An Implementation of Various Languages through MFCC, LPCC and Fomant Frequency Analysis

Chandra Chaturvedani[1]
Mtech Scholar (ECE)
Kalinga University, New Raipur, Chhattisgarh, India

Dr. Sunil Kumar[2]
Head of Department (ECE),
Kalinga University, New Raipur, Chhattisgarh, India.

***Abstract*—** Speech is an antique field of study, and research is being done on it to date. Automatic Speech recognition system apportions with analysis and recognition of the input speech signal by the machine or computer in multiple conditions. To enhance the accuracy and capability of the system, different feature extraction techniques are executed. This resulting paper provides a brief overview of the Speech recognition system and its various phases like analysis, feature extraction, modeling and testing, or matching. We also recognize the voice in multiple languages with MFCC , LPC, and formant frequency techniques. This method is used to look at the thoughts of speaker reputation in various styles and understand its uses in classification and verification systems and to assess the recognition ability of different voice functions and factors to find out the technique this is suitable for Automatic Speaker Recognition systems in representations of reliability and computational performance.

Keywords: Speech Recognition, computational efficiency, speaker recognition.

I. INTRODUCTION

The improvements in the speech recognition system have been made for several years. In all audio processing, it is essential to convert the speech input into the feature matrix representation in order to recognize the anonymous speech signal. The process of converting the address signal to the feature matrix design is called the feature extraction of the speech signal. Principle Component Analysis (i.e., PCA), Linear Prediction Coefficients (i.e., LPC), Cepstrum Coefficients derived from LPC (i.e., LPCC), Mel Frequency Cepstrum Coefficients (i.e., MFCC), Wavelet, Reflection Coefficients (i.e., RC) [1], Independent Component Analysis (i.e., ICA), Linear Predictive Coding, Cepstral Analysis, Filter bank analysis, kernel-based methods [2] are the standard feature extraction methods and have been widely employed in speech recognition research field. Among these methods, the MFCC method has been widely used and has achieved high recognition accuracy in speech recognition system [3, 4, 5]. There are many works studying MFCC, especially in improving recognition accuracy [3]. To the most helpful of our experience, none have considered the consequences of different versions of MFCC feature extraction methods in terms of recognition accuracy and calculation speed of the training HMM model, which is an essential process in the speech recognition system. Principle Component Analysis (i.e., PCA) is the usual successful dimensionality reduction techniques [6]. It has much utilization in many areas, such as pattern recognition, computer vision, statistics, and data analysis. It operates the eigenvectors of the covariance matrix of the data to project on a lower-dimensional subspace. This will lead to a decrease of noises in the data and the low time complexity of the training HMM process of the speech recognition system. In this paper, we measure the completion of the conventional MFCC feature extraction method in terms of recognition accuracy and calculation speed of training HMM process of the speech recognition system. We also propose the two modified versions of the MFCC feature extraction method and evaluate their performances in terms of recognition accuracy and calculation speed. In the two proposed versions of the MFCC feature extraction method, we try to apply the Principle Component Analysis (i.e., PCA) techniques [4] to the MFCC feature matrices to reduce its dimension in order to improve the recognition accuracy and reduce the time complexity of the training HMM process. In other words, in these two versions of the MFCC feature extraction method, we combine the PCA technique and the MFCC feature extraction method. To the most beneficial of our information, this combination has not been investigated. MFCC used Filter bank coefficients and get output More data about decrease frequencies than better frequencies because of Mel spaced filter banks subsequently behaves more like a human ear compared to different strategies, based totally on STFT which has fixed time-frequency decision. A mixture of MFCC and LPCC has been proposed for audio function extraction. One of the best benefits of MFCC is that it's far able to figure out capabilities even in the lifestyles of noise and henceforth, it's now combined with the advantage of LPCC which enables in extracting skills in low acoustics.

During the recognition phase, the test speech data is matched with the training models and the result is given according to the best match. Thus, the speech recognition system can be classified mainly into four phases.

- Analysis
- Feature Extraction
- Modeling
- Testing or Matching

Objective

The main objective of the project is to design and comparison an algorithm by which we can recognize the various local language of people using Spectrum Analysis and comparison. In this implemented project, we can compare many voice signals in different languages with different users. We will be using the Matlab platform for this system. We used MFCC, LDA and Formant frequency scheme for this project.

