Digital Architecture For Computing Energy of Speech Frames Using Configurable MAC

B. Sushma
Assistant Professor
Department of Electronics and Communication Engineering
PSCMR College of Engineering and Technology
Vijayawada, Andhra Pradesh, India
sushmaoct02@gmail.com

S. Hemanth Kumar
Department of Electronics and Communication Engineering
PSCMR College of Engineering and Technology
Vijayawada, Andhra Pradesh, India
hemanthkumarsanagapalli17@gmail.com

G. Sai Susmitha
Department of Electronics and Communication Engineering
PSCMR College of Engineering and Technology
Vijayawada, Andhra Pradesh, India
gadamsettysusmitha@gmail.com

B. Sivaparvathi
Department of Electronics and Communication Engineering
PSCMR College of Engineering and Technology
Vijayawada, Andhra Pradesh, India
sivaparvathi115500@gmail.com

T. Gnana Hari Krishna
Department of Electronics and Communication Engineering
PSCMR College of Engineering and Technology
Vijayawada, Andhra Pradesh, India
tghari12@gmail.com

D. Krishna
Professor
Department of Electronics and Communication Engineering
PSCMR College of Engineering and Technology
Vijayawada, Andhra Pradesh, India
krishnadharavath4u@gmail.com

***Abstract*— Abstract: This study introduces a novel digital hardware architecture tailored for speech applications, focusing on a configurable Multiple-and-Accumulate (MAC) unit. In digital speech processing, signals are typically segmented into frames for subsequent analysis. These frames are then categorized based on their signal response, commonly into voiced, unvoiced, and silence segments. Such classification, known as V/UV/S, holds significant importance across various speech-based applications. A key parameter widely utilized for distinguishing between speech activity and silence is Short-Time Energy (STE).**

The proposed MAC unit serves the purpose of computing STE for speech frames. Given the variability in frame sizes encountered in speech frames, the configurability of the MAC unit in terms of frame size enables efficient computation of STE using streaming samples. Implemented in Verilog HDL and utilizing the Xilinx Vivado tool, this paper elaborates on the hardware architecture of the proposed MAC unit and provides comprehensive insights into its performance metrics.

I. INTRODUCTION

In speech-based applications, voice signals are typically divided into short-duration overlapping chunks that are sequentially sequenced. Salient traits are extracted by processing the resulting speech segments. To create recognition models for a broad range of speech-based applications, characteristics derived from a sizable collection of speech signals are combined. However, not every speech signal segment provides useful information for activities like system modeling and recognition. Speech segments are categorized as voiced (V), unvoiced (UV), and silence (S) in a more general sense. Voiced speech is produced when the vocal cords' vibrations alter the air coming out of the lungs, giving the impression of quasi-periodic excitement [1]. The resulting sound is primarily made up of oscillations at a low frequency.

Silence and unvoiced portions are included in the non-voiced speech. When air from the lungs goes through a tiny constriction in the vocal tract, it becomes turbulent and noiselike excitement without any dominating low-frequency oscillations, resulting in unvoiced speech that is non-periodic and uncorrelated in character [1]. Conversely, the quietness happens when the vocal tract system is not stimulated. Therefore, in many speech-based applications, knowing the type of speech segment is useful. For instance, the frontend of the feature extraction stage of speech recognition systems uses the V/UV/S detector to remove UV/S speech segments [2].

The literature has offered a number of strategies for classifying speech segments during the past few decades [3]–[6]. These techniques primarily focus on offering software domain solutions. As far as the authors are aware, there aren't many published works about the creation of specific, unique hardware designs for the V/UV/S classification of speech segments This inspired us to conduct research in this area and to propose a digital. architecture for speech segment instantaneous V/UV/S categorization. We decided to create the aforementioned architecture using the short-time energy (STE) and short-time average zero-crossing rate (STAZCR), two widely-used time-domain-based speech characteristics. The following justifies the selection of these criteria above others: First off, mel-frequency cepstral coefficients (MFCC) are thought to be the state-of-the-art features and are frequently employed in feature extraction approaches. [7, 8]. The energy of the speech segment is represented by the first coefficient in the MFCC feature vector. Since the computed value satisfies the conditions of both the MFCC feature vector building and the V/UV/S classification, we choose to utilize STE. Second, silence and speech activity can be effectively classified using STE. However, it is challenging to further categorize the speech as V/UV. As a result, we decided to make STAZCR the second parameter. Thirdly, in comparison to other methods, the computing complexity of the STE and STAZCR is lower. However, their mathematical formulas provide hardware realization feasibility and satisfy the recurrence property.

