iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55124

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Advancements in Speaker Recognition: Exploring Mel Frequency Cepstral Coefficients (MFCC) for Enhanced Performance in Speaker Recognition

V. Sai Nitin Varma[1], Abdul Majeed. K. K[2]
Department of Electronics (SENSE), Vellore Institute of Technology, Vellore- 632014

Abstract: ***Speaker recognition, a fundamental capability of software or hardware systems, involves receiving speech signals, identifying the speaker present in the speech signal, and subsequently recognizing the speaker for future interactions. This process emulates the cognitive task performed by the human brain. At its core, speaker recognition begins with speech as the input to the system. Various techniques have been developed for speech recognition, including Mel frequency cepstral coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP). Although LPC and several other techniques have been explored, they are often deemed impractical for real-time applications. In contrast, MFCC stands out as one of the most prominent and widely used techniques for speaker recognition. The utilization of cepstrum allows for the computation of resemblance between two cepstral feature vectors, making it an effective tool in this domain. In comparison to LPC-derived cepstrum features, the use of MFCC features has demonstrated superior performance in metrics such as False Acceptance Rate (FAR) and False Rejection Rate (FRR) for speaker recognition systems. MFCCs leverage the human ear's critical bandwidth fluctuations with respect to frequency. To capture phonetically important characteristics of speech signals, filters are linearly separated at low frequencies and logarithmically separated at high frequencies. This design choice is central to the effectiveness of the MFCC technique. The primary objective of the proposed work is to devise efficient techniques that extract pertinent information related to the speaker, thereby enhancing the overall performance of the speaker recognition system. By optimizing feature extraction methods, this research aims to contribute to the advancement of speaker recognition technology.***

Keywords: Speech recognition, Mel frequency cepstral coefficients (MFCC), Feature extraction, Speech signal processing, Speaker identification, Cognitive computing.

I. INTRODUCTION

Human speech is a fundamental means by which individuals convey their emotions, opinions, thoughts, and ideas orally. The process of speech production involves intricate coordination of articulation, voice generation, and fluency [1]. This natural motor ability, observed in regular adults, allows for the production of approximately 14 distinct speech sounds per second, facilitated by a harmonized network of around 100 muscles interconnected through spinal and cranial nerves [2]. Despite the apparent ease with which humans speak, the underlying complexity of this task makes speech highly sensitive to diseases associated with the nervous system, which might offer insights into related pathologies [3]. To improve speaker recognition systems, researchers, such as Xinhui Zhou and Garcia Romero (2012), have explored the comparison of Linear Frequency Cepstral Coefficients (LFCC) and Mel Frequency Cepstral Coefficients (MFCC) [4]. Their investigation was guided by observations from speech production, suggesting that certain speaker characteristics, particularly those related to the vocal tract structure, are more prominent in the high-frequency region of speech [5][6]. Analysis of MFCC and LFCC performances in the NIST SRE 2010 extended core task revealed that while these techniques complement each other, LFCC consistently outperforms MFCC, especially in female trials, owing to its ability to capture spectral characteristics in the high-frequency range [7][8][9]. While both methods exhibit resilience against babble noise, LFCC's performance differs from MFCCs when dealing with white noise [10][11].

In the pursuit of robust feature extraction from MFCC techniques, Sahidullah and Saha proposed the 'Differentiation in frequency domain' approach in 2013, modifying the Hamming window technique to obtain the derivative of Fourier transform [12]. Among the existing techniques for speaker extraction, MFCC remains one of the most well-known and widely used methods[13]. The utilization of cepstrum allows for the computation of resemblance between two cepstral feature vectors [14].

This paper comprehensively explores the intricacies of human speech production and investigates the comparative efficacy of MFCC in the domain of speaker recognition. The findings contribute to a deeper understanding of speech-related mechanisms and offer potential implications for advancing speaker recognition technologies.

II. SYSTEM DESCRIPTION

A. *Block Diagram of MFCC*

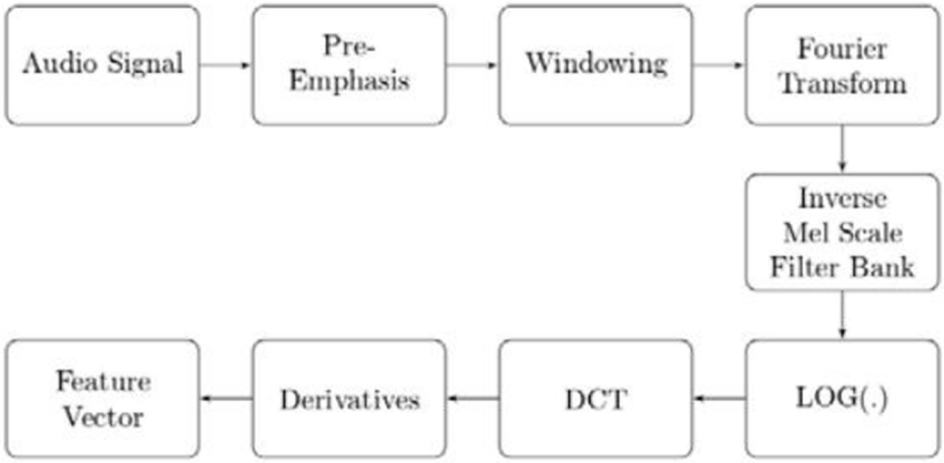


Fig 1.

