fourier transform. The second step is to generate a Mel-frequency signal with different bandpass filters. The next process is to take logarithmic values to perceived loudness on its scale, used to mimic human perception of loudness. The fourth step is calculating cepstral coefficients for eliminating speaker-dependent characteristics by suppressing of the source signal.
- **Speech classification:** here the Hidden Markov model plays an important role, in HMMs states are not visible but entering to particular state follows a probability distribution. The observation provides information regarding the sequence of states, HMMs uses estimation method, then these parameters are passed to under three states of HMMs for training.

A flowchart depicting the process of MFCC with HMMs (figure 1)

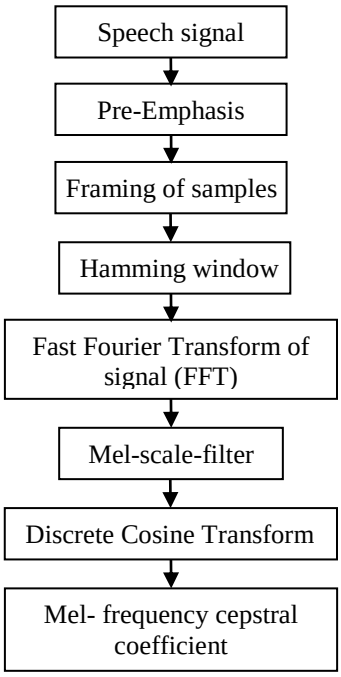

Fig. 1: Flowchart depicting the process of MFCC

4. THE ACCURACY OF TECHNIQUES

The following shows the accuracy table [2]. The accuracy of each technique with respect to language. The recognition rate of each technique is different for every language. For instance, the MFCC algorithm works best with the English language and we have a good recognition rate but with Urdu, the rate is not so much high. So the technique should be used with the language to be used within the system and considering overall factors with respect to their accuracy rate, precision values and their merits and demerits. RASTA technique is usually used with PLP method so that it matches with the human auditory system and MFCC uses cepstrum based domain signals to match with the human auditory system depending upon the coefficients to be used.

Table 1: Technique

Technique	Language in which recognition rate to be measured	Recognition rate
LPC	English	91.40%
LPC	Devnagari	82.3%
RASTA	English	94.27%
RASTA	Spanish	93.9%
MFCC	English	99.9%
MFCC	Urdu	86.67%

5. ANALYSIS OF TECHNIQUES

Hidden Markov Models (HMM's) are used in huge vocabulary speech recognizer system which is trained in an automatic manner after some hours of training on large speech the benefit of this model is that it decreases complexity and time of training but it is difficult to examine errors. In neural network modelling approach can control low quality and noisy data but is not the optimal way. Another technique called dynamic time wrapping calculates the similarities between two series that differ in time or speed. Vector quantization maps the voice samples from large space to some definite number of regions each region has some code words which serves as prerecorded words for speaker and used when the speaker is tested in the system.

Pattern matching is most common natural language processing method where some predefined patterns are found out and then the input keywords are mapped to this predefined pattern to match the input words, the number of patterns can be reduced by matching with semantic primitives rather than words. Another technique called syntactically driving parsing in which syntax is the words that fit with each other and form high-level units such as phrases, sentences and clauses. Here a group of words are matched. Semantic grammar technique is similar to syntactic parsing but here the rules of grammar are applied to find out the logistic meaning. And frame-based technique combines all the key elements to increase performance.

The following table shows the comparison chart for LPC, LPCC, MFCC and RASTA filtering with respect to their features their limitations and benefits. It also shows the circumstances in which the techniques can be used so that system performance and capacity is increased.

Table 2: Analysis of LPC, LPCC, RASTA and MFCC [2]

Technique Type	Characteristic	Advantage	Disadvantage
LPC	This is a useful and widely implemented approach for high level frequencies.	This technique is authentic, robust and provides reliability and useful for voice tracts.	Cannot differentiate words with the same vowel sounds
LPCC	LPCC uses conversion of coefficients for recognition	It is more reliable, simple and delivers good performance	LPCC is highly susceptible to the quantizer noise and requires the use of proper ordering.
RASTA	A widely popular method that has environmental noise and disturbance	This technique provides good reliability and robustness to low-frequency modulations.	RASTA needs to be integrated with PLP for better results which increases complexity.
MFCC	Cepstral domain based that matches the human auditory system	This technique is highly efficient and gives good accuracy with an audio signal with low complexities.	Background noise affects quality but with simple filters, this can be overcome.

6. METHODOLOGY TO BE PROPOSED

Smart audio coding will be an artificially created entity that will be recognizing the commands from the developer via a medium as voice and converting them to syntactical code. This system will be running on developer's voice, hence it will be a complete hands-free task to code.

The First step is data recording and preprocessing of the data here involves normalization of data to remove redundant and inconsistent information then comes in feature extraction this involves training the Hidden Markov Models they are based on three problem evaluation problem, decoding and learning problem we implement algorithms like Baum-Welch to solve problems based on probability. Understanding of grammar is important the nouns, actions, verb, sentence tense etc. language and acoustic models are used in the recognition phase where they two complement each other. Language model defines the way search is performed. Apart from word sequence probabilities, language model basically provides a graph on which search is run to arrive at a hypothesized word sequence and acoustic models are used to decide which paths in this graph are viable or not.

