iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46807

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 10 Issue IX Sep 2022- Available at www.ijraset.com

Design and Implementation of Energy-Efficient Floating Point MFCC Extraction Architecture for Speech Recognition Systems

Srinija Lankala[1], Dr. M. Ramana Reddy[2]
[1]*UG student,* [2]*Assistant Professor Department of Electronics and Communication Engineering, Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India*

Abstract: ***This brief presents an energy-efficient architecture to extract mel-frequency cepstrum coefficients (MFCCs) for real-time speech recognition systems. Based on the algorithmic property of MFCC feature extraction, the architecture is designed with floating-point arithmetic units to cover a wide dynamic range with a small bit-width. Moreover, various operations required in the MFCC extraction are examined to optimize operational bit-width and lookup tables needed to compute nonlinear functions, such as trigonometric and logarithmic functions. In addition, the dataflow of MFCC extraction is tailored to minimize the computation time. As a result, the energy consumption is considerably reduced compared with previous MFCC extraction systems.***

Keywords: Mel-frequency cepstrum coefficients (mfccs), speech recognition, floating-point operations, hardware optimization.

I. INTRODUCTION

Increasingly, as they are applied in real world word applications, speech recognition systems must operate in situations where it is not possible to control the acoustic environment. This may result in a serious mismatch between the training and test conditions, which often causes a dramatic degradation in performance of these systems. The aim of the work presented in this thesis is to make automation speech recognition systems robust to these environmental differences.

Speech can be characterized by a slowly changing spectral envelope. This spectral envelope is perceived by humans and translated into words and their associated meaning. Automatic speech recognition attempts to emulate part of this task, that of mapping the spectral envelope into a series of words. There are many problems associated with this process. Not all people speak the same. The spectral envelope will vary due to regional accents and differences in the individual, for example whether male or female and their height. Among diverse human-device interfaces, speech recognition has widely been used in the last decade, and its importance becomes higher as the era of the Internet of Things comes close to reality . Due to the prevalence of energy-limited devices, energy- efficient architecture is inevitably demanded to lengthen the device life. The demand for low-energy architecture leads to the speech recognition system being implemented with dedicated hardware units . A speech recognition system consists of two processes:

1) feature extraction and 2) classification. The feature extraction process picks the characteristics of a sound frame, and a word is selected in the classification process by analyzing the extracted features. This brief mainly focuses on the hardware design of feature extraction. The most widely known feature extraction is based on the mel- frequency cepstrum coefficients (MFCCs), as MFCC-based systems are usually associated with hh recognition accuracy In MFCC extraction was implemented with an optimized recognition program running on a low-power reduced instruction set computer processor platform. To reduce energy consumption further, dedicated architectures have been proposed in and constructed with fixed-point operations. The previous architectures, however, have not fully considered the arithmetic property of the MFCC extraction algorithm. This brief presents a new energy-efficient architecture for MFCC extraction. Investigating the algorithmic property of MFCC extraction, we renovate the previous architecture with optimization techniques to reduce both hardware complexity and computation time. As a result, the energy consumption is remarkably reduced compared with the previous architectures.

II. ARCHITECTURE

This approach is completely different from that have utilized a separate hardware unit for each operation. The proposed architecture is described with setting N to 256, M to 13, and L to 32. For sound signals sampled with 16 bits at 16 kHz, in addition, the bit- widths of F and E in the floating-point representation are determined to 6 and 7 bits, respectively.

By analyzing the dataflow of the modified MFCC algorithm, we propose a new MFCC extraction system implementable with a small hardware cost. The overall architecture of the proposed system is shown in Fig 1, which consists of a multiply- and accumulate (MAC) unit, an address generation unit, a controller, memories, and counters. Though the proposed architecture has one MAC unit, it is sufficient to process.

In terms of memory size, the proposed memory structure is more efficient than those of previous works, since it is shared with several processes. To access an entity of a memory, the corresponding address is computed by employing counters. To fetch data for a MAC operation, each counter is increased by a certain amount. The proposed architecture utilizes two counters to generate two addresses needed to access two memories simultaneously.

A. Floating Point System

Since many operations used in the MFCC algorithm depend on complex functions, such as square and logarithmic functions, their outputs are associated with a large dynamic range. Compared with the fixed-point representation, the floating-point representation can cover such a large dynamic range with a much smaller number of bits. In addition, the operation bit-width can be reduced further, grounded on the property that the resulting feature vectors are influenced by the order of magnitude of interim values. For these reasons, a floating-point representation is employed in this brief to implement the modified MFCC extraction algorithm described above.