II. METHODOLOGY

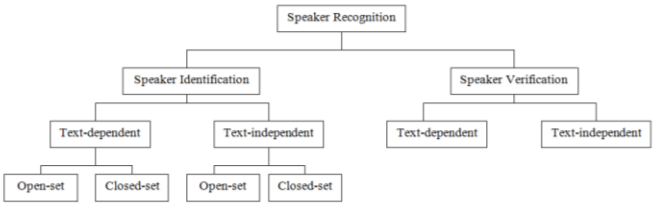

Figure 1: Speaker recognition processing

MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC):

Mel frequency cepstral coefficients (MFCC) are commonly used features for speech recognition. Since MFCC and their modifications will use as the features in our experiments, Individual steps for calculating MFCC are required to know.

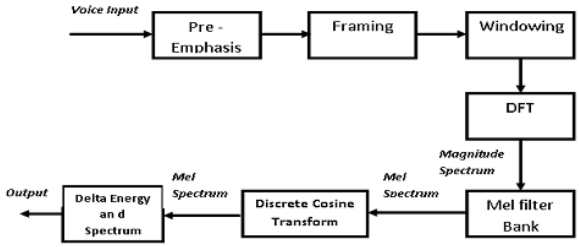

Figure 2: Process to calculate MFCC

The first process in MFCC is preemphasis, which generates energy in a high frequency that previously compressed during the sound generating process. Framing then used to trim the sound signals into smaller sections. The sound data is often stationary with 10 – 30 ms long; therefore, the signal analysis process is conducted in a short period (frame) in the speech recognition system. Thus, it is essential to cut the signals into smaller frames that still contain the original characteristic of the signal analysis process [10]. Windowing is used to avoid discontinuity of signals generated from the framing process. The type of window used in this study was the Hamming window. The function of the Hamming window is expressed in Equation (1). Meanwhile, Equation (2) is the output of each frame after the filter process,

with N is the number of samples per frame, Y[n] is the output signal, X[n] is the input signal, and W[n] is the nth coefficient of the Hamming window [3]. Fast Fourier transform (FFT) is used to convert a signal from the time domain to the frequency domain. FFT is the fast algorithm of discrete Fourier transform (DFT) [11]. Filterbank is the bandpass filter that overlaps each other. Based on the Mel scale, it is linear under frequency 1 kHz and logarithmic on it [12].

Pre-emphasis

In the first step, the speech is passed through a digital filter as in (1) aiming to enhance the signal at frequencies above 1 kHz. In the time domain, the relationship between the output and the input of the pre-emphasis block is shown in

(2) $H(z) = 1 - uz^{-1} \qquad 0.9 < u < 1$

The most characteristic value of a is about 0.95 [7]. The signal spectrum is improved around 20 dB/decade by the pre-emphasis filter.

Framing

Then, the pre-emphasized speech signal is divided into short time segments called "frames" [5]. The length of the overlapping part between adjacent frames, as shown in Fig. 3, is 50% the length of one frame.

Windowing

A window is often applied to increase the continuity between adjacent frames while minimizing the discontinuities between the beginning and end of each frame. In the recognition systems, Hamming window is commonly used as shown in (3):

$$y(n) = x(n)w(n)$$

Hamming window are used for the speech recognition task as

$$w(n) = 0.54 - 0.46\cos(\frac{2n}{N-1})$$

Spectral Estimation

Spectral Feature extraction is an integral part of the Automatic speech recognition system. [18] The performance, quality and accuracy of ASR suffer significantly due to the increase in background noises and linear distortions. Thus, feature extraction is a technique to remove the changeability of the input speech signal while retaining the essential characteristics of the speech. The speech signal is converted into useful parametric- representation, which can be further, analyzed, and classified. [12] This process removes unwanted and redundant information and retains vital information; however, during the removal of some unsolicited information, we could also lose essential details. Feature extraction is not only limited to speech analysis, coding, synthesis, enhancement, and recognition but is also widely used in speaker recognition, voice modification, and language identification. The main goal of feature extraction is to produce a perpetually meaningful representation of digitalized waveform; it changes the input signal into acoustically identifiable components and keeps computations feasible. DFT can be defined as:

$$X(k) = \sum_{n=0}^{N-1} y(n)e^{-\frac{j2\pi n}{N}} \qquad 0 \leq n, k \geq N-1$$