In the parsing phase of the process to analyze and execute the native English statements NLP parser will be used and after parsing of code into tree will be done, parsing of code into tree is to enhance the speed and accuracy of the code. First will come "class name" then "main ()" then "methods and functions". The code which will be generated will be trained so that it can generate advanced level coding also here machine learning a subset of Artificial intelligence is required. Machine learning is set of supervised, un-supervised algorithms and reinforcement algorithms to train the models based on the requirement by categorizing train and test data which is chosen randomly or by fold-inspection Method to get better accuracy for the system.

The java compiler API is used to get a tree or graph for the searching algorithm it is free and included inside java distribution which serves the advanced control of compilation process. This API uses Abstract Syntax Tree (AST) this will help for speedy access the benefits associated with it is the replacement of visitor pattern with a classic approach and enrichment of features used. The code which will be generated will be given to the training model to generate advanced codes with speed.

In speech recognition process the problem of inaccuracy due to noise generates for which there are Least mean square (LMS)

and RLS algorithms are used which checks to mix of audio signals with the generated noise and determine impulse level and then comparing with signal to noise ratio and based on fixed decibel values try to remove the noise introduced in the audio signals.

7. CONCLUSION

Automatic speech recognition is an area of development for many years but still, the techniques developed are not fully efficient and accurate. Through this analysis paper, we did a study and analyzed the speech recognition system in detail and studied techniques like LPC, LPCC, RASTA, MFC coefficient. We have observed that every technique has its own advantages and disadvantages. We can also conclude from the above paper that maximum amount of studies has been done on the native English. We can also say that for the English language the recognition rate ratio is much more effective and accuracy is also good, therefore widely used language. We can combine these techniques together making a hybrid approach to make the system reliable and robust. We will be using MFCC based technique in our project.

8. REFERENCES

[1] Harshita Gupta, Divya Gupta, "Lpc and Lpcc method of feature extraction in speech recognition system 978-4673-8203-8/16, the year 2016.

[2] Kartiki Gupta, Divya Gupta "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech Recognition System "978-1-4673-8203-/16, the year 2016.

[3] Kanchan Nathan I, V.M Thakar, Ashish sexual," English language recognition using MFCC and HMM" 978-1-5386/18, the year 2018.

[4] Ibrahim Patel, Srinivas Rao, "Speech recognition using Hidden Markov Model with MFCC-SUBBAND technique", 978-0-7695-3975-1/10, the year 2010.

[5] Abhishek Dixit, Abhinav Vidwans, Pankaj Sharma, "Improved MFCC and LPC Algorithm for Bundelkhandi Isolated Digit Speech Recognition", 978-1-4673-9939-5/18, the year 2018

[6] Vikramjit Mitra, Horacio Franco, Chris Bartels," speech recognition in unseen and noisy channel conditions ", 978-1-5090-4117-6/17, the year 2017.

[7] M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha3and A. Upendra Raju," Speech Based Human Emotion Recognition Using MFCC", 978-1-5090-4442-9/17, the year 2017.

[8] Dr Hebah H. O. Nasereddin, Ayoub Abdel Rahman Omari," Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real-Time Speech Translation", 978-1-5090-5443-5/17, the year 2017.

[9] Abdella K. Mohammad, Amit Pandey,"Sphinx-based speech recognition application for sidma language, IEEE ISBN 978-1-5386-0807-4" the year 2018.

[10] Tomas Bublik, Miroslav virius"source code recognition by graph algorithm" IEEE 978-4673-4490-6 year 2013.

[11] Kshipra prasdh " Efficiency analysis of Noise reduction algorithms "

[12] Bo Wu, Kehuang Li, Fengpei Ge, Zhen Huang, "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition", the year 2017.

[13] Amitrajit Sarkar, Surajit Dasgupta, Sudip Kumar Naskar, Sivaji Bandyopadhyay," says who? Deep learning models for joint speech recognition, segmentation and diarization", 978-1-5386-4658-8/18, the year 2018.

[14] Ruchismita Tripathy, Hrudaya Kumar Tripathy, "Unalike Methodologies of Feature Extraction & Feature Matching in Speech Recognition" the year 2014.

[15] Vibha Tiwari, "MFCC and Its Applications in Speaker Recognition", International Journal on Emerging Technologies, the year 2010.

[16] Ram Singh, and Preeti Rao, "Spectral Subtraction Speech Enhancement with RASTA Filtering", Proceeding of National Conference on Communications (NCC), Kanpur, India, and the year 2007.

[17] Huang Guopin, Zhao Wei, Zhang Qin "Improvement of Audio Noise Reduction System Based on RLS Algorithm" the year 2013

[18] Kunxia Wang, Ning An, Bing-Nan Li, Yanyong Zhang, and Lian Li, "Speech emotion recognition using Fourier parameters," IEEE Transactions on Affective Computing, the year 2015

[19] Nuzhat atiqua Nafis, Md. Safaet hassain,"Speech to text conversion in real time",ISSN 2351-8014 year 2015.