B. DDMFCC

DDMFCC refers to Delta Mel Frequency Cepstral Coefficients. It is obtained as the second derivative of MFCC (Mel Frequency Cepstral Coefficients). MFCC is one of the most important features, which is required among various kinds of speech applications. It shows high accuracy results for clean speech. They can be regarded as the "standard" features in speaker as well as speech recognition. However, experiments show that the parameterization of the MFC coefficients is best for discriminating speakers from the one usually used for speech recognition applications .It is the most common algorithm that is used for speaker recognition system. It is possible to obtain more detailed speech features by using a derivation on the MFCC acoustic vectors. This approach permits the computation of the DMFCCs, as the first order derivatives of the MFCC. Then, the DDMFCCs are derived from DMFCC, being the second order derivatives of MFCCs. The speech features which are the time derivatives of the spectrum-based speech features are known as dynamic speech features.

C. DDMFCC Algorithm

This is the block diagram for the feature extraction process apllying DDMFCC algorithm. Each blocks of figure 1 are described below.

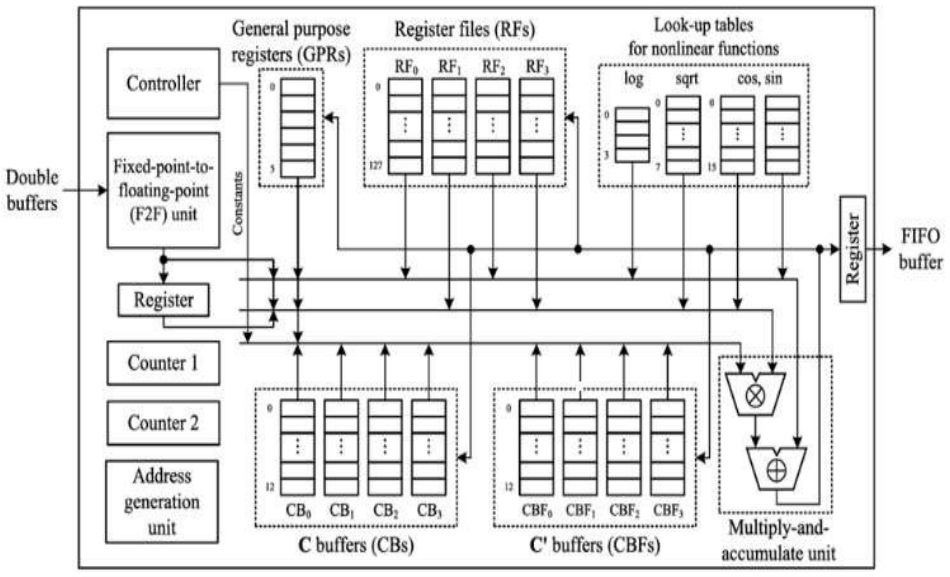


Fig. 1 Modified Architecture of MFCC

III. DDMFCC FLOW DIAGRAM

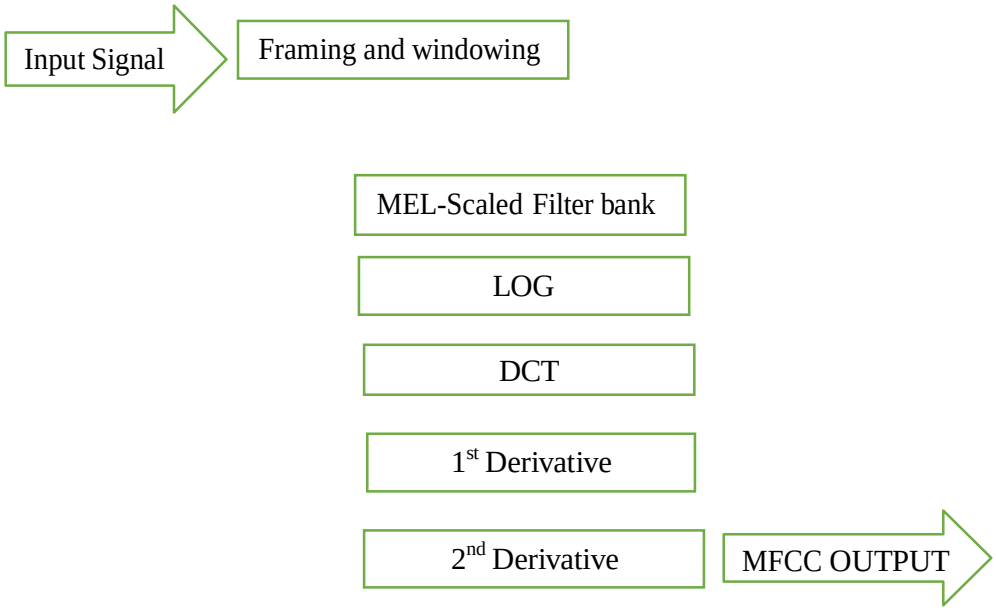


Fig. 2 DDMFCC Flow Diagram

1) *Framing and Windowing:* In figure 3, First the signal is split up into several frames such that we are analyzing each frame in the short time instead of analyzing the entire signal at once, at the range (10-30) ms the speech signal is for the most part stationary .Also an overlapping is applied to frames. This is called the Hop Size. In most cases half of the frame size is used for the hop size. The reason for this is because on each individual frame, a hamming window is applied which will get rid of some of the information at the beginning and end of each frame. Overlapping will then reincorporate this information back into our extracted features.