The following sums up the important contributions made in this paper: First off, the suggested architecture can be reconfigured to accommodate speech chunks with varying lengths up to 1024 powers of 2. Second, in order to map the needed functionality in hardware, we used the algorithmic state machine with datapath (ASMD)-based design technique. The suggested architecture is able to operate with window function architectures that are pipelined or recursive coordinate rotation digital computer (CORDIC) based. These architectures generate windowed speech samples sequentially at a rate of one sample per clock cycle or one sample per L clock cycles, respectively [9]. STE and STAZCR can process clear, noise-free speech signals because they employ thresholds derived from empirical observations. Consequently, the intended hardware architecture is meant to be used in low-noise voice recognition systems.

II. LITERATURE SURVEY

S. Sunil Kumar et al. proposed, the task of determining whether vocal fold activity zones are present or absent in the speech signal is known as voice/non-voice detection. The majority of current state-of-the-art techniques rely solely on the signal's amplitude, either in the time or frequency domains, and this has a substantial impact on how well they function during weakly voiced laryngeal transitions and noisy speech segments. In this study, we present a robust technique based on the source signal's phase harmonics for identifying voice and non-vocal areas in the speech signal. Here, zero frequency filtering (ZFF) is used to extract the voice signal's source signal from the vocal tract resonances. The experimental findings show how reliable the suggested approach is for correctly identifying voiced and non-voiced regions. [2]

R. Bachu et al. The voiced-unvoiced choice is typically made in speech analysis in order to extract information from the speech signals. Two techniques are used in this research to distinguish between the voiced and unvoiced portions of the speech signals. They are energy and zero crossing rate, or ZCR. Here, we divided the speech sample into segments and utilized energy and zero crossing rate computations to distinguish between voiced and unvoiced speech segments in order to assess the findings. The findings imply that while energy is high for voiced parts and low for unvoiced parts, zero crossing rates are low for voiced parts and high for unvoiced parts. These techniques have therefore been shown to be successful in differentiating between voiced and unvoiced speech.[3]

B. Atal et al. The voiced-unvoiced determination in speech analysis is typically carried out in tandem with pitch analysis. Pitch analysis and voiced-unvoiced (V-UV) decision-making are linked, which adds needless complexity and complicates the classification of brief speech fragments that last less than a few pitch intervals. We provide in this study a pattern recognition method for classifying, based on signal measurements, a given segment of a speech signal as voiced speech, unvoiced speech, or silence. The speech segment that has to be categorized is measured using five distinct techniques in this procedure.[4]

F. Ykhlef et al. In this study, we have evaluated many temporal domain characteristics for speech signal categorization into voiced and non-voiced categories. To create three distinct classifiers, we have seamlessly selected the autocorrelation function (ACF), weighted ACF (WACF), and average magnitude difference function (AMDF). Experiments were carried out in both clean and noisy situations using the TIMIT database. The generated classifiers have been validated by the use of white noise taken from the NOISEX92 database. The average value of the percentage of classification accuracy (Pc) has been used to rank these classifiers overall.[5]

S. Ahmadi et al. An enhanced algorithm for pitch determination and voice identification based on cepstrum is showcased. A multifeature voiced/unvoiced classification technique based on statistical analysis of the energy, zero-crossing rate, and cepstral peak of short-time speech signal segments is used to determine the voice. A modified cepstrum-based technique is used to extract pitch frequency information, which is then meticulously adjusted utilizing pitch tracking, correction, and smoothing algorithms. A thorough investigation of performance on a sizable database shows a significant improvement over the traditional cepstrum approach. Furthermore, demonstrated is the suggested algorithm's resistance to additive noise.[6]

N. S. S. Srinivas et al. Recognition of language from voice utterance is known as spoken language identification (LID) or spoken language recognition (LR). This research proposes a novel Fourier parameter (FP) model for spoken language recognition that is independent of the speaker. The analysis and comparison of the suggested FP features' performance with the legacy mel-frequency cepstral coefficient (MFCC) features is done. The two multilingual databases utilized to extract FP and MFCC characteristics are the Oriental Language Recognition Speech Corpus (AP18-OLR) and the Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus (IITKGP-MLILSC). Three classifiers—feed-forward artificial neural networks, deep neural networks, and support vector machines—are used to create spoken LID/LR models utilizing the retrieved FP and MFCC characteristics.[7]

N. Sugan et al. Speech emotion recognition, also known as speech utterance analysis, is the act of determining the speaker's emotional state from the speech. Numerous other cepstral qualities have been suggested in previous studies as a means of developing SER systems. Two widely utilized variations of cepstral features are the mel-frequency cepstral coefficients (MFCC) and the human-factor cepstral coefficients (HFCC). Mel and human-factor filter banks are used, respectively, to extract MFCC and HFCC characteristics from voice signals. Each filter in these filter banks has a triangle-shaped magnitude response. Because of this, these filter banks are known as triangular filter banks (TFB), and the derived cepstral coefficients that correspond to them are designated as TFBCC-M (for MFCC) and TFBCC-HF (for HFCC).[8]

V. Kumar et al. In signal processing, communication systems, and image processing, popular window techniques like Hanning, Blackman, Blackman-Harris, and flat top windows are a logical choice before fast Fourier transform (FFT) to minimize the undesired phenomenon, such as spectral leakage and picket fence effect, which arise due to direct truncation by rectangular window. It is vital to select a certain window function and window length based on the applications since window characteristics vary depending on the kind of function and length of window employed. On the other hand, effective and adaptable very-large-scale-integration (VLSI) architecture is required for the window function to be implemented in real-time. Therefore, this article provides a novel hardware efficient VLSI architecture based on coordinate rotation digital computer (CORDIC) that can be reconfigured to select a window function from the aforementioned popular window. [9]

III. METHODOLOGY

In this article an approach is used to calculate the short time energy (STE) of the speech signal using a configurable MAC. Here new digital hardware architecture with a customizable Multiple-and-Accumulate (MAC) unit designed for voice applications. Signals in digital voice processing are usually divided into frames for further examination. The signal response of these frames is then used to classify them, usually into spoken, unvoiced, and quiet segments. This kind of categorization, called V/UV/S, is very important for many speech-based applications. Short-Time Energy (STE) is a critical characteristic that is frequently used to differentiate between speaking activity and quiet.

The computation of STE for voice frames is the function of the proposed MAC unit. The MAC unit's frame size configurable feature allows for the efficient computation of STE utilizing streaming samples, especially considering the wide range of frame sizes seen in speech frames. Upon power-up, the controller switches into the idle state, and waits for the start signal. Before the start signal arrives, the controller must be provided with the MAC configuration data, denoting the frame size of the speech signal, for appropriately computing the STE.

Whenever start signal is asserted (for one clock cycle), then the controller switches into the computation state. At the time, its internal counter register and accumulator register are reset to zero. Next The controller waits to receive a valid speech frame sample on the data Sample input. When a speech sample arrives on the said input, the dataValid input is asserted for one clock cycle. When dataValid at low logic then the MAC unit is disabled.

When the dataValid input is at high logic, then the MAC unit is enabled. When MAC is enabled, then the accumulator register value is added to the resultant of the multiplier and the final resultant is updated in the MAC. Upon updating the MAC, the controller increments its counter register. The controller compares its counter value with the frame value. When the count value is less than the frame value, then the counter repeats the previous steps to repeat the process

If the count value is equal to the frame value, then the contents of the accumulator register is the desired STE value. The computed value is sent out via the output line and the corresponding outputValid signal is asserted for one clock cycle. Finally, the controller switches back to the idle state.

This is a special instance that yields speech segments that do not overlap. There are very few speech-based programs that support this version. Ultimately, segment-based speech parameters are computed using the acquired speech segments. It is frequently challenging to categorize the voice segments in the V/UV/S classification situation with a single parameter.

Consequently, the segment-based STE is used in this article jointly to accomplish the intended goal. These parameters have the following definitions:

A speech segment's STE is indicated by [10],

$$= \sum | \ [m] \cdot h[n - m] \ |^2, n \in [0, Nf - 1], \ (1)$$

where h denotes a Hamming window function defined as [11]

$$h[\ ] = 0.54 - 0.46 \cos (\quad \quad ^2) , \quad = , \quad (2)$$

where the window length is indicated by N. A speech segment is categorized as a speech (Sp) segment if its E is comparatively high in relation to a predetermined threshold (Eth); otherwise, it is classified as a quiet segment. Figures 1(a) and 1(b) provide contour graphs that demonstrate this. The E estimations are shown in a linear scale in the former and in decibel and log-compressed scales in the latter.

To minimize space overheads and computational complexity in the event of hardware implementation, the E estimations are, however, given in a linear scale. Moreover, the speech segments are categorized as S/Sp using the E estimates derived on xp [m]. The categorization result is combined to determine the borders of transition between S/Sp areas in xp [m].

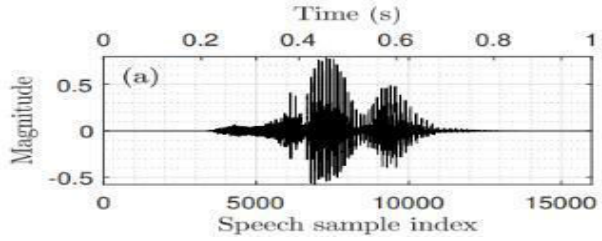

Fig.1(a)

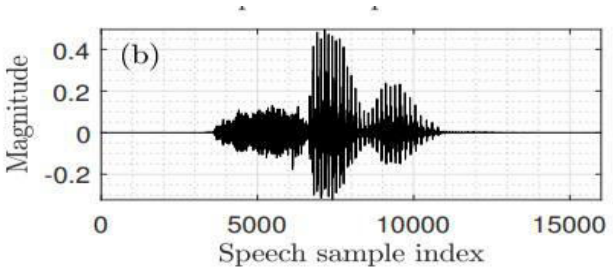

Fig.1(b)

After entering the idle state, the controller awaits the start signal. In order for the controller to calculate the STE

correctly, it has to be given the MAC configuration data, which indicates the speech signal's frame size, prior to the start signal.

After asserting the start signal for one clock cycle, the controller enters the calculation stage. Its accumulator register and internal counter register are reset to zero at that point.

On the dataSample input, the controller waits to receive a valid voice frame sample. The dataValid input is asserted for one clock cycle upon the arrival of a voice sample on the specified input. The MAC unit is deactivated while dataValid at low logic.

Calculating the Short-Time Energy (STE) of speech signals using a MAC unit that may be configured. The frame size characteristic of this MAC unit may be adjusted, which is important for effective STE calculation utilizing streaming samples and the classification of speech frames into spoken, unvoiced, and silent segments (V/UV/S categorization). When activated, the MAC unit is used by the controller in this design to analyze speech samples and transition between the idle and calculation modes. The computed STE value is then output by the controller. This design is specifically made to handle the range of frame sizes that are frequently used in voice processing.

The following figure 2 differentiates between the speech signal, voiced signal, unvoiced signal, silence. This classification of signal into these categories helps the controller to perform the desired processing to the signal.

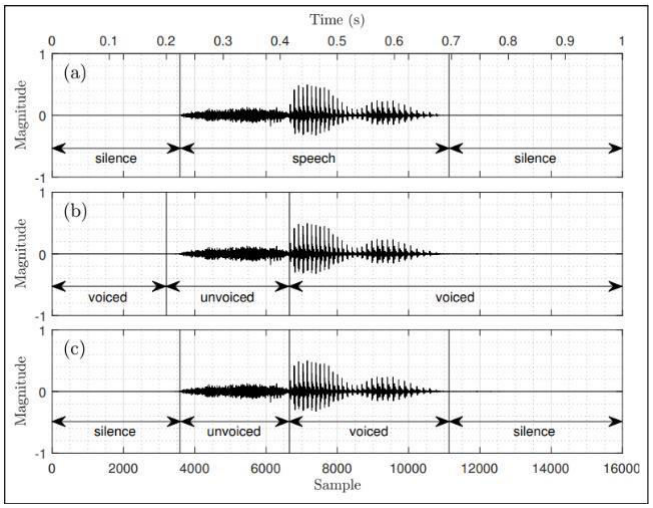

Fig.2

The portion of signal which represents the silence does not contain any information. Remaining part of the signal is combination of voiced and unvoiced signal. Unvoiced signal is a mixture of information and noise together. The controller shifts to operational mode from idle state whenever controller input is fed with a speech signal. Another input to the controller is given with the configurable MAC, which contains information about the number of frames to be divided for the further processing.

Before the controller process the signal that has to be pre-emphasized and should undergo window technique to eliminate the sudden spikes or raises in the signal. From figure 4 whenever the data valid signal is high start signal is made high to proceed with the operation. The process begins to start after the activation of start signal. MAC register counts the number of frames processed and increments the counter each time after completion of frame processing.

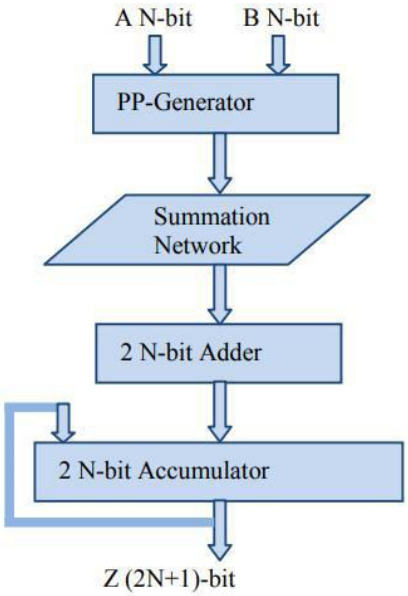

Fig.3

Above figure represents the MAC which is a part of the controller . This MAC unit helps the controller initiating and terminating the process of the controller.

In digital signal processing and computer architectures, a Multiple Accumulate Unit (MAC) is an essential part. It carries out the action of multiplying two input values and adding the outcome to the sum of the previous values. Typically, a multiplier, an accumulator, and control logic make up a MAC unit.

The data input and the coefficient input are the two values that the MAC unit gets when it is operating. These inputs are multiplied by the multiplier to create a product, which is then added to the current value of the accumulator. The accumulated total is kept in the accumulator and is updated with every calculation cycle.

The capability of the MAC unit is essential for many applications, including matrix computations, convolution, and filtering. Compared to independent multiplier and accumulator units, it lets complicated mathematical operations be computed efficiently with less hardware complexity. Furthermore, MAC units are frequently utilized in microprocessors and digital signal processors (DSPs) for the quick and energy-efficient processing of data and signals.

IV. RESULTS

Figure 4 represents the simulated results of the controller output of the speech signal, various components and signals that are given to the controller and MAC. Data Valid

signals becomes logic high whenever the input is given with a speech or voice signal.

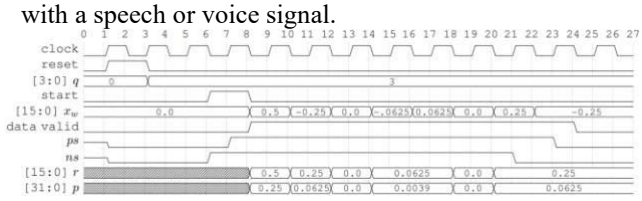

Fig.4

Verilog-HDL is utilized for the proposed architecture's simulation duties. The outcomes of the simulation provide a clear explanation of the controller's operation for every clock cycle. To put the controller in the idle state, use the reset signal. The voice signal is where the controller begins processing when the start signal changes from logic low to logic high. This procedure keeps on until all frames have been processed. Upon completion of this procedure, Short-Time Energy (STE) is acquired and utilized for further uses.

V. CONCLUSION

This study suggests a digital architecture that uses a MAC unit that may be configured to calculate the energy of voice signals or frames. Using the programmable MAC, this design calculates the segment-based short-time energy (STE). Short voice signal energy calculations can be analyzed and used for future uses.

VI. REFERENCES

[1] S. Sunil Kumar and K. Sreenivasa Rao, "Voice/non-voice detection using phase of zero frequency filtered speech signal," Speech Commun., vol. 81, pp. 90 – 103, 2016.

[2] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in Adv. Tech. Comput. Sci. Softw. Eng. Springer Netherlands, 2010, pp. 279–282.

[3] B. Atal and L. Rabiner, "A pattern recognition approach to voicedunvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust., Speech, Signal Process., vol. 24, no. 3, pp. 201– 212, 1976.

[4] F. Ykhlef and L. Bendaouia, "Evaluation of time domain features for voiced/non-voiced classification of speech," in Int. Conf. Signals, Electron. Syst. (ICSES), 2012, pp. 1–4.

[5] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," IEEE Trans. Speech Audio Process., vol. 7, no. 3, pp. 333–338, 1999.

[6] N. S. S. Srinivas, N. Sugan, N. Kar, L. S. Kumar, M. K. Nath, and A. Kanhe, "Recognition of spoken languages from acoustic speech signals using Fourier parameters," Circuits Syst. Signal Process., vol. 38, no. 11, pp. 5018–5067, 2019.

[7] N. Sugan, N. S. S. Srinivas, L. S. Kumar, M. K. Nath, and A. Kanhe, "Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on Bark and ERB frequency scales," Digit. Signal Process., vol. 104, p. 102763, 2020.

[8] V. Kumar, K. C. Ray, and P. Kumar, "A VLSI architecture of CORDIC based popular windows and its FPGA prototype," Int. J. High Performance Systems Architecture, vol. 7, no. 2, pp. 57–69, 2017.

[9] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," ArXiv e-prints, Apr. 2018.

[10] L. Rabiner and R. Schafer, Digital Processing of Speech Signals. Prentice Hall, 1979.

[11] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-Time Signal Processing; 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

BIBILOGRAPHY

B Sushma working as Assistant. Professor in PSCMRCET, Vijayawada. She completed her M. Tech with Embedded systems specialization at K L University. She holds memberships in IAENG, IFERP, and SDIWC. She published 7 papers in reputed journals and attended national and international conferences.

S. Hemanth Kumar is pursuing his final year in the Department of Electronics and Communication Engineering, PSCMR College of Engineering and Technology, Vijayawada, Andhra Pradesh, India. He was born on 5 December, 2002, with a keen interest in VLSI, Image Processing, Embedded Systems, , and IoT.

T. Gnana Hari Krishna is pursuing his final year in the Department of Electronics and Communication Engineering, PSCMR College of Engineering and Technology, Vijayawada, Andhra Pradesh, India. He was born on 05 november, 2002, with a keen interest in VLSI, Embedded systems, and IoT.

B. Siva Parvathi is pursuing her final year in the Department of Electronics and Communication Engineering, PSCMR College of Engineering and Technology, Vijayawada, Andhra Pradesh, India. She was born on 01 july, 2002, with a keen interest in VLSI, Embedded Systems and IoT.

G. Sai Susmitha is pursuing her final year in the Department of Electronics and Communication Engineering, PSCMR College of Engineering and Technology, Vijayawada, Andhra Pradesh, India. She was born on 13 march, 2002, with a keen interest in VLSI, Embedded Systems and IoT.