B. *Pre-emphasis*

Pre-emphasis is performed to enhance the higher frequencies of the spectrum. Pre-emphasis flattens the signal making it less susceptible to finite precision.

C. *Framing*

The speech signal exhibits quasi-periodic characteristics, displaying a repetitive pattern over time. To effectively analyze the speech signal, it is divided into several frames, each lasting approximately 20 to 30 milliseconds. Within each frame, the speech signal is considered to be stationary, facilitating meaningful analysis. In order to ensure a comprehensive examination without loss of crucial information, a 50% overlap is applied between successive frames [16].

D. *Windowing*

Windowing involves the slicing of the audio waveform into sliding frames. let, w is the window applied to the original audio clip in the time domain. x[n]=w[n]s[n]
Here, w[n] is a sliced frame whereas s[n] is an original audio clip.
The corresponding equations for w are:

$$w[n] = (1-\alpha) - \alpha\cos\left(\frac{2\pi n}{L-1}\right)$$

L = window width

Hamming(α=0.46164) or Hanning(α=0.5) window.

E. *FFT*

The frequency content of the windowed signal is estimated using the Fast Fourier Transform (FFT), which facilitates the conversion of each frame from the time domain to the frequency domain. This process allows for the computation of the short-term frequency content of the signal through the application of FFT on individual frames. Subsequently, the power spectrum is computed by evaluating the squared magnitude of the windowed signal.

F. Mel-Filter Bank

The windowed signal is multiplied with a mel-filter bank. The Mel scale is an auditory scale similar to the frequency scale of the human ear (similar to how the human ear perceives sound). The scale is roughly linear below 1 kHz and logarithmic above 1 kHz. The relationship between linear frequency and mel scale is given by following formula:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right)$$

G. DCT

In the last step, logarithm operation is performed followed by discrete cosine transform to de-correlate the log energies. The DCT compresses the signal. The cepstrum is calculated using discrete cosine transform(DCT) or inverse Fourier transform to obtain MFCCs [16].

H. Correlation Technique

Syntax for cross correlation: r=xcorr(x,y) Correlation is a measure of similarity of two series as a function of the displacement of one relative to the other. This is also known as a sliding dot product or sliding inner product [17].

III. RESULTS

A. Input Audio signals

- *Voice 1*

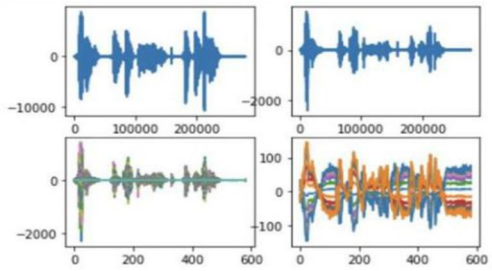


Fig 2.

- *Voice 2*

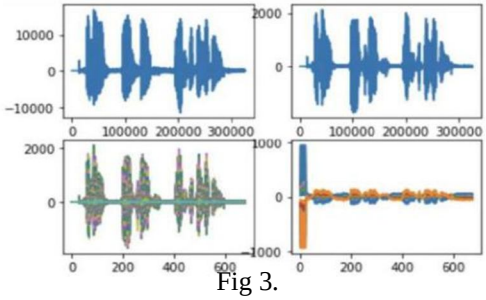


Fig 3.

- *Voice 3*

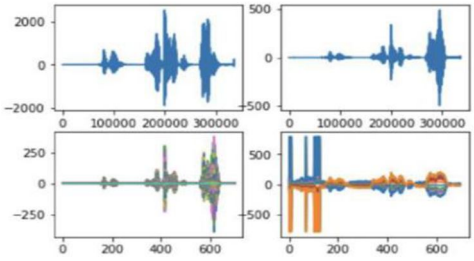


Fig 4.

- *Voice 4*

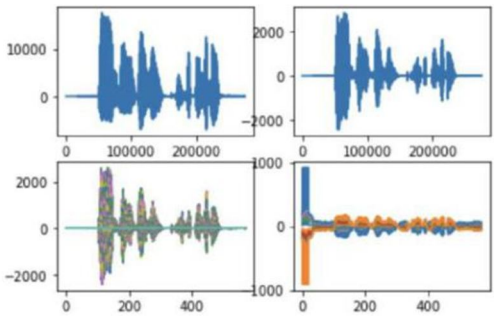


Fig 5.

1) Voice 1 is set as Input voice

Correlation of Voice 1 in database with input voice

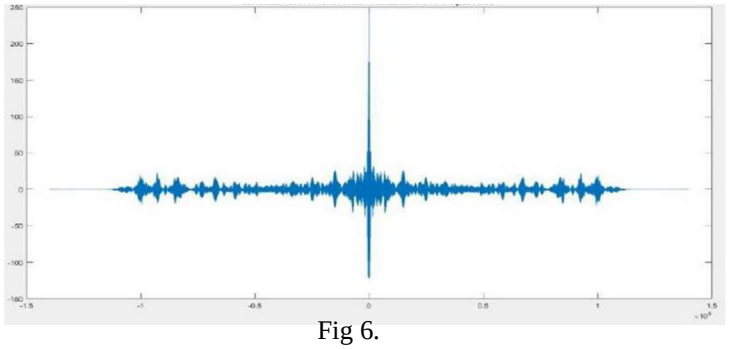


Fig 6.

Correlation of Voice 2 in database with input voice

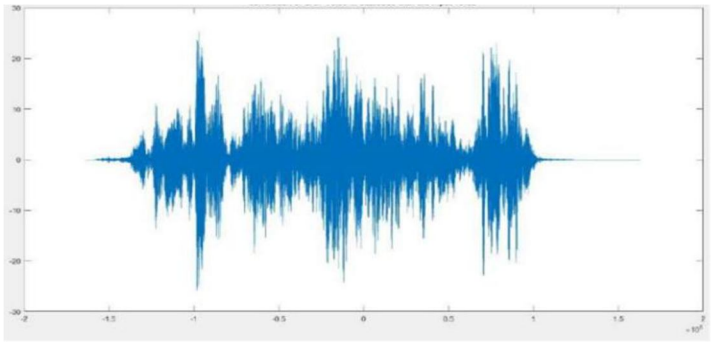


Fig 7.

Correlation of Voice 3 in database with input voice

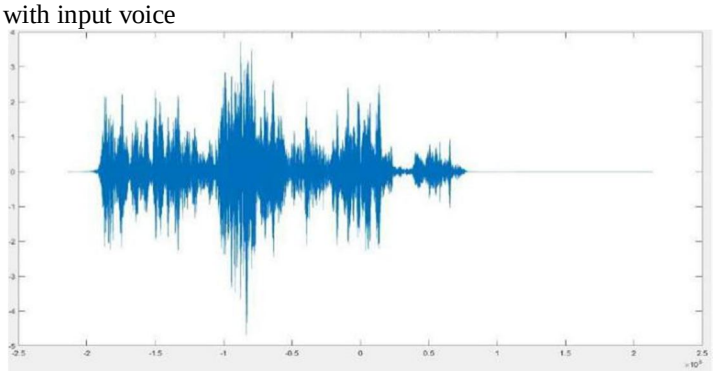


Fig 8.

Correlation of Voice 4 in database with input voice

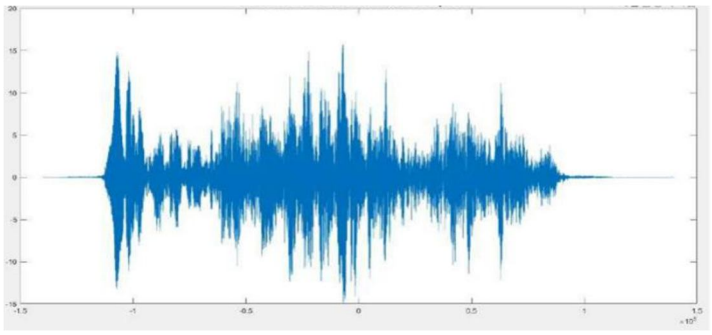


Fig 9.

Peaks

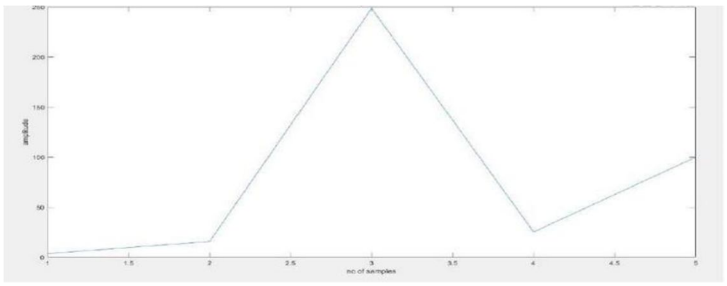


Fig 10.

2) *Voice 2 is set as Input voice :*

Correlation of Voice 1 in database with input voice

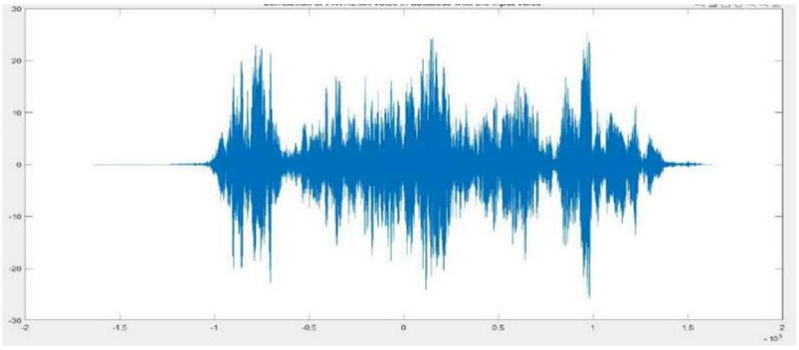


Fig 11.

Correlation of Voice 2 in database with input voice

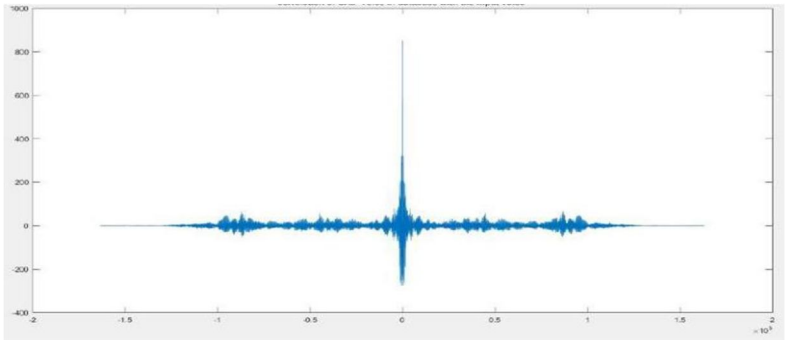


Fig 12.

Correlation of Voice 3 in database with input voice

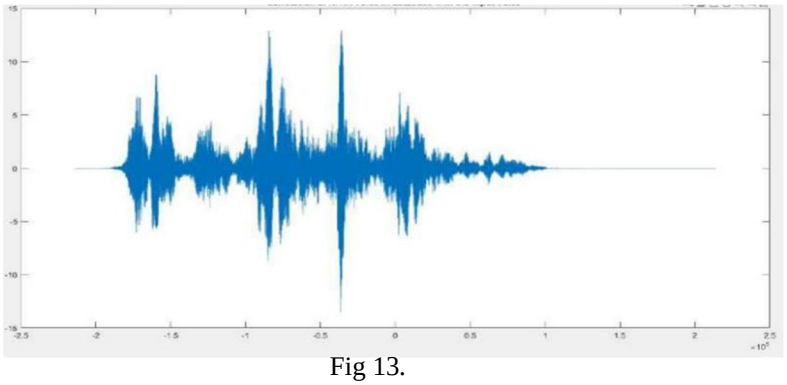


Fig 13.

Correlation of Voice 4 in database with input voice

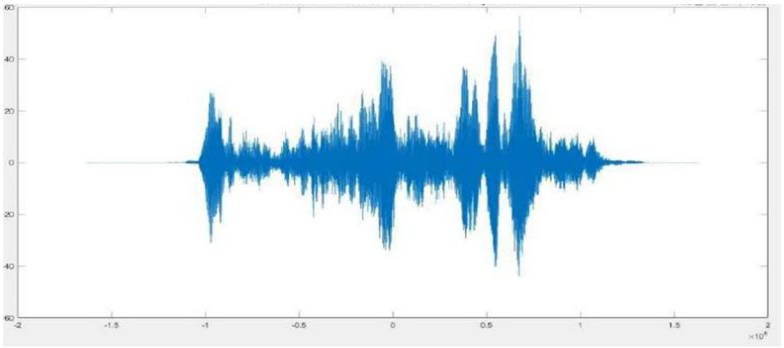


Fig 14.

Peaks

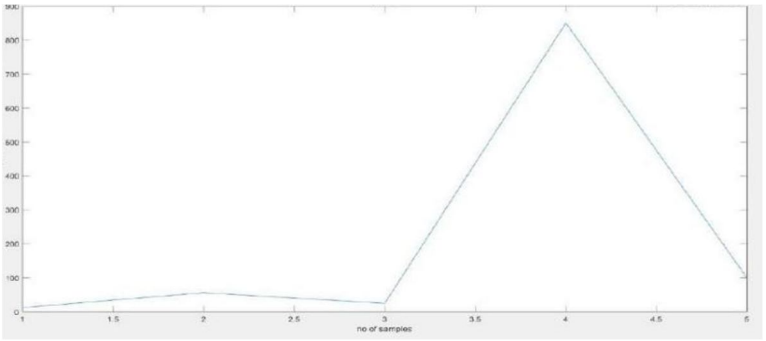


Fig 15.

3) Voice 3 is set as Input voice :

Correlation of Voice 1 in database with input voice

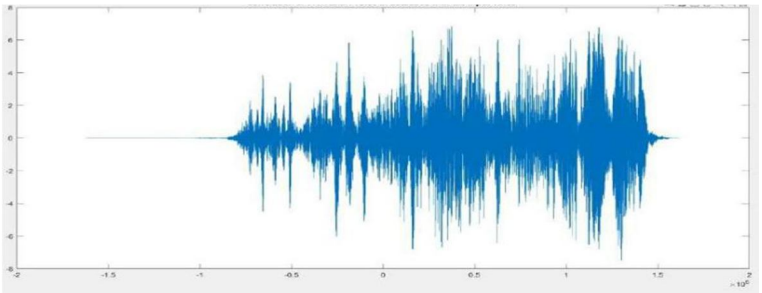


Fig 16.

Correlation of Voice 2 in database with input voice

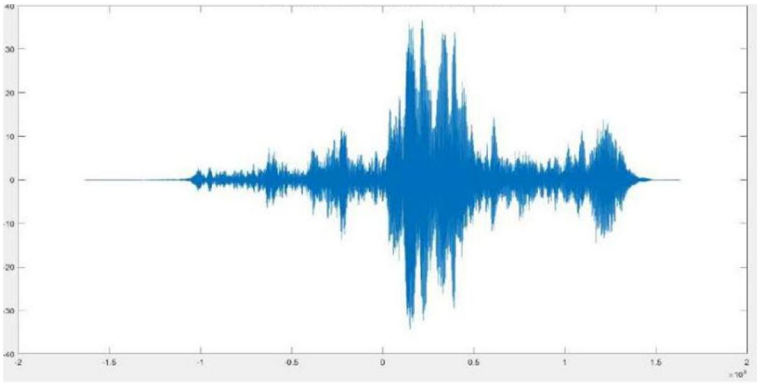


Fig 17.

Correlation of Voice 3 in database with input voice

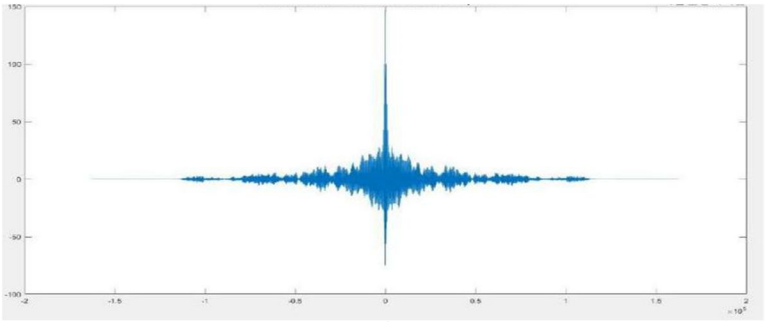


Fig 18.

Correlation of Voice 4 in database with input voice

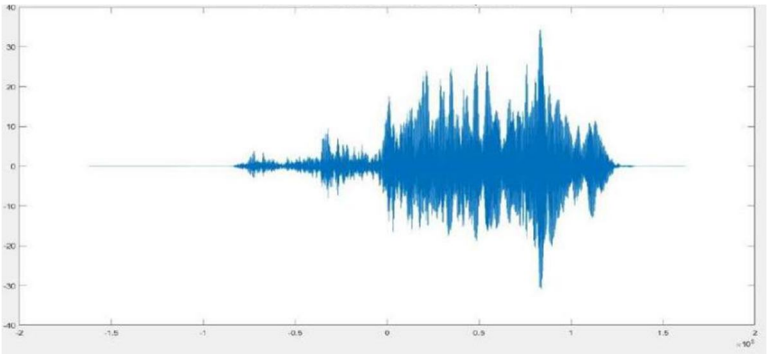


Fig 19.

Peaks

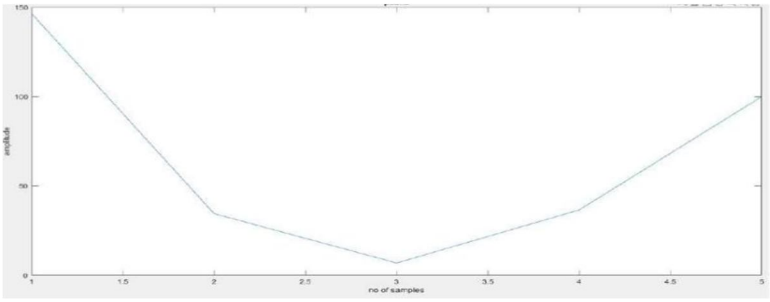


Fig 20.

4) Voice 4 is set as Input Voice

Correlation of Voice 1 in database with input voice

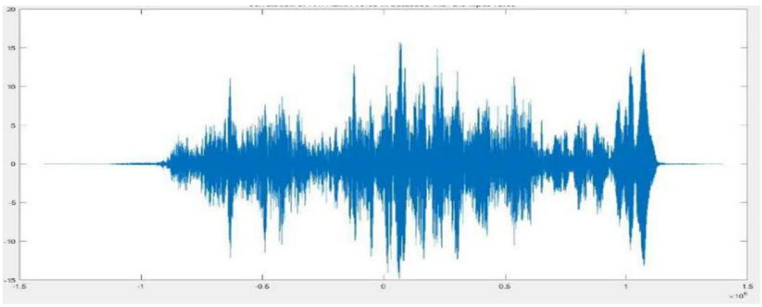


Fig 21.

Correlation of Voice 2 in database with input voice

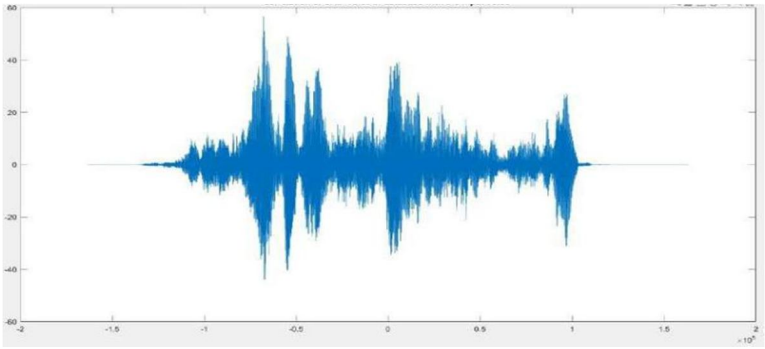


Fig 22.

Correlation of **Voice 3** in database with input voice

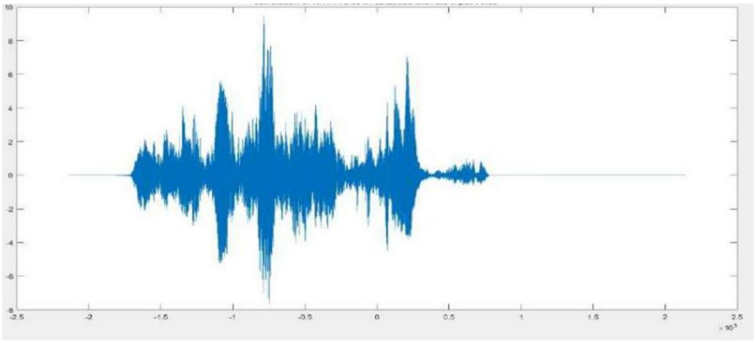


Fig 23.

Correlation of Voice 4 in database with input voice

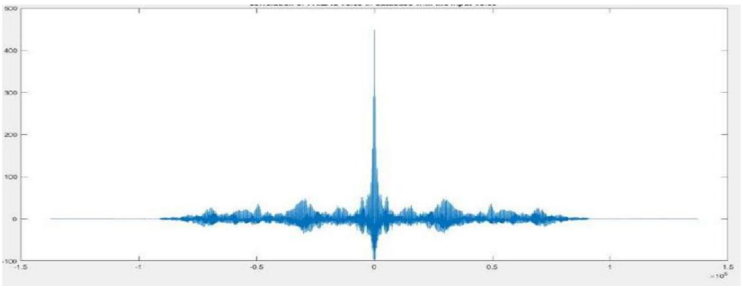


Fig 24.

Peaks

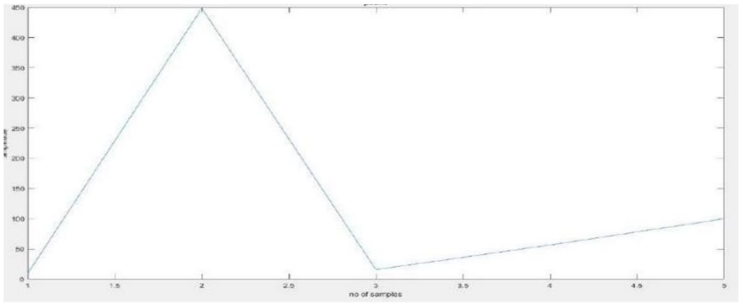


Fig 25.

B. Random Voice Input

Correlation of Voice 1 in database with input voice

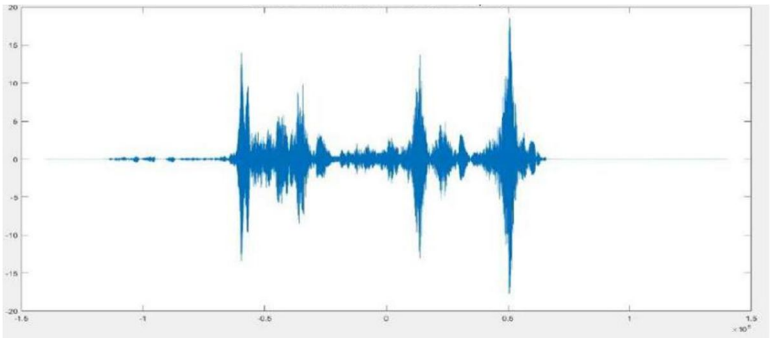


Fig 26.

Correlation of Voice 2 in database with input voice

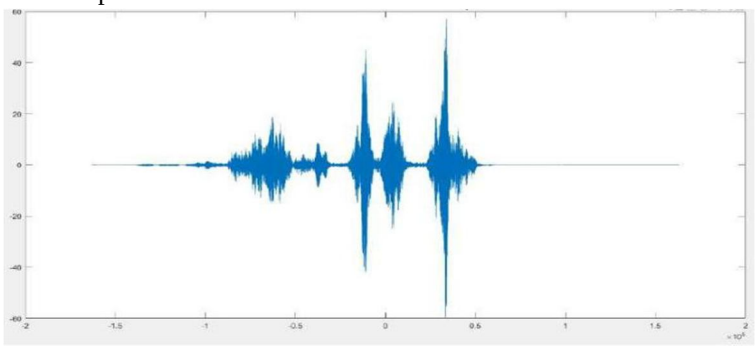


Fig 27.

Correlation of Voice 3 in database with input voice

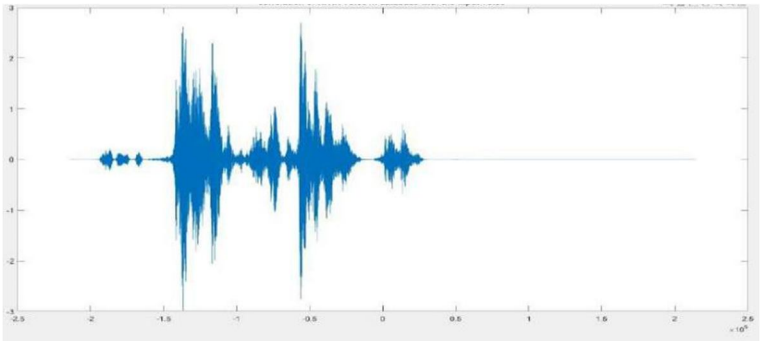


Fig 28.

Correlation of Voice 4 in database with input voice

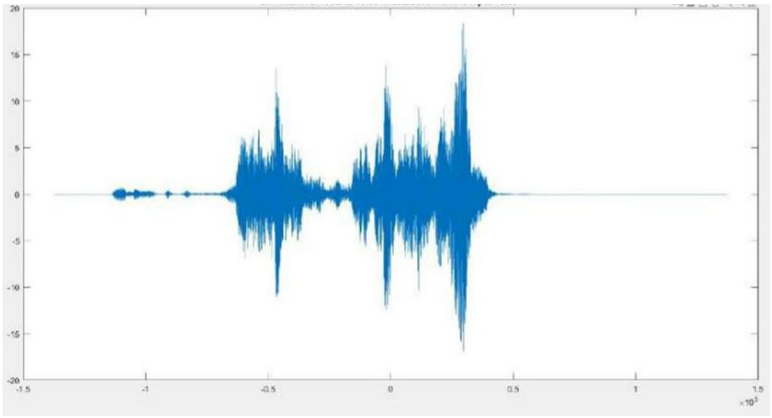


Fig 29.

Peaks

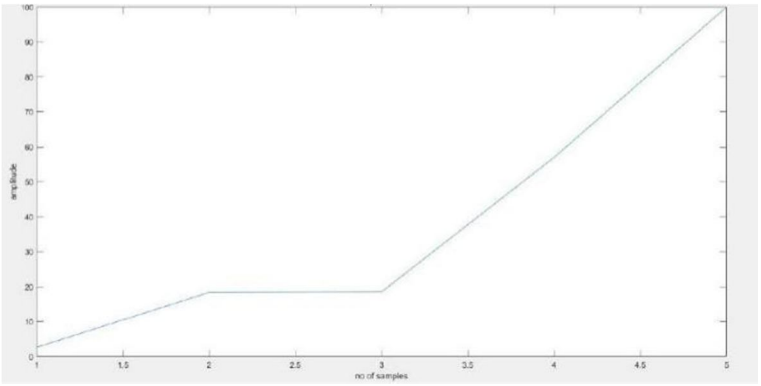


Fig 30

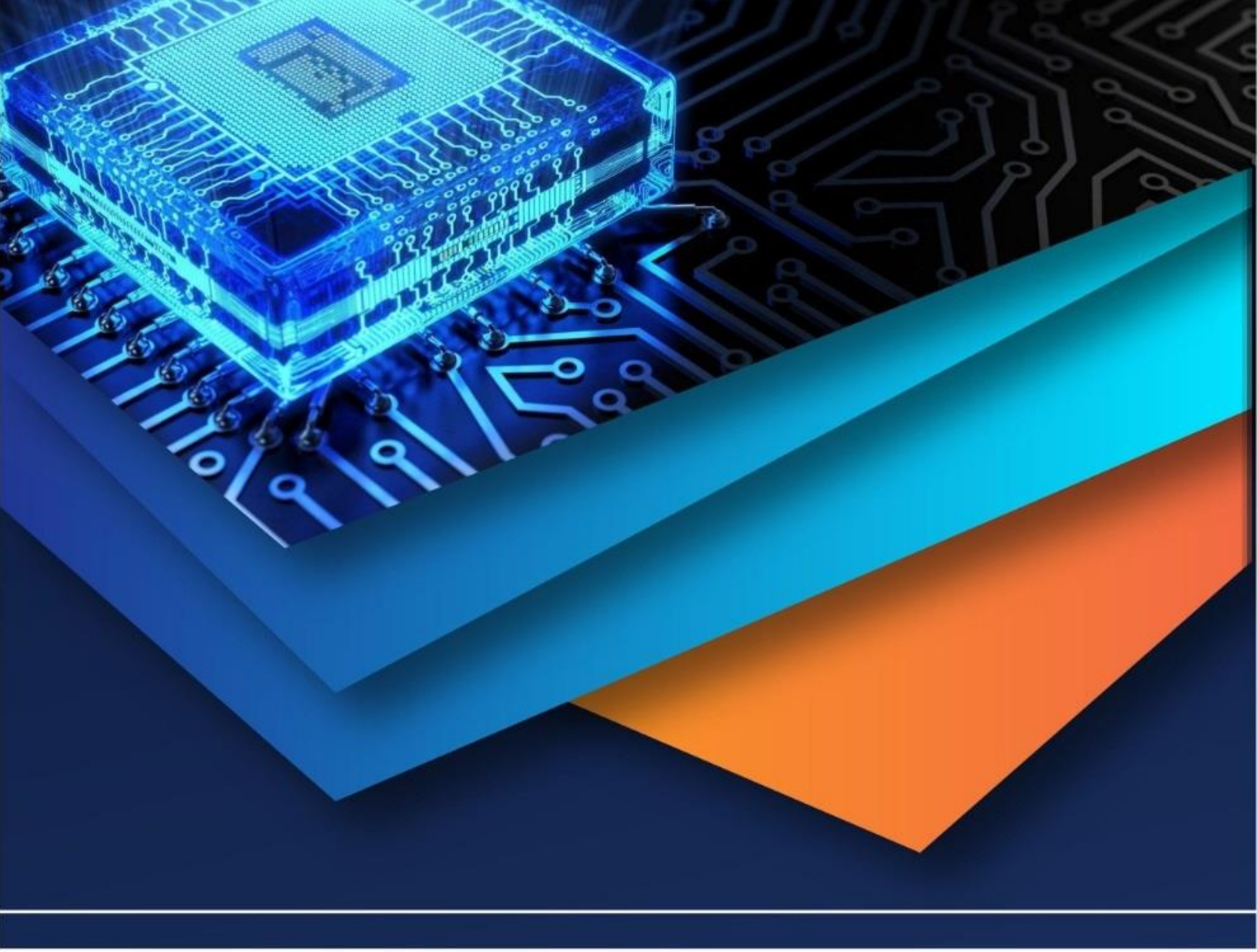
IV. CONCLUSION

We have reviewed feature extraction technique Mel-Frequency Cepstral Coefficients (MFCCs) for speaker recognition. The factor's channel mismatch, background noise affects the performance of MFCC technique. The scheme for robust feature extraction is necessary. Speech signals are extracted by using MFCC technique where features are extracted using linearly spaced filters in Mel scale. Compared to other existing techniques, this method of MFCC provides the beer feature extraction. The weighted vector quantization is suitable for 2D acoustic signals, leading to higher material recognition accuracy than that of other systems. Correlation process is also a nice upon for speech recognition but it's not accurate as of the MFCC process. The whole process of speech recognition using correlation methods has been explained by our project. Application of methods explained above. These methods can be used for security purposes of door unlocking using speech recognition.

REFERENCES

[1] Yıldız and U. Arıöz, "Analyzing human voice and classification of voice frequencies according to smoking effect," 2016 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, Turkey, 2016, pp. 653-656, doi: 10.1109/SIU.2016.7495696.
[2] K. Shiomi, "Voice processing technique for human cerebral activity measurement," 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 2008, pp. 3343-3347, doi: 10.1109/ICSMC.2008.4811813.
[3] B. Roark, M. Mitchell, J. -P. Hosom, K. Hollingshead and J. Kaye, "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2081-2090, Sept. 2011, doi: 10.1109/TASL.2011.2112351.
[4] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 2011, pp. 559-564, doi:10.1109/ASRU.2011.6163888.
[5] A. Alwan, "Modeling speech production and perception mechanisms and their applications to synthesis, recognition, and coding," ISSPA '99. Proceedings of the Fifth International Symposium on Signal Processing and its Applications (IEEE Cat. No.99EX359), Brisbane, QLD, Australia, 1999, pp. 7 vol.1-, doi: 10.1109/ISSPA.1999.818096.

[6] J. Yamauchi and T. Shimamura, "Noise estimation using high frequency regions for speech enhancement in low SNR environments," Speech Coding, 2002, IEEE Workshop Proceedings., Ibaraki, Japan, 2002, pp. 59-61, doi: 10.1109/SCW.2002.1215723.
[7] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 5292-5295, doi: 10.1109/ICASSP.2011.5947552.
[8] S. P. Dewi, A. L. Prasasti and B. Irawan, "The Study of Baby Crying Analysis Using MFCC and LFCC in Different Classification Methods," 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, 2019, pp. 18-23, doi: 10.1109/ICSIGSYS.2019.8811070.
[9] H. Cao, T. Huang, Z. Zhou and C. Yu, "LFCC-based transformer sound feature extraction and analysis," 12th International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE 2022), Emeishan, China, 2022, pp. 1618-1622, doi: 10.1049/icp.2022.3099.
[10] M. Mohammadi and H. R. Sadegh Mohammadi, "Study of speech features robustness for speaker verification application in noisy environments," 2016 8th International Symposium on Telecommunications (IST), Tehran, Iran, 2016, pp. 489-493, doi: 10.1109/ISTEL.2016.7881869.
[11] X. Fan and J. H. L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 4553-4556, doi: 10.1109/ICASSP.2009.4960643.
[12] M. Sahidullah and G. Saha, "A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition," in IEEE Signal Processing Letters, vol. 20, no. 2, pp. 149-152, Feb. 2013, doi: 10.1109/LSP.2012.2235067.
[13] M. Sadeghi and H. Marvi, "Optimal MFCC features extraction by differential evolution algorithm for speaker recognition," 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS), Shahrood, Iran, 2017, pp. 169-173, doi: 10.1109/ICSPIS.2017.8311610.
[14] A. Kumar, S. S. Rout and V. Goel, "Speech Mel Frequency Cepstral Coefficient feature classification using multi level support vector machine," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, India, 2017, pp. 134-138, doi: 10.1109/UPCON.2017.8251036.
[15] C. Paseddula and S. V. Gangashetty, "DNN based Acoustic Scene Classification using Score Fusion of MFCC and Inverse MFCC," 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 2018, pp. 18-21, doi: 10.1109/ICIINFS.2018.8721379.
[16] Bhadragiri Jagan Mohan and Ramesh Babu N., "Speech recognition using MFCC and DTW," 2014 International Conference on Advances in Electrical Engineering (ICAEE), Vellore, India, 2014, pp. 1-4, doi: 10.1109/ICAEE.2014.6838564.
[17] A. Tamaki, Feng-Hui Yao and K. Kato, "On the detection of feature on the line segment by using cross correlation," Proceedings of the 35th SICE Annual Conference. International Session Papers, Tottori, Japan, 1996, pp. 1089-1094, doi: 10.1109/SICE.1996.865415.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY