I. J. Engineering and Manufacturing, **2022, 6, 14-21**
Published Online on December 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2022.06.02

Extricate Features Utilizing Mel Frequency Cepstral Coefficient in Automatic Speech Recognition System

Gaurav D. Saxena
Department of Computer Science, Kamla Nehru Mahavidyalaya, Nagpur, India
E-mail: gauravsaxena711@gmail.com
ORCID iD: https://orcid.org/0000-0002-2204-4261?lang=en

Nafees A. Farooqui*
Department of Computer Science, Era University, Lucknow, India
E-mail: nafeesf@gmail.com
ORCID iD: https://orcid.org/0000-0002-7583-3467
*Corresponding Author

Saquib Ali
Department of Computer Science, Era University, Lucknow, India
E-mail: saquibali99@gmail.com
ORCID iD: https://orcid.org/0000-0001-6765-9780

Received: 05 August, 2022; Revised: 20 September, 2022; Accepted: 31 October, 2022; Published: 08 December, 2022

Abstract: As of late, Automatic speech recognition has advanced on account of instruments, for example, natural language processing, and deep learning, among others. It is a framework or put in another way, a gadget that changes a raw signal into computer comprehensible text. The genuine creation of speech is comprised of changes in air pressure that outcomes in pressure wave that our ear and cerebrum comprehend. The vocal tract is utilized to deliver a human speech, which is adjusted by teeth, tongue, and lips. Speech recognition alludes to a machine's ability to perceive human speech and transform it into a computer comprehensible text. Speech recognition is a magnificent illustration of good interaction between humans and computers. In this paper, we introduce the process to extricate the feature from the signal utilizing Mel-frequency cepstral coefficients. Mel-frequency cepstral coefficients are a genuinely far wide and proficient methodology for feature extraction from a sound file. This technique improved the speech recognition process and removes the distortion in the voice. In this manuscript we applied the Mel-frequency filtration process to improve speech and remove the background noise. Therefore, the proposed methodology gives better performance in the automated speech recognition system.

Index Terms: Features, Mel filter banks processing, Mel frequency cepstral coefficients, Sound file, Speech Recognition.

1. Introduction

At the point when we talked about the Automatic speech recognition system, we saw that the speech recognition system involves two parts the first one is extraction of the features from the audio file where the raw signal is passed through the set of the procedures known as the Mel- Frequency cepstral coefficients technique and the other one is classification of the input signal where we use the classifiers to classification purpose. We utilize the classifier for example, Gaussian Mixture Model, Hidden Markov Model, Artificial Neural Network, and Support Vector Machine well. there are varied feature extricating techniques are available for example, Linear Predictive Cepstral Coefficient, Mel frequency cepstral coefficients, Linear Predictive Coding, Perceptual Linear predictive coefficients and some more. The best among the different feature extraction strategies is Mel frequency cepstral coefficients.

In Mel Frequency cepstral coefficients extraction, we used to figure out the elements from a raw signal, The Mel frequency cepstral coefficients are certifiably not a solitary interaction it involves a different sub process like, pre-emphasis, framing, windowing, fast Fourier transform, Discrete cosine transform, and last one to improve the quality of the speech using dynamic features such as delta and delta-delta derivatives, Before the Mel frequency cepstral coefficients came into light, the other feature extricating techniques that Linear Predictive Cepstral Coefficient, Linear Predictive Coding, and zero crossing rate were utilized to the extricate the features from the raw or input signal [1].

The major objective of this paper to improve the quality of speech recognition system and enhance the frequency pitch at different stages. There are already various solutions are existed, but they do not enough to solve the distortion problems in the voice therefore our proposed technique will be more effective the other existing approaches.

After the consummation of the feature extraction strategy, the quality signal is passed to the classifier for the characterization purposes.

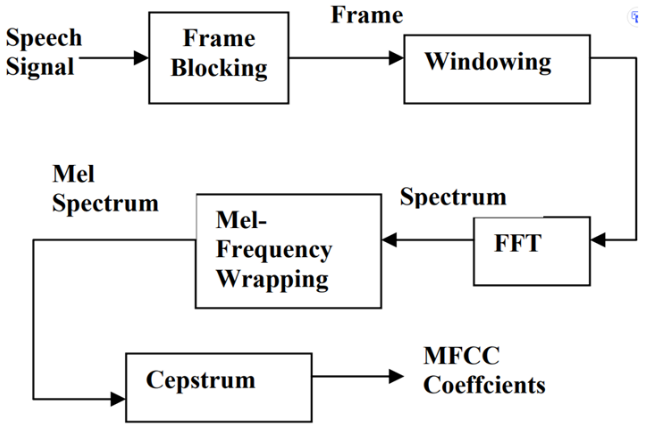

Fig. 1. Block Diagram of Mel Frequency Cepstral Coefficients.

2. Literature Review

Frequency spectral data is added to the traditional Mel spectrum-based speech recognition method. The Mel frequency technique takes use of the frequency observation for speech signals at a certain resolution, which leads to resolution feature overlaps and a recognition limit [22].

The feature extraction approach is a crucial tool for classifying the features of voices as well as for speech and speaker recognition. There are several kinds of feature extraction approaches, including PLP, LPC, various MFCC types, etc. In comparison to other forms of MFCC, Delta-Delta MFCC has superior Cepstral coefficients and the lowest standard deviation [23].

One of the most practical and well-known techniques in the area of signal processing is MFCCs. However, if the number of test dataset strange speakers is big, it requires a large training dataset and a lengthy execution time to extract the relevant characteristics. For this, based on the development of the operator which is adaptable to any input signal. An benefit of dealing with a sparse matrix where most of its elements are zero is that it makes the calculation and execution time of the method shorter [24]. This is true even if its formation goes through numerous iterations log2 N iterations where N is the length of the signal.

3. Mel Frequency Cepstral Coefficient Extraction stages

First, the signals is windowed, then the discrete Fourier transform (DFT) is applied, then logarithmic of the magnitude is calculated, the frequencies are warped using the Mel scale, and last, the inverse discrete Fourier transform is applied. This is the MFCC feature extraction method. Step-by-step instructions for exécution extracting MFCC features are provided below.

3.1. Pre-Emphasis

The initial phase in most component or feature calculation procedure is pre-emphasis [2]. It is the most often channel utilized in any speech recognition system to work on the nature of signal, as it will in general diminish the encompassing commotion and further develop the voice quality. The information signal to be inspected procured involving mouthpiece as the information source, and afterward Pre-emphasis channel is applied. The glottal impact is eliminated from the vocal parcel attributes by pre-emphasis.

Pre-emphasis is ordinarily a first-order high-pass channel that attempted to reestablish the amplitude of high-frequency regions. At the end of the day, Pre-emphasis supports the signals of high-frequency integration, whereas clearing out the low-frequency integration in their unique state. The Transfer function of the pre-emphasis factor α is

$$H(z)=1-\alpha z-1 \quad (1)$$

The most well-known scope of values for α is somewhere in the range of 0.95 and 0.97, in spite of the fact that qualities in the scope of 0.9 and only under 1.0 have additionally been utilized in various frameworks [3,4].

3.2. *Framing*

The subsequent stage is to do framing of the pre-emphasis signal acquired in the preceding stage. The speech signal is partitioned into little segments of length 20-30 ms with half cross-over to keep away from any deficiency of data. It accepts that over this brief length, speech signal remaining parts fixed or stationary. In past case as speech is non-fixed signal, it's challenging to manage this signal in consistent domain [5,6,7]. Generally, Framing is the overall term for the division of a voice signal into a short time frame length. Practically speaking, the voice signal is partitioned into N samples, with back-to-back frames isolated by M. Typically, the upsides of N and M are viewed as N=256 and M=100, individually.

3.3. *Windowing*

This is a significant stage wherein every individual frame is windowed to limit the discontinuities of the signal toward the start of the frame and toward the finish of the frame. In this, the standards used to limit the intermittence or otherworldly contortion is involving the window to tape the signal by a no worth toward the start and end of the frame. Windowing is performed on each frame with one of the well-known signal handling windows like the hamming window [8,9,10].

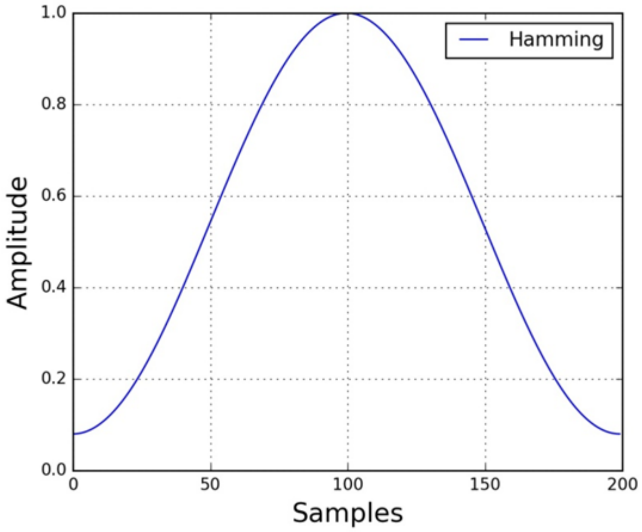

Fig. 2. Hamming window.

As talked about in the previous segment, there might be discontinuities of the signal at both the finishes. Altogether, to make the signal smoother, the frames procured in the previous methodology are multiplied by window function [11]. Different windowing strategies are applied like Hann windowing, Hamming Windowing and so forth, however for the most part windowing is finished by choosing the hamming window as opposed to the next windowing procedures.

$$Y(n) = X(n)W(n) \tag{2}$$

Hamming window is numerical addressed by the situation as given underneath.

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\Pi n}{(N-1)}\right), 0 \le n \le N-1 \tag{3}$$

Documentations,
W (n): Windowed frame
n: Current casing/frame
N: absolute number of casings/frames.

Generally, while playing out the Mel frequency cepstral coefficients feature extraction in MATLAB, at the windowing stage, windowing is made more straightforward with the predefined work hamming accessible in MATLAB which complete the entire course of windowing.

3.4. Fast Fourier Transform

The Discrete Fourier Transform of a sequence is computed using a Fast Fourier Transform. Fourier analysis transforms signal's original domain typically time domain into frequency domain representation [12].

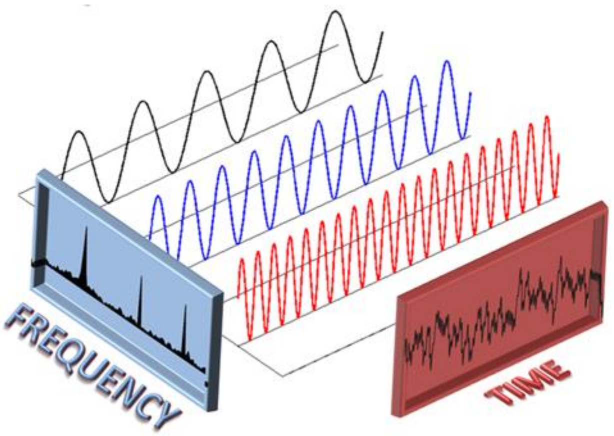

Fig. 3. Fast Fourier Transform.

3.5. Mel filter banks processing

After the signal changed from time domain to frequency domain, this progression required for assessment of Mel-filter banks. The human hearable framework is unequipped for getting discourse on a direct scale [13].

The Mel scale depends on how individuals see pitch. Mel Scale Channels are triangular, straight beneath 1 KHz and logarithmic above 1 KHz. The Mel Scale filter bank is displayed in Fig.6 [14]. A progression of band pass channel is utilized in the filter bank investigation. The channel bank is an assortment of covering triangular band pass channels organized in a Mel frequency scale [15,16].

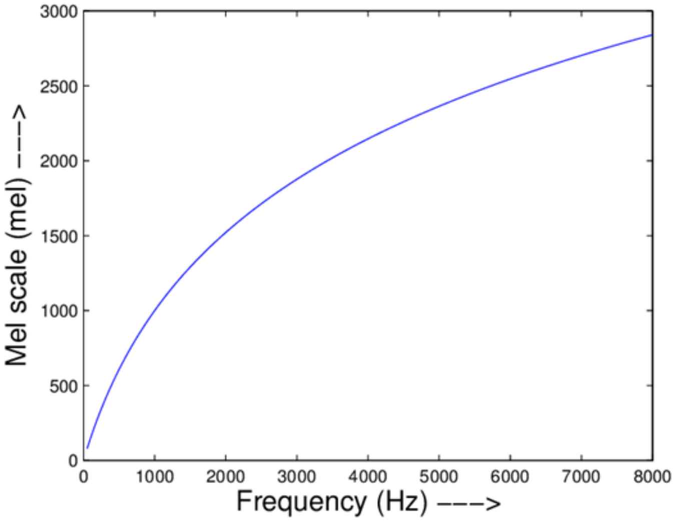

Fig. 4. Relationship between frequency scale and Mel scale.

$$f_{mel} = 2595\log_{10}\left(1+\frac{f}{100}\right) \tag{4}$$

The condition of above shows the connection between Mel scale frequencies and direct frequencies. For MFCC calculation, filter banks are for the most part carried out in frequency domain. The greatness of recurrence reference of each channel is triangular, and the result of each filter is the amounts of its separated range parts [17].

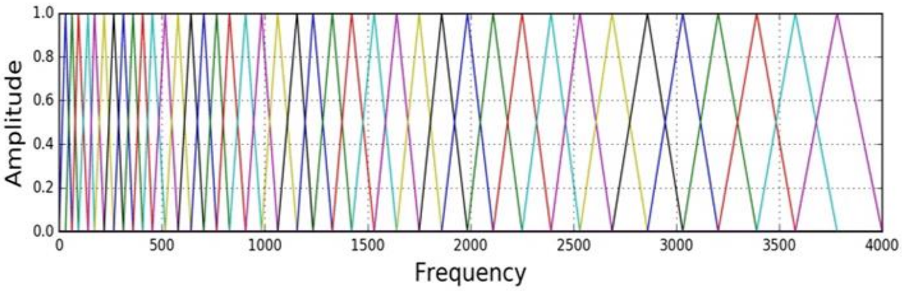

Fig. 5. Filter bank on Mel scale.

3.6. *Cepstrum*

The Mel Spectrum is eventually translated or converted to time domain. Mel frequency Cepstral Coefficients are the end consequence of this process.

For the predetermined casing investigation, the cepstral depiction of the discourse range precisely addresses the signals of the neighborhood otherworldly highlights. The discrete cosine transform may be used to transfer the Mel spectrum coefficients into the time domain because they are real values [20].

$$C(n) = \sum_{n=0}^{M-1} \log_{10}(s(m)) \cos\left(\frac{\Pi n(m-0.5)}{M}\right) \tag{5}$$

Discrete Cosine Transform removes the pitch contribution. The log Mel spectrum in the long run gets changed over to time domain in the last stage. Mel frequency cepstral coefficients are the term referring to the outcome [18,19].

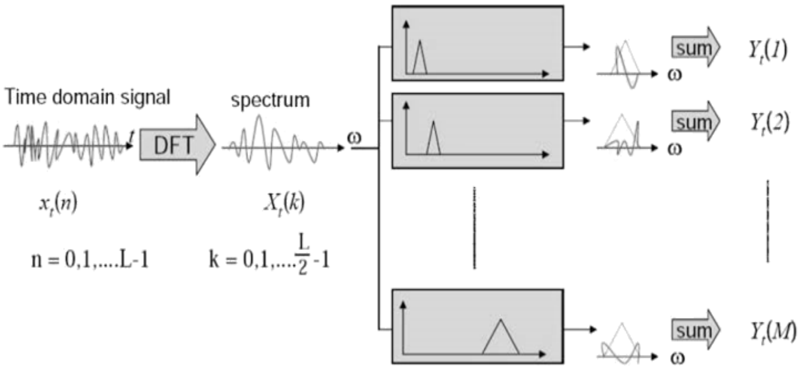

Fig. 6. Mel-filter creation.

3.7. *Dynamic Features: Delta Derivatives*

It is feasible to acquire more subtleties speech highlight by utilizing a determination on the Mel frequency cepstral coefficients highlights, this approach allows the calculation of the first order delta and then the second order delta derivatives [21] that are gotten from the first order delta derivatives.

The motivation behind applying a delta and the delta-delta derivatives or subsidiaries is to further improve the voice recognition. Differential and speech increase coefficients are alluded to as delta and the delta-delta coefficients.

$$d_{t,k} = \frac{\sum_{n=1}^{N} n\left|c(t+n,k) - c(t-n,k)\right|}{2\left|\sum_{n=1}^{N} n^2\right|} \tag{6}$$

Where, dt signifies the delta coefficients assessed with regards to static coefficients at time t. The speed increase or acceleration or delta-delta coefficients [21] are registered similarly, however rather than using the static coefficients, the differential coefficient is utilized.

As demonstrated in the accompanying conditions, the first order delta derivatives and double delta derivatives not entirely set in stone. The delta coefficient portrays the discourse rate, though the double delta coefficients depict the data that is connected with discourse speed increase.

4. Results

In wake of executing each course of extricating the Mel-frequency Cepstral Coefficients let us look at the representation of MFCC Coefficients. It is somewhat looks like a bidirectional cluster or the framework. The representation of the MFCC is displayed in the accompanying delineation that were elapsing the inspecting rate and the y axis have different coefficients. Presently we are planning between the varieties and mathematical qualities.

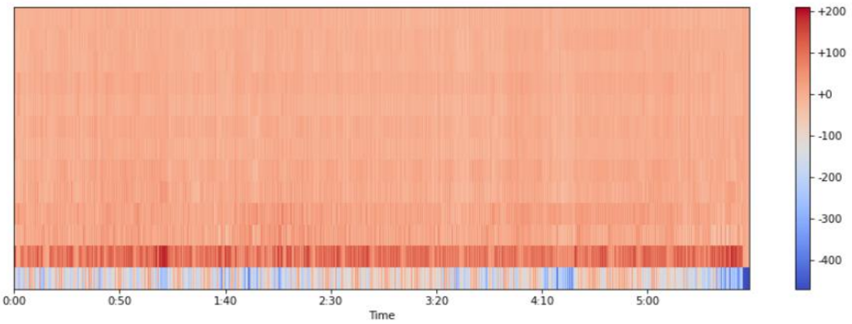

Fig. 7. MFCC Extracted

In this segment, we can envision the delta and double delta coefficients concept displayed in the representation after the utilization of the delta derivatives on the Mel frequency cepstral coefficients. we get the visualization of the MFCC after working on the MFCC coefficients. we will get the first and the second order derivatives of MFCCs.

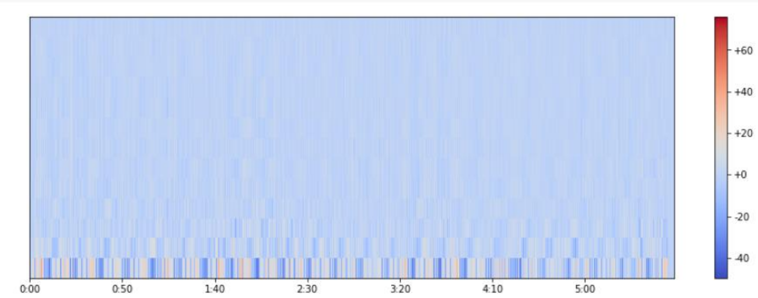

Fig. 8. MFCC Delta -First Order Delta Coefficient Extracted visual

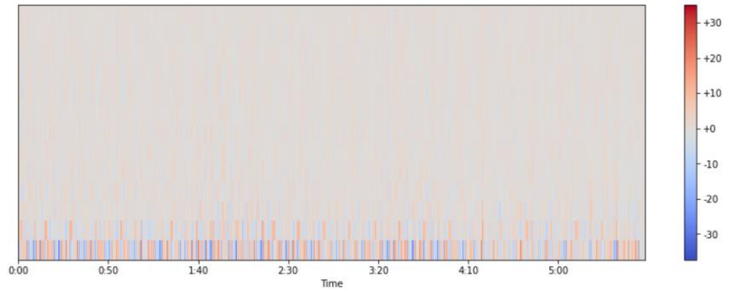

Fig. 9. MFCC Delta -Second Order Delta Coefficient Extracted visual

After using the Mel-frequency cepstral coefficients technique we find that the distortion in second order delta coefficient is very less and good performance. Therefore, the pitch of the speech is easily recognized at every stage of the frequency of voice.

5. Conclusion

We have endeavored to introduce a point-to-point investigation of Mel frequency cepstral coefficients technique in this paper, which will help the specialists or researcher in facilitating their comprehension in this domain. The fundamental objective of this paper is to give a concise outline of the means associated with extricating features from raw signals through a set of procedures. A great deal of examination is being done in this field to work on the presentation of speech recognition systems. Overall, the paper emphasize the clarity of speech and easily recognize the voice with the help of MCC techniques. In future there will be use spectrogram windowing with the hybrid algorithm of deep learning techniques that may be more improve the performance of the automated speech recognition system.

References

[1] Bharthi B, Deepalakshmi V, NelsonI. A neural network-based speech recognition system for isolated Tamil words. In Proceedings of International Conference on Neural Networks and Artificial Intelligence, Brest, Belarus 2006 Jun.

[2] Rajput N, Nanavati AA. Speech in Mobile and Pervasive Environments. John Wiley & Sons; 2012 Jan 26.

[3] Du JX, Guo YL, Zhai CM. Recognizing Complex Events in Real Movies by Audio Features. In International Conference on Intelligent Computing 2012 Jul 25 (pp. 218-223). Springer, Berlin, Heidelberg.

[4] Beigi H. Speaker recognition. In Fundamental of Speaker Recognition 2011 (pp.543-559). Springer, Boston, MA.

[5] Pangaonkar S, Panat A. A Review of Various Techniques Related to Feature Extraction and Classification for Speech Signal Analysis. ICDSMLA2019. 2020 (pp.534-549).

[6] Dhonde SB, Chaudhari A, Jagade SM. Integration of Mel-frequency cepstral coefficients with log energy and temporal derivatives for text-independent speaker identification. In Proceedings of the international Conference on Data Engineering and Communication Technology 2017 (pp. 791-797). Springer, Singapore.

[7] Mohdiwale S, Sahu TP. Nearest Neighbor Classification Approach for Bilingual Speaker and Gender Recognition. In Advances in Biometrics 2019 (pp. 249-266). Springer, Cham.

[8] Goyal S, Batra N, Batra NK. An integrated Approach to Home Security and Safety Systems. CRC Press; 2021 Oct 14.

[9] El-Samie FE. Information security for automatic speaker identification. Information security for automatic speaker identification. 2011:1-22.

[10] Kalpana Chowdhary M, Jude Hemanth D. Deep Learning Approach for Speech Emotion Recognition. In Data Analytics and Management 2021 (pp. 367-376). Springer, Singapore.

[11] Fayek H. Speech processing for machine learning: Filter banks, Mel-frequency cepstral coefficients MFCCS) and what's in between. URL: https://haythamfayek.com/2016/04/21/speechprocessingfor-machine-learning.html. 2016 Apr.

[12] Tree Spirit: Illegal logging detection and alerting system using audio identification over an IoT network-Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Fast-Fourier Transformation-2-Fast-Fourier-Transformation-To increase-the-performance_fig6_323281289[accessed Mar 6, 2022].

[13] Srivastava S, Chaudhary G, Shukla C. Text-Independent Speaker Recognition Using Deep Learning. In Concepts and Real-Time Applications of Deep Learning 2021 (pp. 41-51). Springer, Cham.

[14] Li L, Li Y, Wang Z, Li X, Shi G. A Reliable Voice Perceptual Hash Authentication Algorithm. In International Conference on Mobile Multimedia Communications 2021 Jul 23 (pp. 253-263). Springer, Cham.

[15] Kamble VV, Deshmukh RR, Karwankar AR, Ratnaparkhee VR, Annadate SA. Emotion recognition for instantaneous Mrathi Spoken Words. In Proceedings of the 3rd international Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 2015 (pp. 335-346)

[16] Music and Speech Analysis Using the 'Bach' Scale Filter Bank- Scientific Figure on ResearchGate. Available from: https://wwww.researchgate.net/figure/The-Mel-scale-as-a-linear-and-b-semilog-plots_fig3_282609758 [accessed 6 Mar 2022].

[17] A Mel-Filter bank and MFCC-based Neural Network Approach to Train the Houston Toad Call Detection System Design-Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/An-example-of-Filter-Bank-on-Mel-Scale-12_fig3_330477843 [accessed Mar 6, 2022].

[18] Isolated speech recognition using MFCC and DTW. International Journal & Magazine of Engineering, Technology, Management and Research. April 2017.

[19] Mohamad Jamil MH, AI-Haddad SA, Kyun Ng C. A flexible speech recognition system for cerebral palsy disabled. In International Conference on Informatics Engineering and Information Science 2011 Nv 14 (pp. 42-55). Springer, Berlin, Heidelberg.

[20] Coskun H, Yigit T. Artificial Intelligence Applications on Classification of Heart Sounds. In Nature-Inspired Intelligent Techniques for Solving Biomedical Engineering Problems 2018 (pp. 146-183). IGI Global.

[21] Kasiviswanathan U, Kushwaha A, Sharma S. Development of human speech signal-based intelligent human-computer interface for driving a wheelchair in enhancing the quality -of-life of the persons. In Intelligent Systems for Healthcare Management and Delivery 2019 (pp. 21-60). IGI Global.

[22] Patel, I., 2010. Speech recognition using HMM with MFCC-An analysis using frequency specral decomposion technique. *Signal & Image Processing: An International Journal (SIPIJ) Vol*, *1*.

[23] Ranjan, R., & Thakur, A. (2019). Analysis of feature extraction techniques for speech recognition system. *International Journal of Innovative Technology and Exploring Engineering*, *8*(7C2), 197-200.

[24] Abakarim, F., & Abenaou, A. (2022). Comparative study to realize an automatic speaker recognition system. *International Journal of Electrical and Computer Engineering*, *12*(1), 376-382.

Authors' Profiles

Gaurav D. Saxena currently pursuing the master's in computer science from Kamla Nehru Mahavidyalaya, Nagpur, Maharashtra. He is currently indulged in research in the field of artificial Intelligence and Machine Learning. He has presented various papers in International and National Conferences, with one paper in conference proceedings, and other papers are published in esteemed journals. He has also filed four Indian patents. He is also working as a peer reviewer of various UGC internal journals.

Nafees A. Farooqui was born at Siddharth Nagar, Uttar Pradesh, India on January 01, 1984. He had completed BSc (Hons) in Statistics from AMU, Aligarh, MCA from Integral University, Lucknow, UP and PhD (CS) from DIT University, Dehradun, UK. He is having more than 13 years of experience and presently he is working as Assistant Professor in the Department of Computer Science, Era University, Lucknow, UP, India. His research interests include Data Mining, Machine Learning, Data Science and Artificial Intelligence. He has 25 publications in a total of which are 13 journals, 1 book and 1 book chapter’s publication, and 10 are conferences publication. He is a member of International Association of Engineers (IAENG) since 2017, ACM since 2011. He guided various projects of UG and PG level Students. He received awards from various professional bodies.

Saquib Ali had completed MCA from Integral University, Lucknow UP. He is having 4-year industrial experience and 3-year academic experience. He is NET qualified. Presently he is working as Assistant Professor in the Department of Computer Science, Era University, Lucknow UP, India. His research interest including Data Mining, Machine Learning and Artificial Intelligence. He has published four research paper in reputed journals and conferences. He is active academician and good skill in the content writing. His Two book chapter in under publication.

How to cite this paper: Gaurav D. Saxena, Nafees A. Farooqui, Saquib Ali, "Extricate Features Utilizing Mel Frequency Cepstral Coefficient in Automatic Speech Recognition System", International Journal of Engineering and Manufacturing (IJEM), Vol.12, No.6, pp. 14-21, 2022. DOI:10.5815/ijem.2022.06.02