Real Time FPGA Implementation for Speech Feature Extraction

Aditya Deshmukh
Student of E&TC
PCCOER
Pune, India
Aditya.Deshmukh_entc21@pccoer.in

Arman Khan
Student of E&TC
PCCOER
Pune, India
Arman.Khan_entc21@pccoer.in

Vedika Patil
Student of E&TC
PCCOER
Pune, India
Vedika.patil_entc21@pccoer.in

Kishor Bhangale
Assistant Professor
PCCOER
Pune, India
Kishor.Bhangale@pccoer.in

Abstract - This project aims to develop a system for extracting Mel-Frequency Cepstral Coefficients (MFCCs) from raw audio signals, and subsequently transmitting this feature data via USB to a laptop for further processing. MFCCs is widely used features in speech recognition, speaker identification, and other audio-based applications due to their ability to capture essential characteristics of the human voice. By extracting these features efficiently and transmitting them to a more powerful computing platform, this project seeks to provide a foundation for advanced speech processing tasks. The system will consist of an embedded platform (here, FPGA). It will be responsible for acquiring audio signals, performing pre-processing, extracting MFCC feature, and transmitting the feature data via USB to the laptop.

Keyword: Mel frequency cepstral coefficient, speech recognition

I. INTRODUCTION

Speech processing is used in the day-to-day life of every one of us. It is used when we communicate with our mobile phones through “Siri” or “Ok Google”, etc. It is also used when using speech enhancement during studio recordings, podcasts, music production and Artificial Intelligence and Machine Learning applications such as music classification, segregation, etc. MFCC is widely used feature in speech recognition, speaker identification and other audio- based applications due to their ability to capture essential characteristics of the human voice with subtle details. Extracting these features efficiently and transmitting them to a more powerful computing platform provides a foundation for advanced speech processing tasks. Current research in the domain of Digital Signal Processing and Speech Processing are aimed at efficient implementation of Mel Frequency Cepstral Coefficients (commonly known as MFCC) for deep learning applications. The MFCC models after the human auditory capabilities to describe the trends in speech signal. It has a wide variety of applications. This paper investigates the MFCC extraction techniques and its implementations. The paper is aimed at understanding the algorithms behind feature extraction specifically MFCC. The need for precise and efficient audio processing has grown as speech recognition technologies have developed. These systems are necessary for applications such as accessibility tools and virtual assistants. The computationally demanding job of extracting pertinent features from audio signals—in particular, the computation of Mel- Frequency Cepstral Coefficients (MFCCs), which capture speech characteristics—remains a significant difficulty.

II. METHODOLOGY

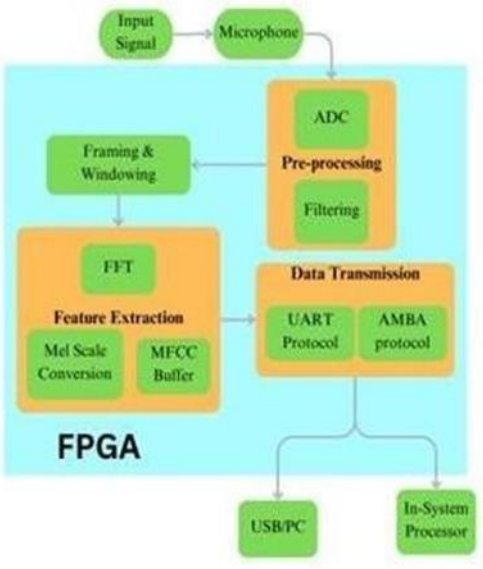

Fig 2.1. Block Diagram of proposed system

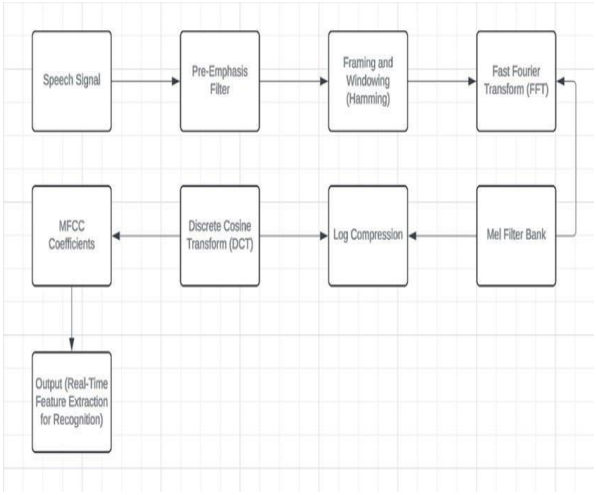

Fig 2.2. Process Flow

A speech is the combination of sounds which depict meaning when heard by a listener. It is produced by the vibration of vocal cords in humans. The audible frequency range for humans is 20Hz to 20kHz. The fundamental frequency range of human speech is 350Hz to 3kHz. Also, from 3.5kHz to 17kHz is called the Harmonics. Thus we choose a mono-channel microphone supported for the frequency range of at least 3.5kHz. The microphone is connected to the FPGA board via the audio port. The audio port is configured to take the input and feed it to the onboard XADC which will digitize the signal at a sampling rate near 8kHz. A speech is the combination of sounds which depict meaning when heard by a listener. It is produced by the vibration of vocal cords in humans. The audible frequency range for humans is 20Hz to 20kHz. The fundamental frequency range of human speech is 350Hz to 3kHz. Also, from 3.5kHz to 17kHz is called the Harmonics. Thus, we choose a mono- channel microphone supported for the frequency range of at least 3.5kHz. The microphone is connected to the FPGA board via the audio port. The audio port is configured to take the input and feed it to the onboard XADC which will digitize the signal at a sampling rate near 8kHz.

A. INPUT DEVICE

This includes a dedicated hardware for Capturing the input signals within the necessary parameter range such as frequency, number of channels, etc. The device is selected by considering the human speech parameters such as frequency range, psychoacoustics, etc.

B. PREPROCESSING

The initial stage of the system involves the pre- emphasis of the received digital audio input. This stage involves selective filters to remove unwanted noises in the signal. Here, the unwanted frequencies include ambient noise, humming noise, and harmonics. The hum noise lies in the range of 100-500Hz. Thus, we want all the frequency components above this range. We use a first order High Pass Filter in this stage to remove the hum. The cut-off frequency is set to 500Hz with a gain of 6dB.The filter design must be chosen so that it gives fast roll-off and utilizes less hardware resources. It is a First Order High Pass filter which uses a set of 100-150 weighted taps which remove the hum noise at lower frequency ranges. The filter is designed using Verilog and Vivado. The algorithm for the filter includes the flash memory which holds the filter weight taps. The filter module is connected using custom data path.

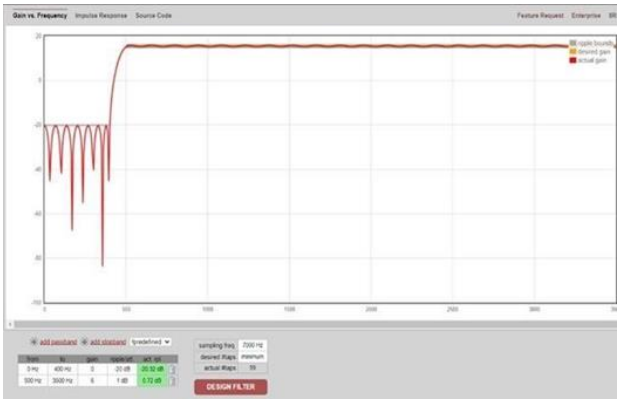

Fig 2.3. Filter Design

C. FILTERING

This stage consists of a filter component which is designed to select only necessary components within the signal. The filter design involves parameter considerations and accuracy requirements. The most commonly used filters include the Finite Input Response (FIR) filter with pre-determined weights for the delays and buffers and Infinite Input Response (IIR) filter with pre-defined weights and error correction with feedback system from the past outputs.

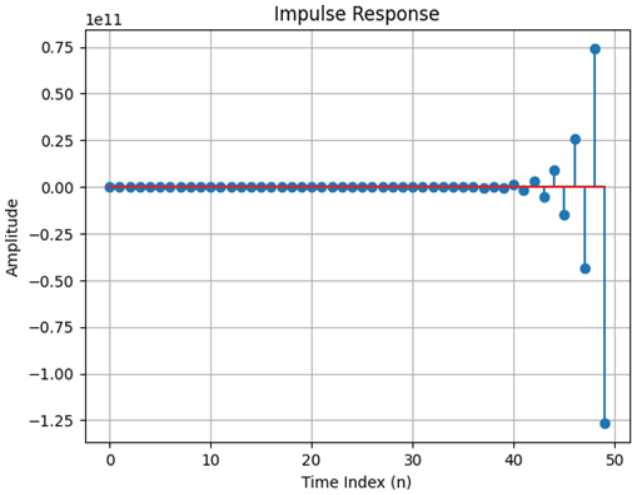

Fig 2.4. Filter design analysis

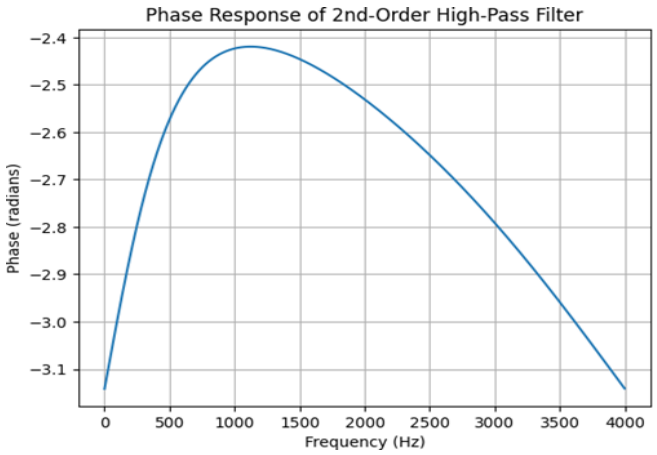

Fig 2.5 Filter design analysis

D. FRAMING

Framing is the segmentation of the signal to make it easier to analyze. It is generally determined in terms of milliseconds. The framing duration depends on the average duration of the phonetics in the language. In English, the average duration of each phoneme is nearly 50-100ms. Thus, the standard frame duration of Around 30ms is considered for most audio processing applications. After the framing is done, a window function is multiplied to the signal to get individual blocks of signal which can be easily analyzed. The windowing is done to reduce the spectral leakage of the frame after abrupt truncation of the signal. The windowing function can be of different types such as rectangular, triangular, hanning, and hamming, etc. According to the speech processing application, the choice of window function changes This stage involves dividing the digital signal into short overlapping signals. The overlapping is mostly 50the edges of the signal to a more integral value and reduce possible chances of errors between frames.

E. FAST FOURIER TRANSFORM

The data controller sends the data after pre-processing to the FFT compute module. The FFT algorithm considered in the proposed system is split-radix. The split radix is a combination of radix-2 and radix-4 FFT algorithms. This algorithm combines the strengths of both radix-2 and radix-4 algorithms. It is more efficient for wider range of input lengths which are not a power of 2 or 4. Discrete Cosine Transform is used for domain conversion of the signal from time domain to frequency domain of a cosine signal. This domain conversion is effective for data compression. The processed signal values are converted into a Mel scale where they are approximated to the required and understandable data formats. These values are the final output of the processor.

F. CUSTOM DATA PATH

The main novelty for the proposed system is the custom data path and data controller designed on the FPGA fabric. The FPGA board used includes a dual core ARM cortex A9 which is integrated in the fabric. The data path is responsible for moving the data through the entire hardware system. The data path is optimized for low latency and large data manipulation.

G. HARDWARE SPECIFICATIONS

The proposed solution is an attempt for ASIC development focusing on leveraging the DSP aspects involved in deep learning models. The system spreads out the DSP components onto the FPGA fabric for faster computation and reduced load on processor for instruction-based applications.

TABLE 2.1 SPECIFICATIONS

Sr. No	**Hardware**	**Specification**
1	FPGA/SOC	Zynq-7000 series
2	Processor architecture	Dual-core Cortex A9
3	Clock	650MHz
4	Logic Slices	13,300
5	DSP slices	220
6	Power Supply	7V-15V external supply
7	Audio Input	Microphone with PDM interface
8	Debugger	JTAG/Quad-SPI Flash
9	Supported Development Environment	Xilinx Vivado, Vitis Unified IDE, Pynq Python-compatible library
10	Additional features	On-chip analog-to-digital converter (XADC)
11	Peripheral Communication	AXI3 Protocol

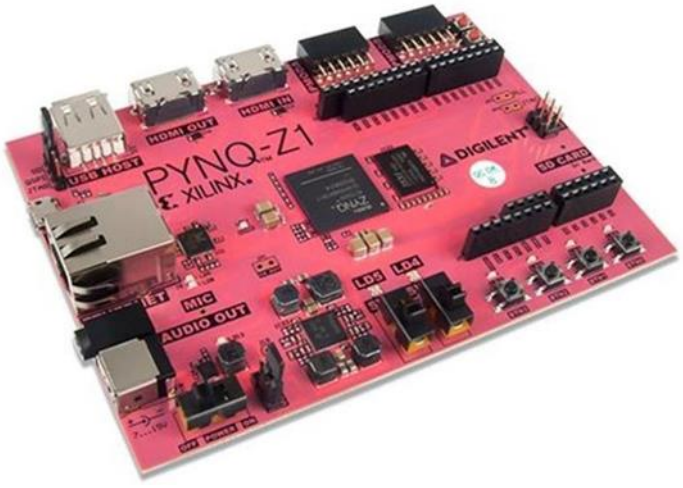

Fig 2.6. Pynq Z1 FPGA Board

III.FINDINGS FROM LITERATURE SURVEY

1. Jinghong1, Tian Yanan1, Zhang Lijia1 [1]: The suggested approach incorporates key processes for speaker recognition, such as Mel Frequency Cepstral Coefficient (MFCC) extraction for feature representation and Voice Activity Detection (VAD) for speech segment identification. Vector Quantization (VQ) recognition is then utilized to identify the speaker using the retrieved MFCC features. Techniques like pipelining, module reuse, and Ping- Pong operation are used to maximize the system's performance on FPGA, enabling high-speed and parallel processing.

2. Zhu, Jianchen, and Zengli Liu [2] The suggested Approach incorporates key processes for speaker recognition, such as Mel Frequency Cepstrum Coefficient (MFCC) extraction for feature representation and Voice Activity Detection (VAD) for speech segment identification. Vector Quantization (VQ) recognition is then utilized to identify the speaker using the retrieved MFCC features. Techniques like pipelining, module reuse, and ping-pong operation are used to maximize the system's performance on FPGA, enabling high-speed and parallel processing. For speech recognition applications that demand low latency, this makes it possible for the system to manage real-time data streams effectively.

3. Sujuan Ke, Yibin Hou, Zhangqin Huang, Hui Li[3]: Creating an embedded voice recognition system on an FPGA platform is the research methodology. With an emphasis on creating an IP core to execute the forward algorithm—a crucial step in the HMM process—the system uses the HMM algorithm for voice recognition. In contrast to conventional software-based methods, the implementation of this algorithm on specialized hardware—more especially, an FPGA—is meant to shorten processing times. The forward algorithm is designed and simulated in the FPGA environment as part of the technique. The system's efficacy for real-time, low-latency speech recognition is then evaluated by experimental testing and analysis of the results.

4. Junchang Zhang,Yuanyuan Chen,Jian Zhang [4] First, the method divides speech signal frames into several clusters according to their similarities using Kernel-based Fuzzy K-means Clustering (KFKC). KFKC is more adaptable and more suited for complicated or noisy speech data since it permits each frame to have a degree of membership to many clusters, in contrast to classic clustering algorithms. Following clustering, the clustered frames are subjected to Kernel Principal Component Analysis (KPCA), which lowers their dimensionality and mitigates noise. The quality of the retrieved features is enhanced by this two-step procedure, which increases their suitability for application in subsequent tasks like voice recognition.

5. S J Melnikoff, S F Quigley M J Russell [5]: Using an FPGA, the Viterbi decoding technique is used to speed up the decoding of continuous Hidden Markov Models (HMMs), which are used for speech recognition. When modeling monophonies, continuous HMMs are used, with each state in the model representing a distinct phoneme. To increase speed and efficiency, the Viterbi algorithm which is essential to HMM-based voice recognition—is tailored for FPGA hardware. The system can process voice data in real time since the FPGA implementation lowers the temporal complexity of decoding. A continuous voice recognition task is used to test this method, and Accuracy and processing speed are assessed. Advantage: High Processing Speed: Achieves 75 times real-time processing speed, demonstrating significant acceleration over software implementations

6. W Liu, Q Liao, F Qiao, W Xia, C Wang, F Lombardi [6]: In order to achieve a dynamic balance between the precision of the transform and the FPGA implementation's performance, the paper presents two algorithms that modify the word length utilized in FFT phases. By decreasing the precision of less important FFT stages, the word length adjustment seeks to lessen computational complexity while maintaining adequate accuracy for important components. An FPGA platform is used to construct the algorithms, and a detailed assessment of the hardware utilization—including processing speed and resource consumption is conducted. Performance criteria including accuracy trade-offs and speedup are taken into account while evaluating the efficacy of the suggested techniques

7. Joysingh, S.J., Vijayalakshmi, P. and Nagarajan [7]: The process entails substituting the chirp magnitude spectrum, which is based on the Chirp Z- transform, for the Fourier transform magnitude spectrum, which is commonly employed in conventional MFCC extraction. This modification attempts to better capture the signal's underlying spectral characteristics, especially when non-stationary components like music and speech are present. Three different real-world tasks—speech- music classification, speaker identification, and speech command recognition—are used to calculate and assess the Chirp MFCC. In order to evaluate performance, the study compares the accuracy and recognition reliability of Chirp MFCC with classical MFCC and analyzes class separation using the product of likelihood Gaussians

8. S Misra, TK Das,P Saha,U Baruah,RH Laskar [8]: The accuracy and decision time of MFCC and LPCC for speaker verification on an uncontrolled environment database are compared in this work. Whereas LPCC coefficients are acquired via linear predictive modeling of the articulatory process, MFCC coefficients are produced from the warped frequency scale intended to correspond with human auditory perception. Accuracy and equal error rate (EER) were the two criteria used to assess the systems' performance during speaker authentication testing. In order to assess the

effectiveness of MFCC and LPCC, the amount of time needed for the system to reach a decision was also measured. To comprehend the shortcomings of the method, the misclassification of real and fake samples was examined

9. TRJ Kumari, HS Jayanna [9]: The work deals with text-independent speaker verification using speech data that is shorter than 15 seconds. Gaussian Mixture Model (GMM) and GMM-Universal Background Model (UBM) approaches are used to model the features after they have been retrieved using Mel- Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). The NIST- 2003 database is used in experiments to assess performance under constrained training and testing circumstances.

10. AO Salau, S Jain [10]: The study examines a broad variety of feature extraction methods used in image-related fields such computer vision, data mining, image processing, and image retrieval. It focuses on how to extract important features from photos using Generalized Local Derivative Statistics (GLDS). Contrast, homogeneity, entropy, mean, and energy are important characteristics that may be extracted using GLDS. These features each reveal important details about the texture and structure of the image. The study examines the computation of these features, their importance in tasks such as image identification and classification, and an assessment of their applicability in other fields.

11. C Xu, W Rao, ES Chng, H Li [11]: The study introduces a time-domain speaker extraction network (SpEx) that extracts a target speaker's voice from a mixture using multi-scale embedding coefficients, avoiding traditional phase estimation. The SpEx network comprises four components: a speaker encoder, a speech encoder, a speaker extractor, and a speech decoder, facilitating efficient reconstruction of the target speaker's speech.

12. Heriyanto, Heriyanto, and Dyah Ayu Irawati[12]: In this work, the Dominant Weight Normalization (NBD) approach was used for feature selection after MFCC was used for feature extraction. The frames were selected from 0 to 10 or 11 and the cepstral coefficients were selected from 0 to 24. For testing, 300 voice samples in all, captured in 16-bit stereo at 44.1 kHz, were employed. Finding the frame and feature selection combination that would optimize speech recognition accuracy—especially in Shahada recitations—was the goal of the study. The accuracy results for various frames and selections were compared in order to assess the method's performance and outcomings of the method, the misclassification of real and fake samples was examined.

IV. GAP IDENTIFICATION

Although Convolutional Neural Networks (CNNs) and Mel-Frequency Cepstral Coefficients (MFCCs) are frequently used in Speech Emotion Recognition (SER), many previous studies have not fully investigated the potential of combining multiple emotional speech datasets for a more thorough assessment of emotion classification models. Additionally, even though MFCCs work well for feature extraction, the impact of sophisticated preprocessing methods such data augmentation, standardization, and model fine-tuning on enhancing the precision of emotion recognition has not received enough attention. By combining a number of reputable emotional speech datasets, such as CREMA-D, SAVEE, EMO-DB, TESS, and RAVDESS, and implementing strict preprocessing techniques, this study closes the gap. By doing this, it shows how combining several datasets with well-tuned CNN models can greatly improve the ability to identify emotional nuances in speech, resulting in a more reliable and broadly applicable SER system. It also tackles the problem of enhancing SER performance across a range of speaker attributes and emotional expressions, which is frequently a drawback of current models.

V. FUTURE SCOPE

Speech Processing and Machine Learning: Implementing the system for optimized audio processing workloads like speech recognition, speaker identification. Integration with broader AI frameworks: Integrating the design with popular AI frameworks like TensorFlow or PyTorch to simplify its use. Exploring new applications: Applying the system to emerging AI domains, such as natural language processing or robotics. Edge device: Further tailoring the system to specific applications, such as biometric systems and fault detection systems. Scaling up: Investigating ways to scale the system for larger- scale calculation, diverse mathematical computations like convolution, LPCC, etc.

VI. CONCLUSION

By efficiently offloading the speech feature extraction components from conventional speech processing models, the suggested approach greatly lessens the demand on host systems' resources. The system combines several critical functions, including as sampling, filtering, framing, windowing, Discrete Cosine Transform (DCT), Fast Fourier Transform (FFT), and Mel-Frequency Cepstral Coefficients (MFCC) computation, by utilizing specialized DSP IP cores on an FPGA. Every step of the process flow, from recording audio signals within the required frequency ranges to using sophisticated filtering techniques to guarantee that only pertinent components are evaluated, is painstakingly designed to maximize performance. The system's capacity to efficiently convert and compress audio data is further improved by the use of FFT and DCT for signal processing.

The resultant output is a set of MFCC data that can be easily integrated with more robust processing platforms for additional analysis. These results can be sent over UART or AMBA protocols to a laptop or onboard system. This all-inclusive method improves speech recognition systems' accuracy and responsiveness while also streamlining the administration of audio data.

As a result, this study opens the door for more effective and efficient speech processing applications by showcasing the many benefits of using FPGA technology for speech feature extraction. Subsequent improvements can concentrate on optimizing the algorithms and investigating new functionalities to further improve system efficiency and suitability for other audio-related domains.

REFERENCES

[1]li, . T. Yanan and Z. Lijia, "Research and implementation of speaker recognition algorithm based on FPGA," *In 2012 24th Chinese Control and Decision Conference (CCDC),IEEE*, pp. 1155- 1158, 2012.

[2] Zhu, Jianchen and Z. Liu, " Analysis of hybrid feature research based on extraction LPCC and MFCC," *In 2014 Tenth International Conference on Computational Intelligence and Security,IEEE*, pp. 732-735, 2014.

[3]Sujuan ke, Yibin hou, Zhangqin Huang and Hui Li, "A HMM speech recognition system based on FPGA.," In 2008 Congress on Image and Signal Processing,IEEE, vol. 5, pp. 305-309, 2008.

[4]Junchang Zhang, Yuanyuan Chen and Jian Zhnag, "Speech feature extraction of KPCA based on kernel fuzzy K-means Clustering," *In 2011 International Conference on Computer Science and Service System (CSSS),IEEE*, pp. 756-759, 2011

[5] SJ Melnikoff, S F Quigley and M J Russell, "Implementing a simple continuous speech recognition system on an FPGA," *In Proceedings. 10th Annual IEEE Symposium on Field-Programmable Custom Computing Machines,IEEE*, pp. 275-276, 2002

[6]W Liu, Q Liao, F Qiao, W Xia, C Wang and F Lombardi, "Approximate designs for fast Fourier transform (FFT) with application to speech recognition. IEEE Transactions on Circuits and Systems I: Regular Papers," *IEEE Transactions on Circuits and Systems I: Regular Papers* , pp. 4727-4739, 2019

[7] Joysingh, S.J Vijayalakshmi and P Nagarajan, ""Chirp Group Delay-Based Onset Detection in Instruments with Fast Attack (2023): 1639-1662.," *Circuits, Systems, and Signal Processing 42*, pp. 1639-1662., 2023

[8] S Misra, P Saha, U Baruah, Rahul H Laskar and Tushar KantiDas, "Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis.," *In 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015],IEEE*, pp. 1

[9] Kumari, TR Jayanthi and H. S. Jayanna. , "Comparison of LPCC and MFCC features and GMM and GMM- UBM modeling for limited data speaker verification," *In 2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-6, 2016.

[10] Salau, Ayodeji Olalekan and Shruti Jain, "Feature extraction: a survey of the types, techniques, applications.," 10 In 2019 international conference on signal processing and communication (ICSC), IEEE, pp. 158-164, 2019.

[11] Xu, Chenglin, Wei Rao, Eng Siong Chng and Haizhou Li, "Spex: Multi-scale time domain speaker extraction network," IEEE/ACM transactions on audio, speech, and language processing 28, pp. 1370-1384, 2020.

[12]Heriyanto and Dyah Ayu Irawati. , "Comparison of Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction, With and Without Framing Feature Selection, to Test the Shahada Recitation.," In RSF Conference Series: Engineering and Technology, vol. 1, pp. 335-354, 2021.