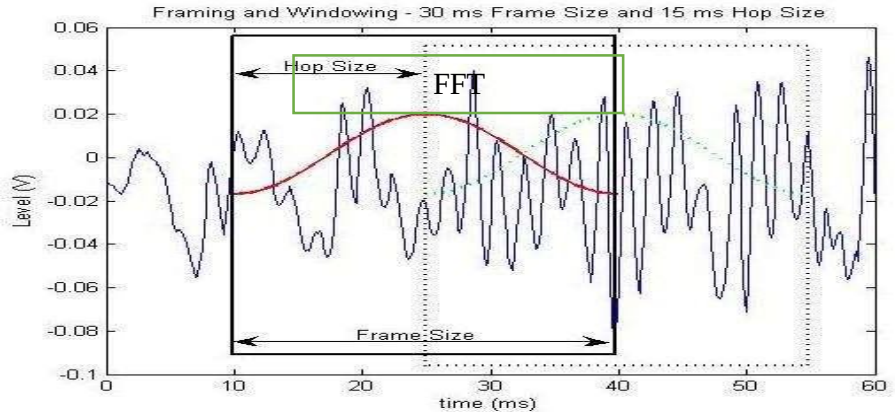


Fig. 3 Framing and Windowing

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. In speaker recognition, the most commonly used window shape is the hamming window. It gradually attenuates the amplitude at both ends of extraction interval to prevent an abrupt change at the endpoints. It produces the convolution for the Fourier transform of the window function and the speech spectrum. The hamming window WH(n) , defined as :-

tire(n) = 0.54 — 0.46 cos (2nπ/N-1)……..(1)

The use for hamming windows is due to the fact that mfcc will be used which involves the frequency domain (hamming windows will decrease the possibility of high frequency components in each frame due to such abrupt slicing of the signal).

2) *Fast Fourier Transform*: To convert the signal from time domain to frequency domain preparing to the next stage (Mel frequency wrapping). The basis of performing fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain. Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame.
3) *Mel Scaled Filter Bank:* The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above l000Hz.

We can use the following formula to compute the mels for a given frequency f in
Mel(f)=2595*1ogl0(1+f700)..... (2)
One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel frequency component. In figure 3, the filter bank has a triangular band pass frequency response. Mel-Frequency analysis of speech is based on human perception experiments. Human ears, for frequencies lower than 1 kHz, hears tones with a linear scale instead of logarithmic scale for the frequencies higher than 1 kHz. The information carried by low frequency components of the speech signal is more important compared to the high frequency components. In order to place more emphasize on the low frequency components, me1 scaling is performed. Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in high frequency regions.

S(1)'ZS(K) M(K)……..(3)
Where,
S(1) : Mel spectrum.
S(K) : Original spectrum.
M(K) : Mel filterbank.
1=0, 1,……..,L-1, Where L is the total number of mel filterbanks.
N/2 = Half FFT size

$$c_n = \sum_{k=1}^{k}(log\ S_k) \qquad -\frac{1}{2})\frac{\pi}{k}]$$

Where n=1,2,……K
The number of mel cepstrum coefficients, K, is typically chosen as (10-15). The first component, c 0 is included from the DCT since it represts the mean value of the input signal which carriers little speaker specific information. Since the log power spectrum is real and symmetric, inverse FFT reduces to a Discrete Cosine Transform(DCT). By applying the procedure described above, for each speech frame of about 30 ms with overlap, a set of mel-frequency cepstrum coefficients is computed. This set of coefficients is called an acoustic vector.

