A High Performance Dynamic ASIC-Based Audio Signal Feature Extraction (MFCC)

Tam Chi Nguyen, Lam Dang Pham, Hieu Minh Nguyen, Bao Gia Bui, Dat Thanh Ngo, Trang Hoang
Fac. of Electrical & Electronic Engineering
HCM Univ. of Technology
Ho Chi Minh City, Viet Nam
{lamd.pham, hieumnguyen, hoangtrang}@hcmut.edu.vn, {phamnguyen170791, thanhdat5494, buigiabao02}@gmail.com

***Abstract*— State-of-the-art speech recognition, speech analysis as well as music modeling have approached Mel-Frequency Cepstral Coefficient (MFCC) and confirmed great performance in comparison to other feature extractions. Based on obtained software performance, a wide range of hardware designs are applied to highly increasing integrated systems achieving real-time performance and ability of mobility. Nevertheless, most hardware approaches witnessing certain configurations have experienced limitation of functions due to fixed-point format, strict silicon requirements or exact applications, which is reasonable for low ability of reusing and high cost of product. As regards MFCC method, there are various concerning parameters such as number of samples, range of filter bands, Fast Fourier Transform (FFT) number, number of cepstrums or even different level of delta coefficients, which significantly affect final performance of entire applications. As a result, a dynamic ASIC-based MFCC hardware architecture is proposed in this paper in order to meet real-time system requiring high performance as well as confirm superiorities regarding to ability of reconfiguration feasibly through an Advance High-performance Bus (AHB) interface in chip level instead of modifying at Register Transfer Level (RTL) in developed duration. Besides, have not only experiments on 130nm technology with full ASIC design flow witnessed high frequency at 500MHz but applying IEEE 754 floating-point format has also confirmed great accuracy between hardware design and software design, which apply in certain application towards Vietnam automatic speech recognition (ASR).**

Keywords—Mel-Frequency Cepstral Coefficient (MFCC), Advance High-performance Bus (AHB), Fast Fourier Transform (FFT), Pre-emphasis, Window wind, Cepstrum, Delta.

I. INTRODUCTION

MFCC has become highly popular as regards ASR systems or music modeling methods due to high performance of audio feature extraction, which has experienced in Table I with conducted survey from [1] to [7]. However, most approaches have applied software-based designs to be able to achieve flexible configurations, which is also reasonable for low performance of speech, reducing ability of integration or even witnessing limitation of real-time systems as considerable disadvantages. Indeed, the authors in [1] and [3] indicated that MFCC was the most classic hardware architecture for ASR systems. Another design proposed by [6] also confirms that dynamic MFCC increases 5 to 6 percent of the recognition rate than a fixed one. Additionally, [10] and [11] have same problem posing a trade-off between energy consumption and system performance, which need to be solve as future work.

Although the contrasting trend performing on hardware-based designs can tackle such issues, other concerning problems associating with ability of reconfiguration, silicon requirement as well as how to achieve high accuracy based on hardware view point with a vast number of operations are considered carefully.

TABLE I. STATISTICS TOWARDS MFCC AND APPLICATIONS

Author	Implementation	Method	Identification Algorithms	Conclusion	Application
[1]	Hardware (FPGA) And Software (C++)	MFCC And LPC And ENH-MFCC	HMM	MFCC better than LPC	Automatic Speech Recognition system (ASR system)
[2]	Software	LPC And MFCC	ANN	MFCC better than LPC	Automatic Language Identification(Arbic, English, French)
[3]	Software	MFCC	HMM	MFCC is more dominant used	ASR system
[4]	Software (Matlab)	MFCC	DTW	-	Voice Identification (Security system)
[5]	FPGA	MFCC	HMM	MFCC is extensively used	ASR system
[6]	Software	MFCC (Dynamic)	GMM	Dynamic MFCC better than 5% to 6%	ASR system with real noisy environment
[7]	-	General	General	MFCC is used mostly	ASR system with Marathi Language

978-1-5090-6144-0/16 $31.00 © 2016 IEEE
DOI 10.1109/ACOMP.2016.21

TABLE II. DIFFERENT MFCC CONFIGURATIONS ON HARDWARE

Author	Implement on	Sample in frame	FFT point	Mel filter	Cepstrum	Delta's order
[8]	FPGA (Xilinx)	1024	1024	24	24	-
[9]	ASIC (0.6µm)	256	256	20	12	-
[10]	ASIC (130nm) And FPGA (Xilinx)	256	256	32	13	2
[11]	ASIC (0.18 µm)	256	256	-	12	1 & 2
[12]	FPGA	160	256	33	12	2
[13]	FPGA (Xilinx Spartan)	512	512	24	12	-
[14]	FPGA	128-256	128-256	24	12	-
[15]	FPGA (Xilinx)	256	256	-	17	1

Particularly, Table II presents collected data relating to not only performance of hardware-based MFCC but also limitation of applications. Most approaches in Table III present fixed configurations which are compatible with exact applications. Another aspect that needs to be concerned that such hardware targets are usually built on FPGA to be able to reconfigure feasibly that is cause of low performance in comparison to ASIC-based designs. From such circumstances, a high performance dynamic ASIC-based MFCC architecture in 130nm technology is proposed in this work so to halt almost issues relating to hardware approaches. To be more specific, utilizing diverse range of techniques such as IEEE 754 floating-point format, Booth algorithm, internal memories, multi state machines, ability of setting parameters through AHB interface and pipeline technique makes effort to achieve high performance in comparison to other designs. Besides, ASR system integrated proposed MFCC hardware is also completed to confirm high performance of such architecture.

The rest of paper is organized follows. Section II proposes MFCC architecture. Next, Section III describes the experiment results to witness the successes. Finally, Section IV consults the paper.

II. HARDWARE-BASED MFCC ARCHITECTURE

A. MFCC mathematic model

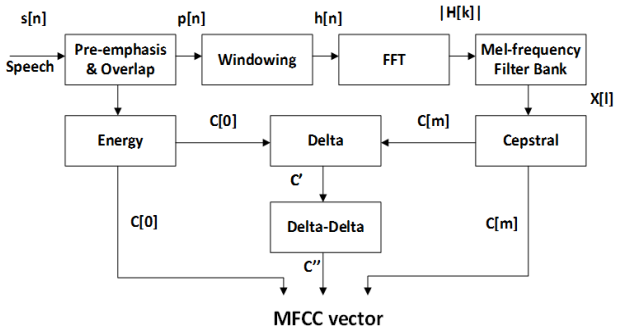

Fig. 1. General MFCC model

In Fig. 1, perspective of MFCC structure is described by general blocks called as Pre-emphasis filter, Energy calculator, Window filter, FFT implement, Cepstrum conductor, Delta implement accompanying following mathematic equations.

1) Pre_emphasis & Overlap:

Pre_emphasis block is used to amplify the signal at high frequencies. In time domain, relationship between output and input is shown in (1), with $s[n]$ is the *nth* sample of speech signal and $p[n]$ is the *nth* sample after going through Pre- emphasis filter. After filtering by Pre- emphasis, n samples are separated into many frames with different number of sample per frame and different percentage of overlapping which based on setting data.

$$p[n] = s[n] - 0.97 \cdot s[n-1] \tag{1}$$

2) Energy

Energy of each frame is also a typical component of MFCC. It is calculated as the logarithm of the signal power before pre-emphasis as (2), where $C[0]$ is the first component of C.

$$C[0] = \log\left(\sum_{n=0}^{N-1} s^2[n]\right) \tag{2}$$

3) Windowing

Window filter are applied to increase the continuity characteristic between mutual frames. One of the most window filter commonly used in speech recognition is Hamming window, where N is the length of the window and it is equal to the length of the frame and $h[n]$ is defined as the result of this block.

$$h[n] = p[n] \cdot \left\{0.54 - 0.46 \cdot cos\left(\frac{2\,\pi\,n}{N-1}\right)\right\} \tag{3}$$

4) FFT

Fast Fourier Transform (FFT) is used to transfer signal from time domain to frequency domain. Actually, it is Discrete Fourier Transform (DFT) which achieves high performance with certain conditions that spectrum are evaluated at discrete frequencies. FFT algorithm only requires computational complexity ratio of $N\log(N)$, whereas DFT calculations require the bigger ratio of N^2. FFT transform is one of the most time consuming phases as well as the resources of the system. Hence, FFT hardware approach will effect highly to the entire MFCC architecture. At final state of FFT, DFT results in $H[k]$ from input $h[n]$, then $|H[k]|$, magnitude of $H[k]$ in each frame, is calculated as (4).

$$H[k] = \sum_{n=0}^{N-1} h(n) \cdot e^{j\frac{2\pi\,n}{N}}$$

$$|H[k]| = \sqrt{\left(Re(H[k])\right)^2 - \left(Im(H[k])\right)^2} \tag{4}$$

5) Mel filter

Bandwidth filter of conventional Mel scale in voice recognition includes many triangular band-pass filters which are distributed within a bandwidth signal. The logarithmic power spectrum on the Mel-scale is computed in order to use a filter bank consisting of L Mel filters. Where $l = 0, 1, \ldots, L-1$; $W_l[k]$ is the *lth* triangular filter; k_{ll} and k_{lu} are the lower and upper bounds of the *lth* filter, respectively.

$$X[l] = \log\left(\sum_{k=k_{ll}}^{k_{lu}} |H[k]| \cdot W_l[k]\right) \quad (5)$$

For speech recognition, spectral boundary is more useful than spectral components, which is reasonable for using inverse Fourier transform to find the boundary of spectrum.

6) Cepstrum

Cepstrum is defined as the inverse Fourier transform of the power factor after taking the logarithm. It can be simplified as DCT transformation as (6).

$$C[m] = \sum_{l=1}^{L} X[l] cos\left(\frac{\pi\, m\, (l-0.5)}{L}\right) \quad (6)$$

7) Delta and Delta-Delta

The quality of the voice recognition system can be improved by adding more derivative features over-time to obtain the basic static parameters. In digital signal processing, first and second order derivative over time can be approximated by (7)

$$\begin{aligned} C_i' &= C_{i+2} + C_{i+1} - C_{i-1} - C_{i-2} \\ C_i'' &= C_{i+2}' + C_{i+1}' - C_{i-1}' - C_{i-2}' \end{aligned} \quad (7)$$

B. IEEE 754 floating-point format and Booth algorithm

From above Fig.1 and mathematic equations, it is acknowledged that if most arithmetic operations including addition, multiplication and logarithm apply fixed-point numbers, accuracy is affected significantly. As a result, in order to tackle this issue, floating-point format is proposed in this paper. Indeed, floating-point format is widely applied to variety of hardware designs so as to improve accuracy in [16]. To be more specific, Fig.2 presents such standard with total of 32 bit comprising of 1 sign bit, 8 exponent bits and 23 mantissa bits.

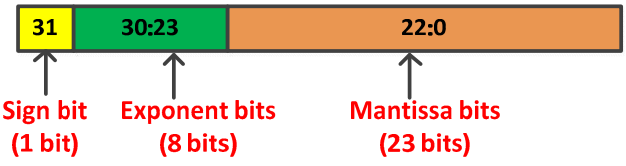

Fig. 2. IEEE 754 floating-point data format

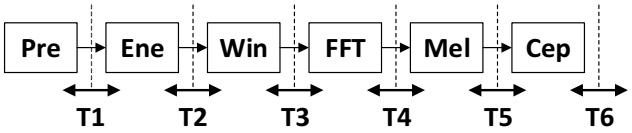

Fig. 3. Timing consumption without pipeline

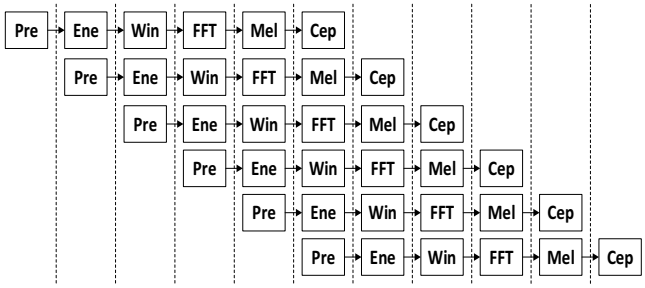

Fig. 4. Pipeline technique

However, applying floating-point format has to face high cost of 23 bit array multiplication. Consequently, Booth algorithm confirmed the high performance compared with array multiplication in [17] is applied to enhance speed of calculation.

C. Pipeline technique and internal memory

Based on MFCC model, if samples are transferred to the Pre-emphasis block firstly and then to next architectures, it consumes much time to solve consequentially before storing expected cepstrums to memory finally. Consequently, if we assume timing consumption per every block is from T1 to T6, total of timing consumption is calculated as Fig.3 and (8) with *n* frames of data, which negatively affect real-time criterion.

$$T = n * (T1 + T2 + \ldots + T6) \quad (8)$$

In order to tackle this issue, pipeline technique is applied to be able to execute all blocks at the same time. Particularly, only one timing parameter T_{Stage} established through AHB interface applied to all blocks to ensure that no block has to wait neighbor one. Moreover, because every block experiencing different formulas from section A for solving a frame of samples instead of one sample, internal memories is essential to store and load intermediate data. Hence, Fig.4 shows how to load and store data toward two neighbor blocks at the same time to ensure pipeline process correctly.

D. Perspective of MFCC architecture

To be able to adapt to a wide range of applications, proposed MFCC architecture witnesses ability of reconfiguration feasibly but not have to modify design at RTL level. Below Table III presents detailed information towards range of configurations in comparison to other references.

Such these parameters are configured by data through AHB protocol that is reasonable for constructing AHB interface as Fig.5. ABH interface takes the role of slave AHB in AHB systems. Totally, 10 internal registers are used to store configured setting from AHB before triggering MFCC.

From such circumstances, complex MFCC architecture is described by Fig.6 and Fig.7. After finishing configuration through AHB, the final register *TRIGGER* is set one to start MFCC operation. Hence, both coefficient data and frame of data is read from external memories and solved simultaneously that is based on combination of pipeline technique and internal memories such Fig.12 until completing all samples and output MFCC feature to *MFCC result memory.*

TABLE III. RANGE OF PARAMETERS IN PROPOSED MFCC

Parameter	**Proposed MFCC**	**Reference MFCCs [8-15]**
Pre-emphasis coefficient	Any value	0.97, 0.975
Sample In Frame	25 to 1024	128, 160,256, 512, 1024
Overlap rate (%)	30 to 70	50
FFT Point	25 to 1024	128, 160,256, 512, 1024
Mel Filter number	1 to 63	20, 24,32, 33
Number of cepstrums	1 to 31	12,13,17,24
Delta's Order	Level 1 and Level 2	Level 1, level 2

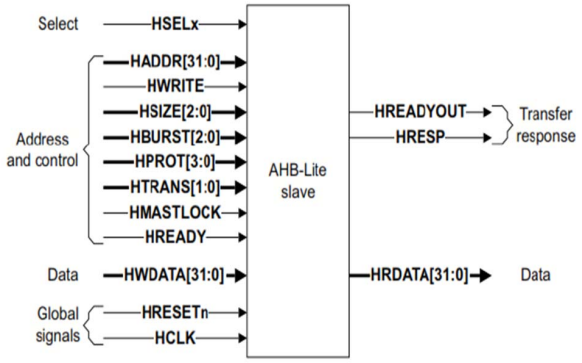

Fig. 5. AHB salve interface

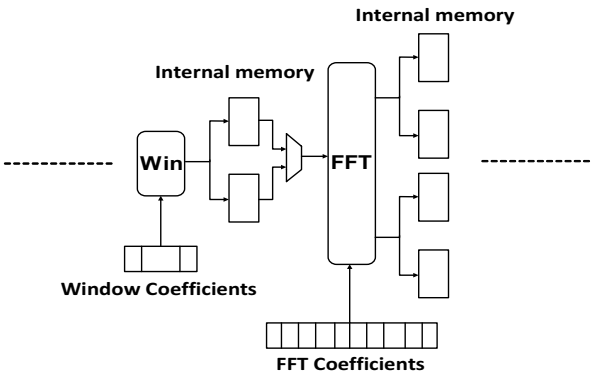

Fig. 6. Pipeline technique between block

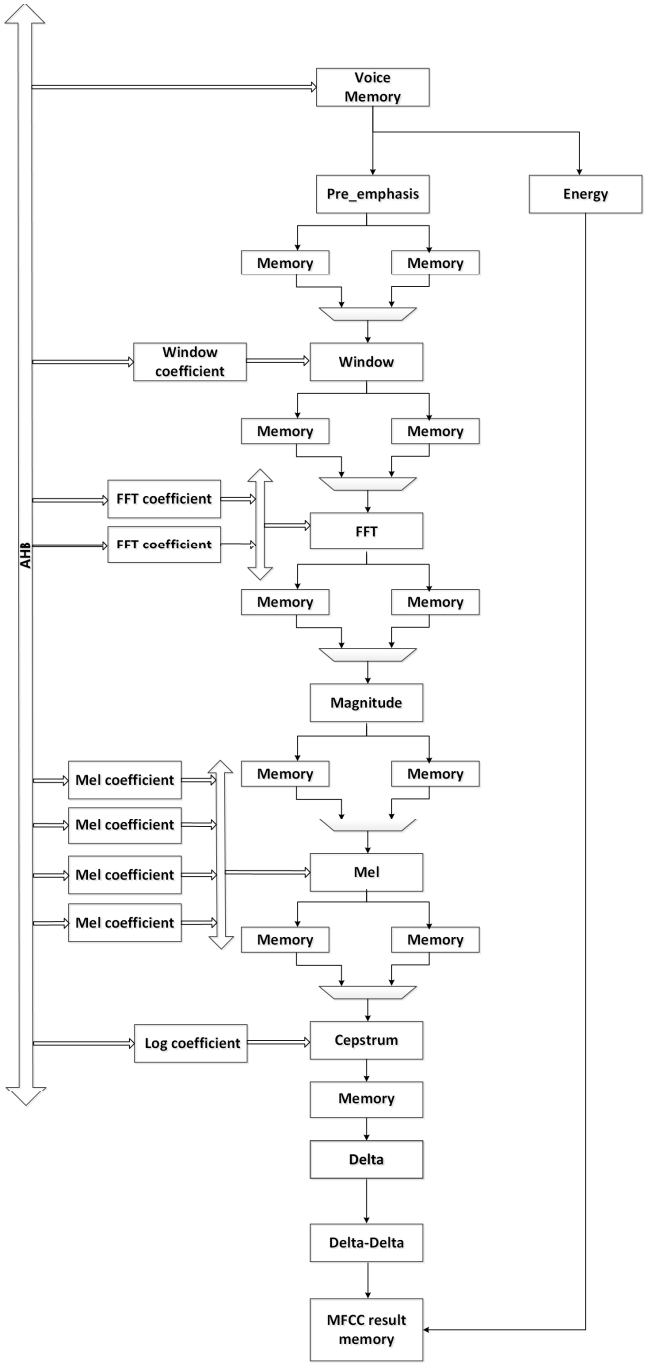

Fig. 7. Complex MFCC architecture

E. Hardware implement of components

Based on MFCC architecture and mathematic model of every block diagrams from section A, detailed hardware description for every block is described by following Fig.8 to Fig.13.

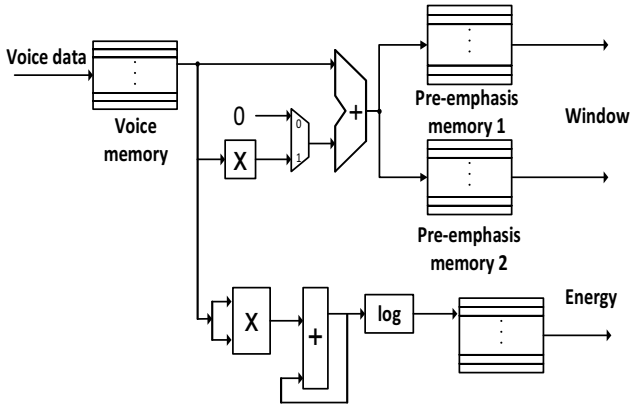

Fig. 8. Pre-emphasis and Energy blocks

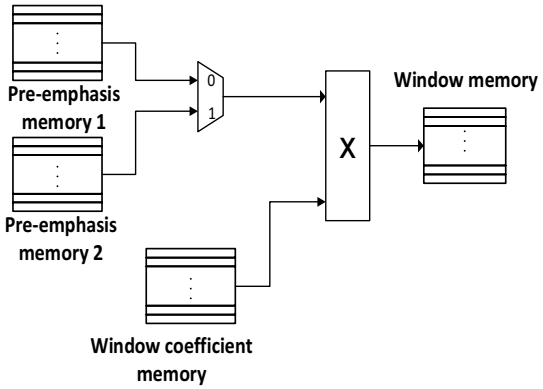

Fig. 9. Window block

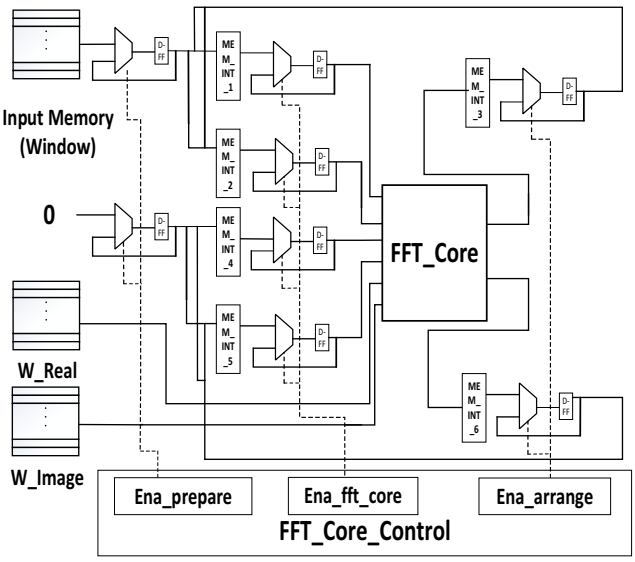

Fig. 10. Hardware architecture of FFT block

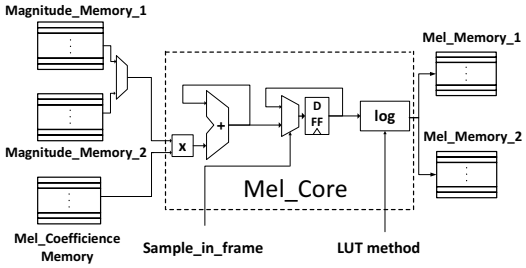

Fig. 11. Hardware architecture of Mel block

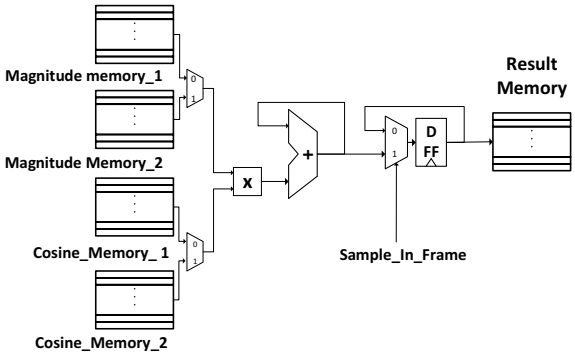

Fig. 12. Cepstrum block

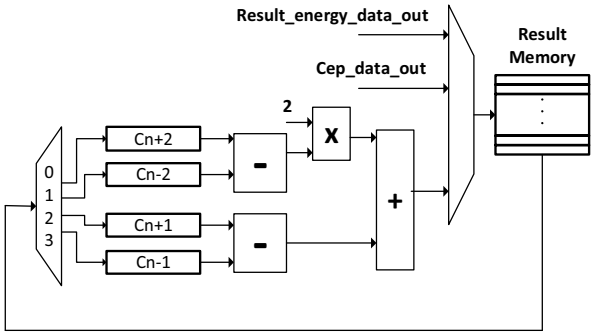

Fig. 13. Delta block

Towards such structure, FFT, Energy and Mel structures affect final performance of whole MFCC architecture not only accuracy but also timing consumption. Actually, although accuracy issue of FFT structure is solved by utilizing floating-point format, applying one butter fly structure to implement complete FFT with larger number makes such block consume the most time in comparison to other structure. As regards Energy and Mel structure, Logarithm with approaching internal memory as look up table reduces the accuracy much. Currently, only 4Kx32 memory is used to look up the value of log (MANTISA) as (9).

$$\text{Value} = 2^{\text{EXPONENT}} \times \text{MANTISSA}$$

$$\log(\text{Value}) = \text{EXPONENT} \times \log 2 + \log(\text{MANTISSA}) \quad (9)$$

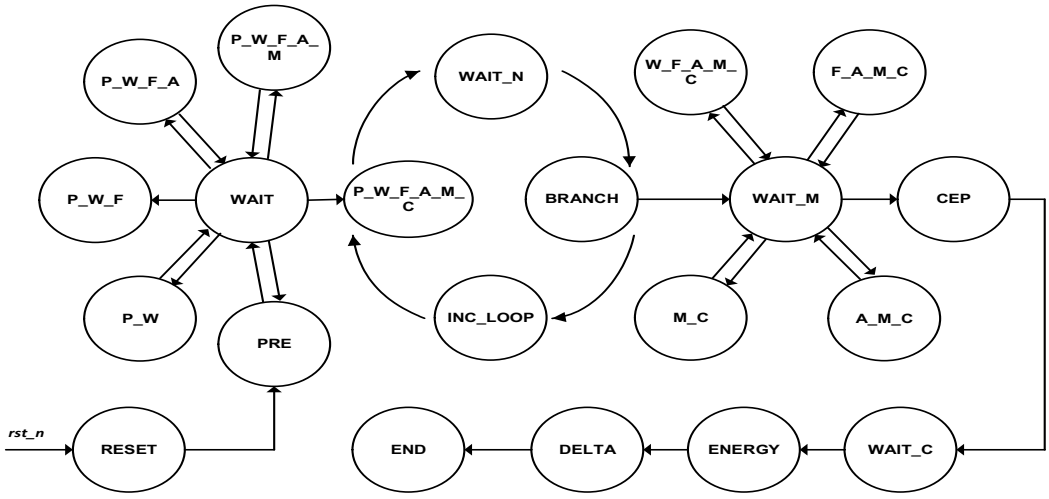

Fig. 14. Main state machine

TABLE IV. STATE MACHINE DESCRIPTION

Present state	Function
RESET	Reset
PRE	Enable Pre-emphasis block
P_W	Enable Pre-emphasis and Window block
P_W_F	Enable Pre-emphasis, Window and FFT block
P_W_F_A	Enable Pre-emphasis, Window, FFT and Magnitude block
P_W_F_A_M	Enable Pre-emphasis, Window, FFT, Magnitude and Mel block
P_W_F_A_M_C	Enable Pre-emphasis, Window, FFT, Magnitude, Mel and Cepstrum block
W_F_A_M_C	Enable Window, FFT, Magnitude, Mel and Cepstrum block
F_A_M_C	Enable FFT, Magnitude, Mel and Cepstrum block
A_M_C	Enable Magnitude, Mel and Cepstrum block
M_C	Enable Mel and Cepstrum block
CEP	Enable Cepstrum block
DELTA	Enable Delta and Delta-Delta block
ENERGY	Enable Energy block
WAIT	General counter for all block
WAIT_N	General counter for all block at N loops
WAIT_M	General counter for all block at M loops
WAIT_C	General counter for final Ceptrum stage
BRANCH	Check the number of necessary loops for P_W_F_A_M_C state
INC_LOOP	Increase number of loops
END	Finish

Such above architectures are connected to perform complete MFCC hardware, but it requires a complex controller structure in order to control harmoniously its behavior based on Fig. 4. In Fig. 14 and Table IV, detailed process of how to control MFCC is presented obviously. Actually, in every state, substates are called to control behavior of complex submodules such as Pre-emphasis, Window filter, FFT, Ceptrums, Energy and Delta calculations.

III. EXPERIMENTS AND RESULTS

In order to satisfy essential silicon requirements correctly, full ASIC design flow with 130nm technology is applied in this work, in which function of MFCC is compared to Matlab model with many different levels such as RTL, gate level netlist (pre_layout and post_layout). Final performance such as area, timing report or power is confirmed after finishing physical design step. Nonetheless, Design For Test (DFT) step is not inserted into this flow due to unnecessary requirement with MFCC architecture as IP level. As regards verifying environment, a multi-language environment comprising of Perl, Bash-Shell, Verilog are established. To be more detail, a setup file as Fig.15 containing necessary parameter as Table III is read firstly. Following every line of such file, such parameters are called by a CPU model, which transfers them through ABH to internal registers in MFCC architecture. Based such convenient environment as Fig.16, a wide range of MFCC are verified automatically and compared with Matlab model exactly.

	#Frame_num	Sample_in_frame	FFT_num	FFT_Stage_Number	Mel_num	Cep_num
Config_1	7'd24	11'd320	11'd512	4'd9	6'd63	7'd31
Config_2	7'd24	11'd320	11'd512	4'd9	6'd12	7'd22
Config_3	7'd24	11'd320	11'd512	4'd9	6'd50	7'd21
Config_4	7'd16	11'd400	11'd512	4'd9	6'd63	7'd31

Fig. 15. Some MFCC configurations in a configured file

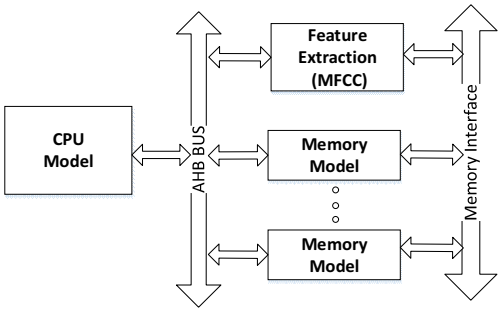

Fig. 16. Verification environment

A. Accuracy estimation

TABLE V. MAXIMUM MFCC CONFIGURATION

Sample in frame	Overlap rate (%)	FFT num	Mel filters	Cep	Delta order
320	50	512	63	31	Two levels

Through above verifying environment, the largest configuration as Table IV has experienced maximum measurement uncertainty with $4.17x10^{-4}$ for Energy, $4.31x10^{-4}$ for Cepstrum and 0.0013 for Delta in comparison to one in [13] with an average of 10^{-3}. This configuration has also witnessed delay 2.348 ms with maximum achieved frequency at 500 MHz, which adapt real-time requirement.

B. Hardware performance

After accomplishing all steps of ASIC design flow in Fig.17, final results of physical performance extracted from layout results such from Fig.17 to Fig.19 are presented and compared with other references as Table VI. Such this Table indicates that high efficiency of proposed architecture is experienced through maximum index of feature number at 96 opposed to only 12 and 48 in [9] and [11], which is reasonable for ability of reconfiguration feasibly. In addition to this, high silicon performance as regards frequency at 500 MHz as well as little area at 1.29x1.29 mm^2 is witnessed in comparison to other references.

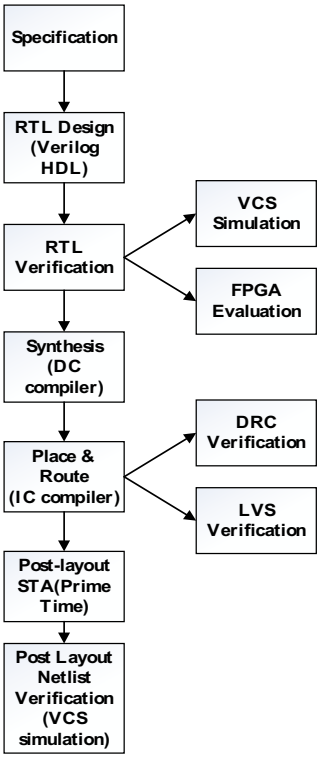

Fig. 17. ASIC design flow

TABLE VI. MFCC PERPORMANCE COMPARISON

Architecture	Proposed Architecture	[9]	[11]
Implementation	ASIC (130nm)	ASIC (0.6µm)	ASIC (0.18µm)
FFT points	8 - 1024	256	256
Mel	1 - 63	20	32
Cep	1 - 31	12	13
Feature Number	96 (Maximum configuration)	12	48
Area (mm^2)	1.29x1.29	3.2x3.3	6.5x3.5
Frequency (MHz)	500	50	30

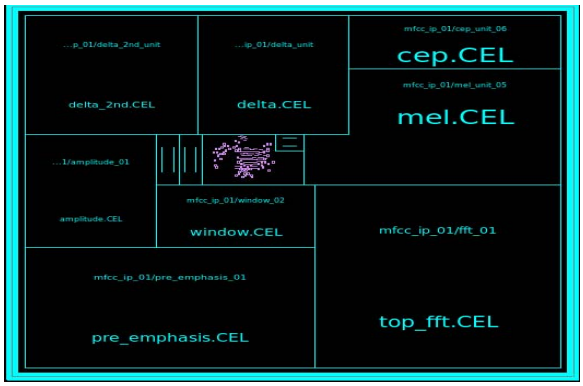

Fig. 18. Result of Floor Plan

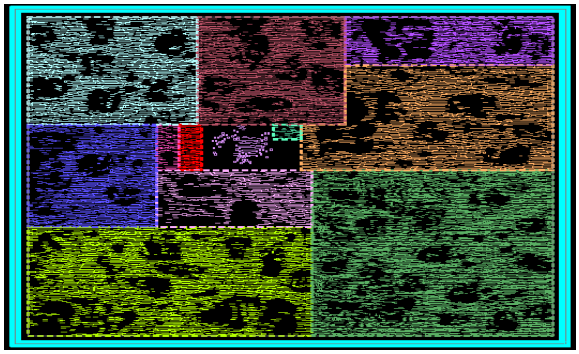

Fig. 19. Result of Placement

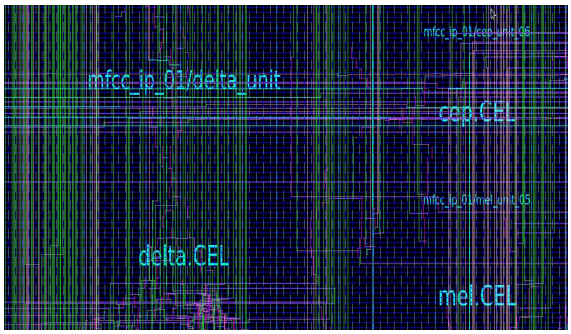

Fig. 20. Result of final Routing

C. Vietnam ASR application

TABLE VII. VIETNAM ASR APLICATION

Number of single words	50
Audio Input	Wav format
Feature Extraction	MFCC – Hardware implementation basing on proposed architecture
Training Algorithm	Hidden Markov Model (8 states, 4 mixture) – Software implementation by Matlab
Identification Algorithm	Viterbi – Software implementation by Matlab
Final result of recognition	98.5 %

Results of feature extraction basing on MFCC scheme obtained by hardware simulations are applied to Vietnam ASR, which is specified in Table VII. According to Table VII, Hidden Markov Model (HMM) algorithm is approached for training procedure and Viterbi algorithm is used for recognition state. Final results of recognition over 50 discrete words has confirmed high performance at 98.5% that indicated efficiency of feature extraction step as regards complete ASR.

IV. CONCLUSION

In this paper, a dynamic ASIC-based MFCC design is proposed to confirm the effective performance compared with other schemes. The promoted MFCC hardware architecture not only solves many applications basing on adapt a wide range of requirement such as accuracy problem, real-time issues, ability of reconfiguration feasibly but also adapts silicon requirements such high frequency matching AHB standard, reused ability as well as resource limitation. The obtained experimental results validated on 130 nm technology have also experienced low cost design for final fabrication and ability of integration towards entire systems. As regards future works, a complete hardware-based ASR system is our target to achieve powerful integrated circuit. Moreover, other applications integrating proposed MFCC will be also implemented to estimate the MFCC hardware performance obviously.

V. REFERENCES

[1] Veton Z. Këpuska, Mohamed M. Eljhani, Brian H. Hight, "Wake-Up-Word Feature Extraction on FPGA," World Journal of Engineering and Technology, 2014.

[2] Eslam Mansour mohammed,Mohammed Sharaf Sayed, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification," Signal Processing, Image Processing and Pattern Recognition, vol. 6, 2013.

[3] Ibrahim Patel, Dr. Y. Srinivas Rao, "Speech Recognition Using Hmm With Mfcc- An Analysis Using Frequency Specral Decomposion Technique," Signal & Image Processing : An International Journal, Vol. 1, 2010.

[4] A. Bala, "Voice Command Recognition System Based On Mfcc And Dtw," International Journal Of Engineering Science And Technology, Vol. 2, No. 12, Pp. 7335-7342, 2010.

[5] Anand Mantri, Mukesh Tiwari, Jaikaran Singh, "Development of FPGA based Human Voice Recognition System with MFCC feature," International Journal of Engineering Trends and Technology (IJETT), vol. 8, 2014.

[6] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, Wang Lihao , " Speaker Recognition Based on Dynamic MFCC Parameters," Image Analysis and Signal Processing IEEE, pp. 406 - 409, 2009.

[7] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications, vol. 10, no. 3, 2010.

[8] Mohammed Bahoura, Hassan Ezzaidi, "Hardware Implementation Of Mfcc Feature Extraction For Respiratory Sounds Analysis," Vol. Signal Processing And Their Applications, Pp. 226 - 229, 2013.

[9] Jia-Ching Wang, Jhing- Fa Wang, Yu-Sheng Weng, "Chip Design Of Mel Frequency Cepstral Coefficients," Acoustics, Speech, And Signal Processing, Vol. 6, Pp. 3658 - 3661, 2000.

[10] Jihyuck Jo, Hoyoung Yoo, In-Cheol Park, "Energy-Efficient Floating-Point Mfcc Extraction Architecture For Speech Recognition Systems," Transactions On Very Large Scale Integration (Vlsi) Systems, No. 99, 2015.

[11] E. Cornu, "An ultra low power, ultra miniature voice command system based on Hidden Markov Models," in Acoustics, Speech, and Signal Processing, IEEE, Orlando, FL, USA, 2002.

[12] Wei Han, Cheong-Fat Chan,Chiu-Sing Choy, "An Efficient Mfcc Extraction Method In Speech," In Ieee International Symposium, Island Of Kos, 2006.

[13] Ngoc-Vinh Vu, Jim Whittington, Hua Ye, John Devlin, "Implementation of The MFCC Front-end for Low-cost Speech Recognition Systems," in IEEE International Symposium, Paris, 2010.

[14] S.M. Ahadi, H. Sheikhzadeh, R.L. Brennan, G.H. Freeman, "An Efficient Front-End For Automatic Speech Recognition," In 10th IEEE International Conference On Electronics, Circuits And Systems, 2003.

[15] Phaklen EhKan, Timothy Allen, Steven F. Quigley, "FPGA Implementation for GMM-Based Speaker Identification," in International Journal of Reconfigurable Computing, 2011.

[16] Damak A, Krid M, Masmoudi D.S, "Neural Network Based Edge Detection with Pulse Mode Operation and Floating Point Format Precision," Design and Technology of Integrated Systems in Nanoscale Era, pp. 1-5, 2008.

[17] Shaifali, Ms. Sakshi, , "Comparison of IEEE-754 Standard Single Precision Floating Point Multiplier's," International Journal of Emerging Trends in Electrical and Electronics (IJETEE), vol. 1, pp. 900-904, March 2013.