Where *X (K)* are the spectral coefficients, and *y(n)* the framed speech signal

Mel Filtering

A group of triangle bandpass filters that simulate the characteristics of the human's ear are applied to the spectrum of the speech signal. This process is called Mel filtering [10]. The human ears analyze the sound spectrum in groups based on a number of overlapped critical bands. These bands are distributed in a manner that the frequency resolution is high in the low-frequency region and low in the high-frequency region as illustrated in Figure

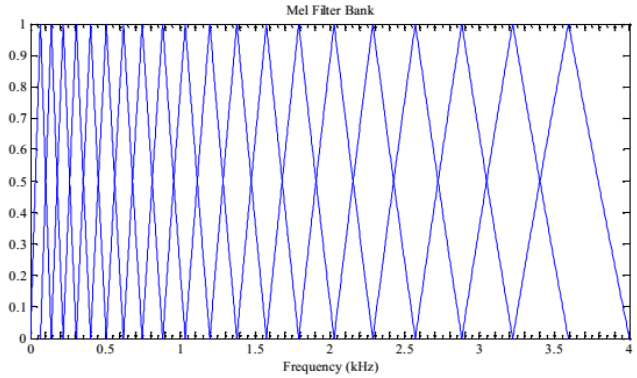

Figure 3: Mel Scale filter bank

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decrease linearly to zero at center frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

$$F(Mel) = [2595 * \log 10[1 + f]700]$$

DCT (DISCRETE COSINE TRANSFORM):

DCT is a Fourier-related transform same like the discrete Fourier transform, but using only real numbers. It is equivalent to DFTs of roughly twice the length, operating on real data. (Since the Fourier transforms give same practical and even function is real and even). A Discrete Cosine Transform computes a sequence of data points at various frequencies gives a summation of cosine functions oscillating. Discrete Cosine Transform on Mel Scale is motivated by speech frequency domain characteristics. The module of the Discrete Cosine Transform reduces the speech signal's repeated information and reaches the speech signal into feature coefficients with minimal dimensions. The final step of the algorithm is to de-correlate the filter outputs. Discrete Cosine Transform (DCT) is applied to the filter outputs and the first few coefficients are grouped together as a feature vector of a particular speech frame.

LPC (LINEAR PREDICTIVE COEFFICIENT):

Linear prediction techniques are the maximum broadly used in speech synthesis, speech coding, speech reputation, speaker identification and verification, and large speech garage. LPC artifices provide correct estimates of speech parameters and do it extraordinarily successfully. The audio signal received from the mic. is sampled, processed for extracting the features. The primary purpose of linear prediction is to predict the output samples with a linear combination of input samples, past samples or both. LPC synthesis imitates human speech production. LPC helps to produce a good model of the audio signal which is right in case of quasi-state voiced regions of speech in which the all-pole model of LPC provides an excellent approximation to the vocal tract spectral envelop. But the LPC model is less suited during unvoiced and transient regions of speech than for voiced regions of speech but it still a useful model for speech recognition.

The idea of Linear Prediction: present-day speech pattern can be closely approximated as a linear aggregate of the earlier samples. LPC is a method that offers a large estimation of the vocal tract spectral envelope and is hazardous in speech evaluation because of the efficiency and pace with which it can be derived. The specific vectors are calculated by way of LPC over each frame. The coefficients used to design the structure typically tiers from 10 to twenty depending on the speech sample, application, and range of poles within the version. However, LPC also has dangers. Firstly, LPC approximates speech linearly in any respect frequencies that are incompatible with the listening to the notion of people. Secondly, LPC may be very susceptible to noise from the heritage, which may additionally cause mistakes within the speaker modeling.

LINEAR PREDICTION CEPSTRAL COEFFICIENTS (LPCC):

LPCC represents the characteristics of positive speech channel, and the equal character with distinctive emotional speech can have multiple channel capabilities, thereby extracting those function coefficients to categorize the feelings contained in the statement. The computational manner of LPCC is often a repetition of computing the linear prediction coefficients (LPC) LPC is one of the maximum powerful speech evaluation strategies and is a beneficial technique for encoding excellent speech at a low bit charge. For estimating the fundamental parameters of a speech sign, LPCC has to turn out to be one of the primary strategies. The central topic at the back of this technique is that one speech pattern on the modern time may be expected as a linear aggregate of past speech samples,

LPCC is a method that mixes LP and cepstral evaluation by means of taking the inverse Fourier rework of the log importance of the LPC spectrum for improved accuracy and robustness of the voice functions extracted.

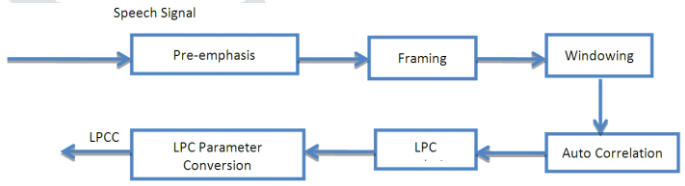

Figure 4: Process of calculating LPCC

FORMANT FREQUENCY

A formant is a convergence of acoustic power around an accurate frequency in the speech wave. There are various formants, each at a different frequency, roughly one in each 1000Hz band. Or, to put it separately, formants happen at approximately 1000Hz intervals. Each formant corresponds to a noise in the vocal tract. Formants can be understood very clearly in a wideband spectrogram, where they are presented as dark bands. The darker a formant is represented in the spectrogram, the more effective it is (more energy there is there, or the more audible it is).

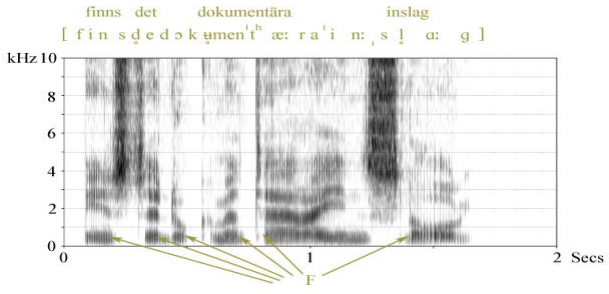

Figure 5: Analysis of the Formant frequency spectrum

The signs at F on this spectrogram point out six examples of the lowest formant. The next formant transpires just above these, between 1 and 2 kHz. Then the following is just above that, between 2 and 3kHz. And so on. When you study at a spectrogram, like this instance, you will see formants universally, in both vowels and consonants. To get why you necessity recall the source-filter theory of speech generation. The vocal tract filters a source sound (e.g., periodic voice fluctuations or aperiodic hissing), and the result of the filtering is the sound you can listen and record outside the lips and show on a spectrogram.

III. TEST AND IMPLEMENTATION

In this project, we implement a schematic GUI using a Matlab platform. It's completed and manages efficiently by the user. User can add their voice using records and save a .wav music file. In this project, we have tried to analyze MFCC, LPCC, and Formant Frequency of the system, which is used for recognition of the user's voice. We used three local languages HINDI, Rajasthani and Marathi. We use a Matlab speech processing toolbox to voice recognition. We try to find MFFC and LPCC parameters of each voice.

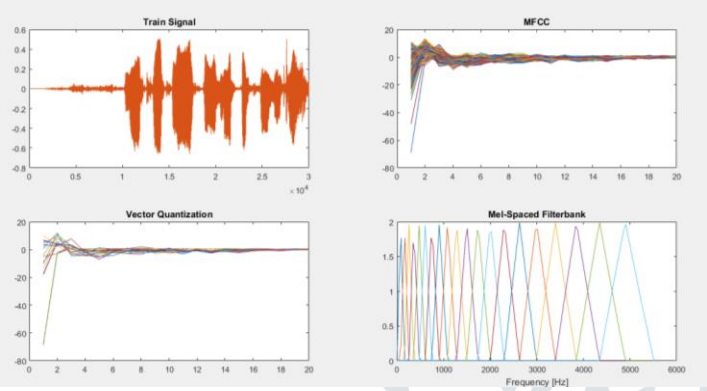

Figure 6: Voice component feature comparison using the MFFC analysis

Using MFFC, we can find the vector quantization level and the mel-spaced filter coefficient this method can fast response as compare to LPCC.

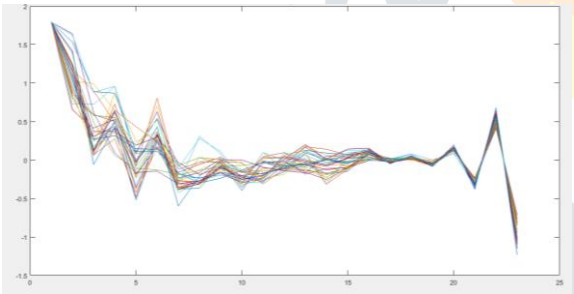

Figure 7: LPPC analysis of voice

IV. RESULTS

The speech recognition System operates in two approaches, that is the Enrollment approach and the Recognition approach. The MFCC feature coefficient is used here for purposes stated earlier. Euclidean distance is used to measure the gap between the feature vectors (Itakura-Saito). The main part of this project work is the implementation and also analysis of the Automatic Speech Recognition algorithms using MATLAB and analyzed their performance. The speech recognition system model has been formed for comparing the two algorithms that are MFCC and LPCC. The production has been estimated by examining the sets of speech signals. It is an analysis that MFCC used in Automatic Speech Recognition system (ASP) gives 90 percent accuracy, whereas LPCC used in Automatic Speech Recognition (ASP) given 83 percentages accuracy. Results and analysis show that the MFCC algorithm gives better results than the LPCC algorithm. From the simulation results, we also add that the MFCC algorithm, which requires more calculation but works better than LPCC in phases of efficiency and accuracy.

We try to analyze many voice signals from various users, which displays in the table below.

SENTENCE	Language	LPCC	MFCC
Speaker 1	HINDI	94.6%	98%
	MARATHI	97.7%	98.7%
	RAJASTHANI	96.7%	98.9%
Speaker 2	HINDI	95.5%	98.4%
	MARATHI	98.5%	98.6%
	RAJASTHANI	96.6%	98.8%
Speaker 3	HINDI	96.6%	98.4%
	MARATHI	94.7%	98.7%
	RAJASTHANI	96.8%	98.9%
Speaker 4	HINDI	94.2%	98%
	MARATHI	97.3%	98.9%
	RAJASTHANI	96.7%	97.9%
Speaker 5	HINDI	94.1%	98.8%
	MARATHI	97.9%	98.7%
	RAJASTHANI	96.4%	96.9%
Speaker 6	HINDI	97.3%	96%
	MARATHI	98.2%	99.7%
	RAJASTHANI	96.8%	98.9%
Speaker 7	HINDI	95.6%	96%
	MARATHI	97.8%	97.7%
	RAJASTHANI	96.9%	98.9%
Speaker 8	HINDI	97.6%	99%
	MARATHI	95.7%	98.7%
	RAJASTHANI	98.7%	99.9%
Speaker 9	HINDI	99.6%	100%
	MARATHI	97.7%	99.9%
	RAJASTHANI	96.7%	99.5%
Speaker 10	HINDI	96.6%	98.7%
	MARATHI	97.7%	98.6%
	RAJASTHANI	96.7%	99.9%

V. DISCUSSION

We have used a decision-based algorithm in our system, which is presented in the paper. In this improved algorithm, we can be successfully analyzed and match the speech signal. We can design an automatic speech recognition system, where the user can be analysis their voice in different languages. We plot the graph for each speech signal. We can also compare each signal.

VI. CONCLUSION

In this implemented methodology, a system is to be designed, which can recognize any language and plot the identical spectrum as per the identified language. The designed curve will signify each word whatever is said by the speaker. Listeners outperform Automatic speech recognition systems in all speech recognition tasks. Advanced high-tech automatic speech recognition systems perform very well in environments where the speech signals are reasonably clean. In most cases, recognition by machines degrades dramatically with a slight adjustment in speech signals or speaking environment. Thus these complex algorithms are used to signify this unpredictability. So, the speech can be simply recognized through the spectrogram.

VII. Future Enhancement

We are allowing for an indoor environment(less noise). We need to see that the classification of sounds into global classifications can be performed with a deficient calculation effort for gender recognition algorithms that needed the low cost and frequency domain features to achieve results. We see for the different categories (Global, Gender, and indoor sound classification) that the use of low-cost algorithms can be equivalently effective for deployment indoors as the application of high-cost algorithms.

References

[1] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, Kong-Pang Pun, "An Efficient MFCC Extraction Method in Speech Recognition", the 2006 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 145-148, Greece (2006).

[2] Renanti, Medhanita Dewi, Agus Buono, and Wisnu Ananta Kusuma. "Infant cries identification by using codebook as feature matching, and MFCC as feature extraction." Journal of Theoretical & Applied Information Technology 56.3 (2013)

[3] Zhang X, Sun J, Luo Z, "One-against-All Weighted Dynamic Time Warping for Language-Independent and Speaker-Dependent Speech Recognition in Adverse Conditions" 2014. PLoS ONE 9(2): e85458, doi:10.1371/journal.pone.0085458 .

[4] S. Sharma, M. Kumar, and P. K. Das, "A technique for dimension reduction of MFCC spectral features for speech recognition," in Industrial Instrumentation and Control (ICIC), 2015 International Conference on, 2015, pp. 99–104.

[5] S. Gaikwad, B. Gawali, P. Yannawar, and S. Mehrotra, "Feature extraction using fusion MFCC for continuous marathi speech
recognition," in India Conference (INDICON), 2011 Annual IEEE, 2011, pp. 1–5.

[6] B. J. Mohan, "Speech recognition using MFCC and DTW," in Advances in Electrical Engineering (ICAEE), 2014 International Conference on, 2014, pp. 1–4.

[7] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," in Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of, 2017, vol. 1, pp. 701–704.

[8] S. Attawibulkul, B. Kaewkamnerdpong, and Y. Miyanaga, "Noisy speech training in MFCC-based speech recognition with noise suppression toward robot assisted autism therapy," in Biomedical Engineering International Conference (BMEiCON), 2017 10th, 2017, pp. 1–5.

[9] Garima Vyas, and Barkha Kumari, "Speaker Recognition System Based On MFCC and DCT", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Vol.2, Issue-5, pp. 145-148, June 2013.

[10] Shumaila Iqbal, Tahira Mahboob, and Malik Sikandar Hayat Khiyal, "Voice Recognition using HMM with MFCC for Secure
ATM", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 6, No 3, pp.297-303, November 2011.

[11] Santosh Gaikwad, Bharti Gawali, and Pravin Yannawar, "Performance Analysis of MFCC & DTW for Isolated Arabic Digit", International Journal of Advanced Research in Computer Science, Vol. 2 (1), pp. 513-518 Jan. –Feb, 2011.

[12] Serajul Haque, Roberto Togneri, and Anthony Zaknich, "ZeroCrossings with Adaptation for Automatic Speech Recognition",
Proceeding of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I.
Watson.University of Auckland, New Zealand. December 6-8, pp. 199-204, 2006.

[13] J. Manikandan, and B. Venkataramani, "Design of a real time automatic speech recognition system using Modified One Against
All SVM classifier" Microprocessors and Microsystems, ELSEVIER, Vol.35, pp. 568–578, 2011.

[14] Jun Wang, Dong Wang, Ziwei Zhu, Thomas Fang Zheng and Frank Soong, "Center for Speaker and Language Technologies
(CSLT)", on I-vectors, a Discriminative Scoring for Speaker Recognition Based, 2014.

[15] Wenjing Liu ; Balu Santhanam, "Large Deviation First Formant Demodulation Via Empirical Mode Decomposition And Multirate Frequency Transformations",2018 52nd Asilomar Conference on Signals, Systems, and Computers.

[16] Yunus Korkmaz ; Aytuğ Boyacı, "Classification of Turkish Vowels Based on Formant Frequencies",2018 International Conference on Artificial Intelligence and Data Processing (IDAP).

[17] Y.V. Srinivasa Murthy ; Shashidhar G. Koolagudi ; Vishnu G. Swaroop, "Vocal and Non-vocal Segmentation based on the Analysis of Formant Structure", 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR).

[18] M. S. Likitha ; Sri Raksha R. Gupta ; K. Hasitha ; A. Upendra Raju, "Speech based human emotion recognition using MFCC", 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).