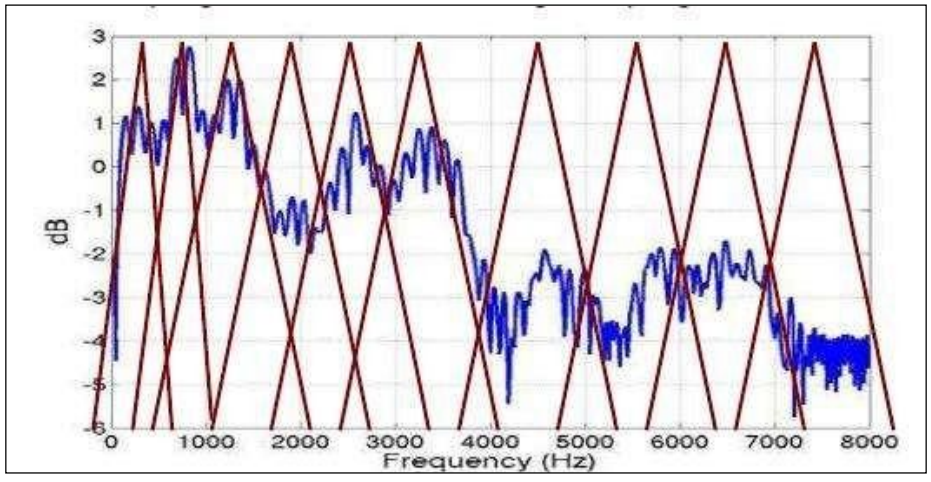


Fig. 4 Filter Bank

4) *Cepstrum:* In the final step, the log me1 spectrum has to be converted back to time. The result is called the melfrequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the me1 spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT).As shown in figure 4 cepstrum is obtained by taking the logarithm and DCT of me1 filter bank output. It is known that the logarithm has the effect of changing multiplication into addition. Therefore we can simply convert the multiplication of the magnitude of the fourier transform into addition.

Then, by taking the inverse FFT or DCT of the logarithm of the magnitude spectrum, the glottal pulse and the impulse response can be separated. The IFFT needs complex arithmetic than DCT. The DCT implements the same function as the FT more efficiently by taking advantage of the redundancy in a real signal. The DCT is more efficient computationally.

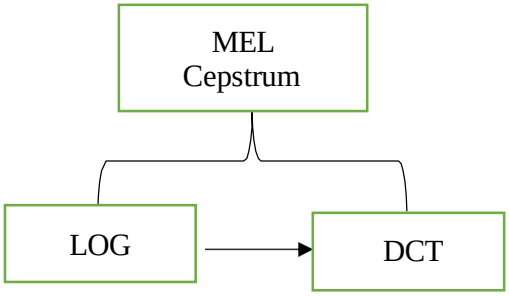


Fig. 5 Mel Spectrum Coefficients

IV. RESULTS AND DISCUSSIONS

A. *The Graphical User Interface*

A graphical user interface (GUI) is a pictorial interface to a program. A good GUI can make programs easier to use by providing them with a consistent appearance and with intuitive controls like pushbuttons, list boxes, sliders, menus, and so forth. The GUI should behave in an understandable and predictable manner, so that a user knows what to expect when he or she performs an action.

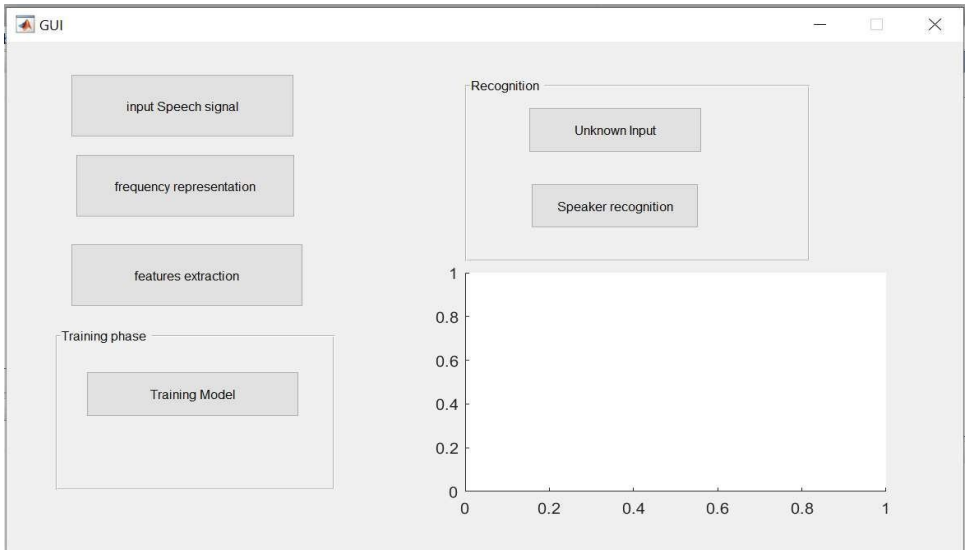


Fig. 5 Graphical user interface

Each item on a MATLAB GUI (pushbuttons, labels, edit boxes, etc.) is a graphical component. The types of components include graphical controls (pushbuttons, edit boxes, lists, sliders, etc.), static elements (frames and text strings), menus, and axes. Graphical controls and static elements are created by the function uicontrol, and menus are created by the functions uimenu and uicontextmenu. Axes, which are used to display graphical data, are created by the function axes.

B. Simulation Process

Train & Test : Dividing the samples into training set and testing set. Training set is a standard sample, and the testing set is the samples from different person in different environment. The author will train the HMM model based on the training sample set and test the system based on the testing set.

Input speech: The input speech signal is given in the format of(.way). The WAV is an audio file format that stores wave form data. What makes the WAV different from other audio formats is it's uncompressed — making it much larger than something like an mp3. It's a raw audio file capable of saving recordings using different bitrates .The speech signal given for input is “Energy Efficient Floating Point MFCC Extraction Architecture For Speech Recognition Systems”. The recorded signal is not a noise free signal.

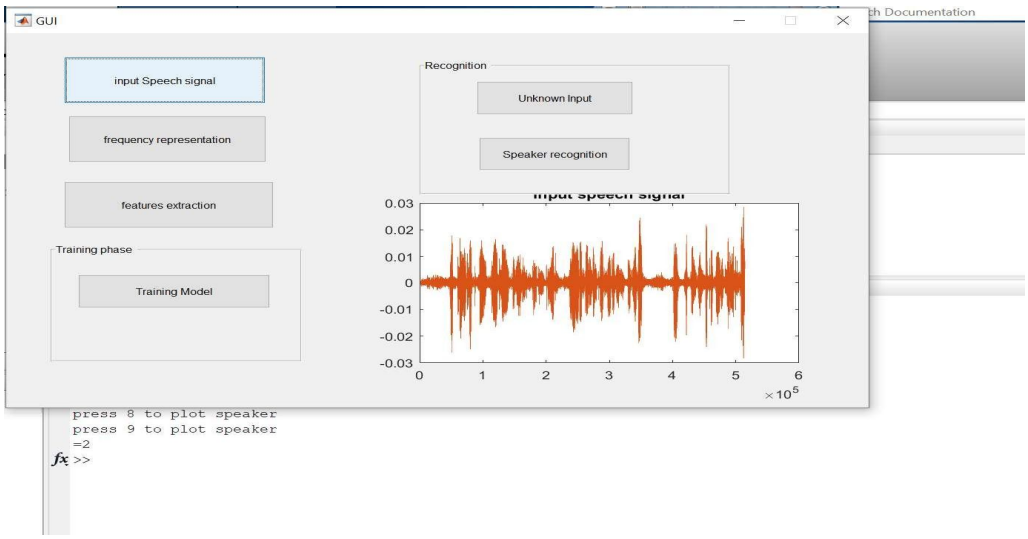


Fig. 6 The plot for a speaker two

1) *Frequency Representation:* In the input speech signal function we convert the speech signal into frames. These frames must undergo windowing process of blackmann window. These signal is converted to frequency domain using short time fourier transform and fast fourier transform. The difference is the short time fourier transform converts only the small signals into speech signal and the FFT can convert only the large signals so we are using to forms. The axis is used for plotting the frequency converted signal of one speaker.
2) *Feature Extraction*: This is the main process where we are converting the speech signal into some coefficients for training and testing the data. Here we use MFCC process for frequency extraction. The output is represented in me1 filter bank format. Here we are extracting two features are MFCC coefficients, the other is logarithmic coefficient.

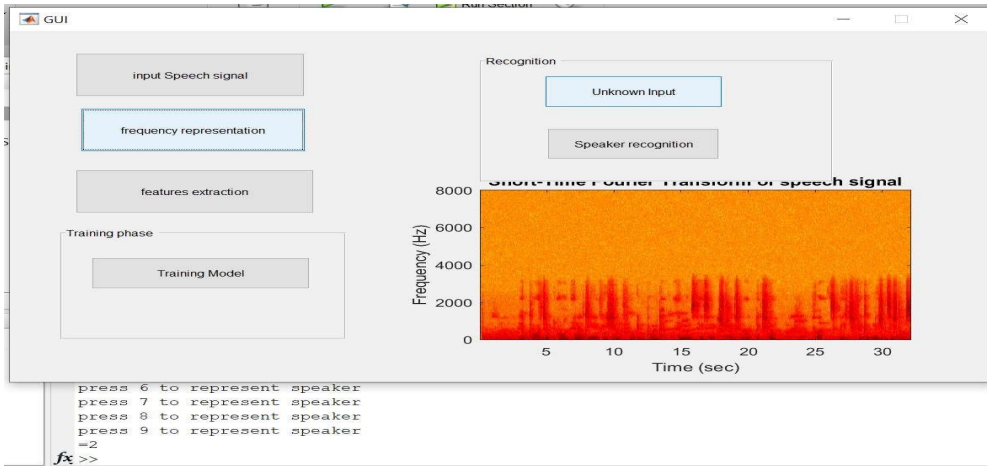


Fig. 7 Frequency representation of speaker two

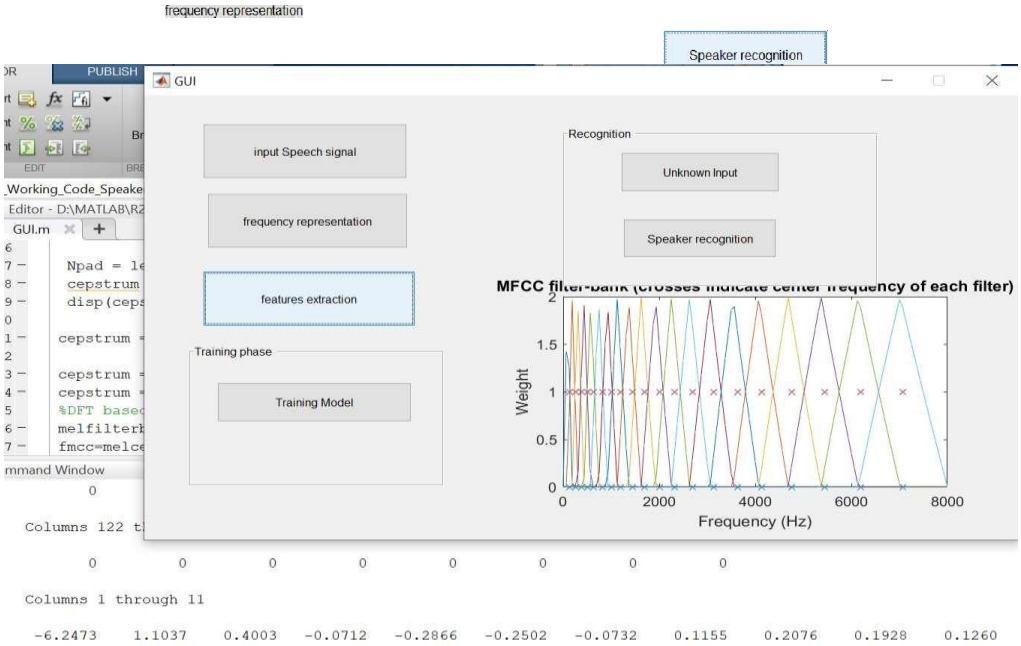


Fig. 8 Plot of coefficients

3) *Training:* The training is performed using HMM model. Training data "teaches" an algorithm to recognize patterns in a dataset. More specifically, training data is the dataset you use to train your algorithm or model so it can accurately predict your outcome.
4) *Testing:* The algorithm can make every gaussian distribution of the HMM model as one kind, the parameter of Mean is the position in Characteristic parameters space of different sample.

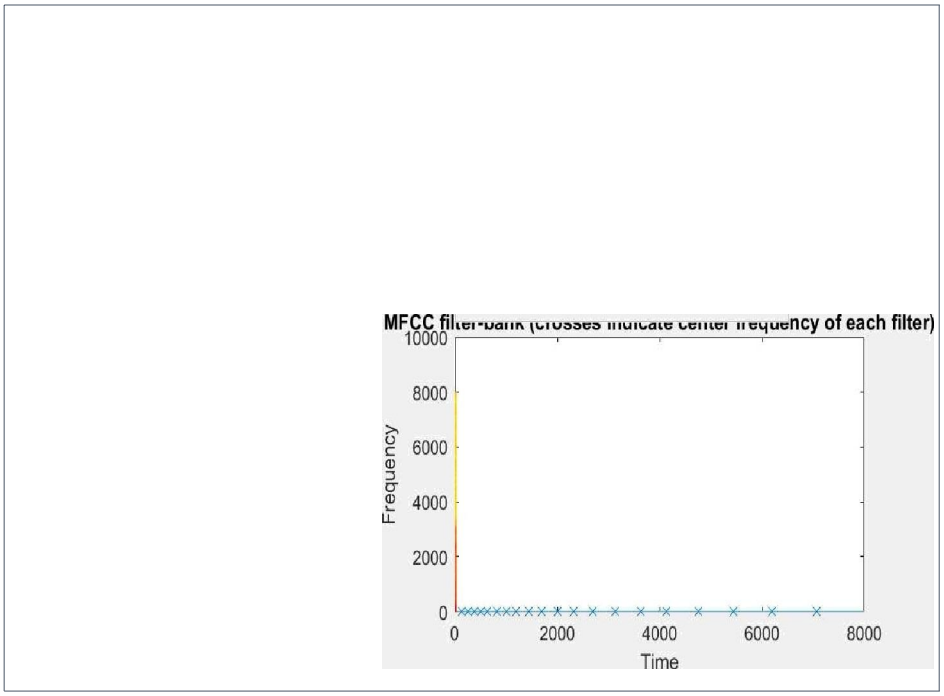
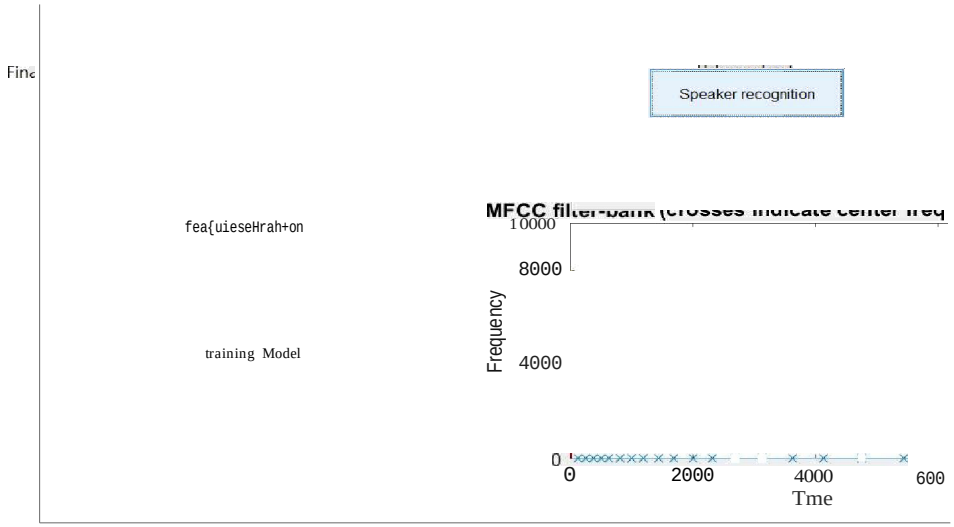


Fig. 9 Recognition of speaker



Press enter to train Authorized speakers
>> GUI
Uzer9
Unauthorized speaker

Fig. 10 Recognition of unauthorized speaker

V. CONCLUSIONS

An energy-efficient MFCC extraction architecture has been presented for speech recognition. The MFCC extraction algorithm is modified to minimize computation time without degrading the recognition accuracy noticeably. In addition, the proposed architecture employs floating-point arithmetic operations to minimize the operation bit- width and the total size of LUTs, while have relied on fixed-point operators. The effectiveness of energy consumption makes the proposed architecture a promising solution for energy-limited speech recognition systems.

VI. FURTHER SCOPE

Feature extraction is the first crucial component in automatic speech processing. Generally speaking, successful front-end features should carry enough discriminative information for enhancements, and it should fit well with the back-end modelling, and be robust with respect to the changes of acoustic environments. As a part of the project work the MFCC feature extraction were analyzed for better understanding of the work. To extend this work for better performances of auto segmentation process in detail analyzes of speech over various cestrum scales like bark scale and its robustness over different noise conditions will be evaluated.

REFERENCES

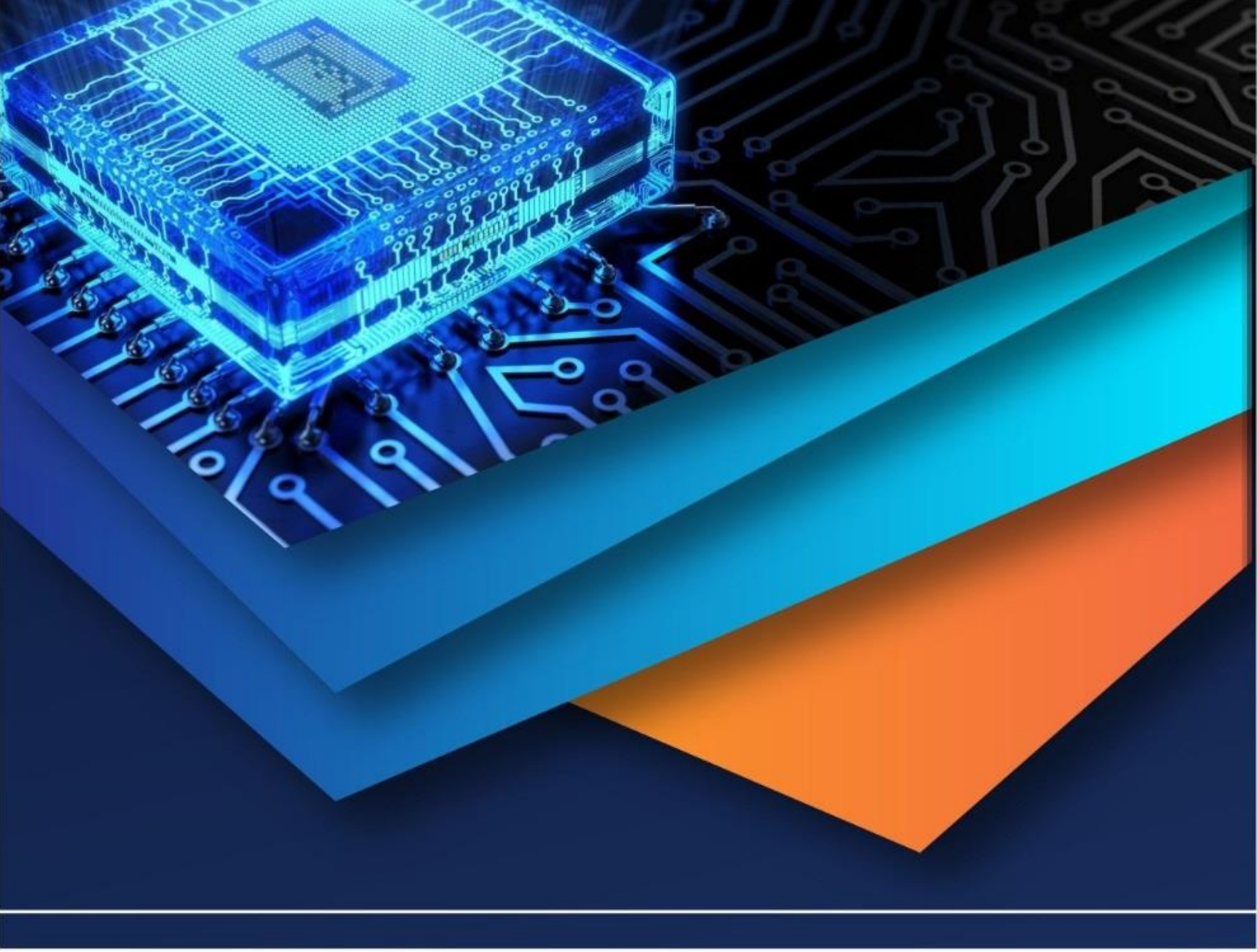
[1] N.-V. Vu, J. Whittington, H. Ye, and J. Devlin, “implementation of the MFCC front-end for low-cost speech recognition systems,” in proc. ISCAS, may/jun. 2010, pp. 2334—2337.
[2] P. Elkan, T. Allen, and S. F. Quigley, “FPGA implementation for gym-based speaker identification,” int. J. Recon fig. Compute., Vol. 2011, no. 3, pp. 1—8, jan. 2011, art. Id 420369
[3] R. Ramos-lara, M. López-garcía, E. Cantó-navarro, and L. Puente-rodriguez, “real- time speaker verification system implemented on reconfigurable hardware,” J. Signal process. Syst., Vol. 71, no. 2, pp. 89—103, may 2013
[4] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, “an efficient MFCC extraction method in speech recognition,” in proc. IEEE ISCAS, may 2006, pp. 145—148.
[5] A. Sunderland, R. A. Strauch, S. S. Wharfield, H. T. Peterson, and C. R. Cole, “CMOS/SOS frequency synthesizer LSI circuit for spread spectrum communications,” IEEE J. Solid-state circuits, vol. 19, no. 4, pp. 497—506, aug. 1984.
[6] Qi.Li, “An auditory-based feature extraction algorithm for Robust speaker Identification under mismatched conditions,” IEEE Transactions on audio, Speech, and language proeessing, Vol.19, No. 6, August 2011.
[7] Tze Fen Li, Shui-Ching Chang, “ Speech recognition of mandarin syllables using both linear predict coding cepstra and Mel frequency cepstra,”2009.

[8] Zhou H., and Aziz A., (1998) "Simultaneous PTL buffer insertion and sizing for minimizing Elmore delay," in Proc. Int. Workshop Logic Synth., pp. 162
[9] Lalitha V, and Kathiravan S, "A Review of Manchester, Miller, and FM0 Encoding Techniques", Smart Computing Review, vol. 4, no. 6, pp.481-490 December 2014.6. V. Hemalatha, P. Srividhya.
[10] Hung.Y.C, Kuo.M.M, Tung.C.-K., and S.-H. Shieh, (2009) "High speed CMOS chip design for Manchester and Miller encoder," in Proc. Intel. Inf. Hiding Multimedia Signal Process., pp. 538—541.

AUTHORS PROFILE

Dr. M. Ramana Reddy is working as an Assistant professor in ECE Department at Chaitanya Bharathi Institute of Technology. His educational qualification is Doctor of Philosophy in VLSI. He has published 2 research papers in International journals. His area of specialization is CMOS VLSI.

Srinija Lankala pursued Bachelor of Engineering in Electronics and Communication at Chaitanya Bharathi Institute of Technology, Hyderabad. Her areas of interest are computer architecture, vlsi design and architecture.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY