Advances in Intelligent Systems and Computing

Volume 538

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

More information about this series at http://www.springer.com/series/11156

Editors
Masato Akagi
School of Information Science,
Area of Human Life Design
Japan Advanced Institute of Science
and Technology
Nomi-shi, Ishikawa
Japan

Thanh-Thuy Nguyen
Department of Computer Science
VNU University of Engineering
and Technology
Hanoi
Vietnam

Duc-Thai Vu
Faculty of Information Technology
Thai Nguyen University of Information
and Communication Technology
Thai Nguyen
Vietnam

Trung-Nghia Phung
Thai Nguyen University of Information
and Communication Technology
Thai Nguyen
Vietnam

Van-Nam Huynh
School of Knowledge Science,
Area of Knowledge Management
Japan Advanced Institute of Science
and Technology
Nomi-shi, Ishikawa
Japan

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-49072-4 ISBN 978-3-319-49073-1 (eBook)
DOI 10.1007/978-3-319-49073-1

Library of Congress Control Number: 2016955568

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Hardware Implementation of MFCC Feature Extraction for Speech Recognition on FPGA

Van-Lan Dao(✉), Van-Danh Nguyen, Hai-Duong Nguyen, and Van-Phuc Hoang

Faculty of Radio-Electronics, Le Quy Don Technical University, Hanoi, Vietnam
kqha1025@gmail.com, vandanhktqs@gmail.com, mta.haiduongnguyen@gmail.com, phuchv@mta.edu.vn
http://www.lqdtu.edu.vn

Abstract. In this paper, an FPGA-based Mel Frequency Cepstral Coefficient (MFCC) IP core for speech recognition is presented. The implementation results on FPGA show that the proposed MFCC core achieves higher resource usage efficiency compared with other designs.

1 Introduction

Nowadays, speech recognition and speaker recognition researched and applied a lot of human life [1,2,4–7]. Feature extraction of speech is one of the most important issues in the field of speech recognition or speaker recognition. This phase is proceeded right after the input speech is pre-emphasized and windowed. There are two dominant acoustic measurements of speech signal. One is the parametric approach, which is developed to match closely the resonant structure of the human vocal tract that produces the corresponding speech sound. It is mainly derived from linear predictive analysis, such as LPC-based cepstrum (LPCC). The other one is the non-parametric method modeling the human auditory perception system. Mel frequency cepstral coefficients (MFCCs) are utilized for this purpose [8]. In recent studies of speech recognition system, the MFCC parameters perform better than others in the recognition accuracy [9,10]. Figure 1 presents the general model of a speech/speaker recognition. The objective of this paper is to design a low area MFCC core on FPGA with an improved MFCC algorithm for speech recognition and speaker recognition applications.

The rest of this paper is organized as follows. Section 2 describes the MFCC algorithm and Sect. 3 shows the MFCC core design. Section 4 presents the implementation results and finally, Sect. 5 concludes of the paper.

2 MFCC Algorithm

The extraction of the best parametric representation of acoustic signals is an important task to produce a good recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior.

M. Akagi et al. (eds.), *Advances in Information and Communication Technology*, Advances in Intelligent Systems and Computing 538, DOI 10.1007/978-3-319-49073-1_27

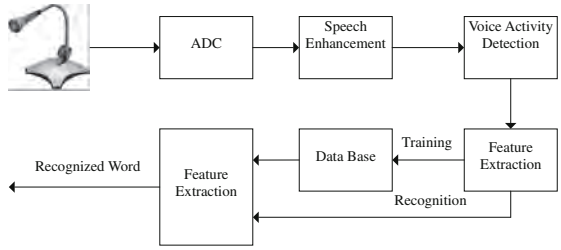

Fig. 1. General model of a speech/speaker recognition.

MFCC is based on the human hearing perception which cannot perceive frequencies over 1 KHz. In other words, MFCC is based on known variation of the human ear's critical bandwidth with frequency [8–10]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000 Hz. A subjective pitch is present on Mel frequency scale to capture the important phonetic characteristic in speech.

The overall block diagram for MFCC algorithm is shown in Fig. 2 [11].

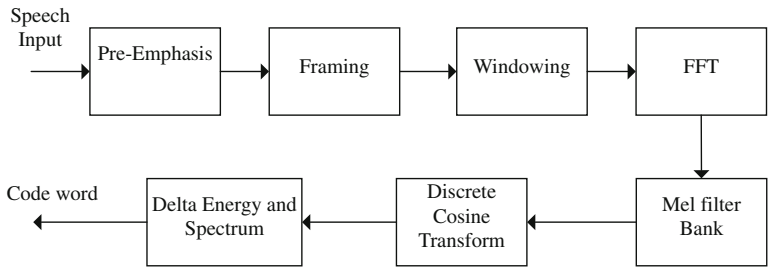

Fig. 2. MFCC block diagram.

MFCC consists of seven computational steps as shown in Fig. 2. The mathematical expression of each step is discussed briefly as follow.

Step 1: Pre-emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - aX[n-1] \tag{1}$$

Lets consider $a = 0.95$, which make 95 % of any one sample is presumed to originate from the previous sample.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 ms to

40 ms. The voice signal is divided into frames of N samples. Adjacent frames are being separated by $M(M < N)$. Typical values used are $M = 100$ and $N = 256$.

Step 3: Hamming windowing
Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

If the window is defined as $W(n), 0 \leq n \leq N-1$ where
N = number of samples in each frame
$Y(n)$ = Output signal
$X(n)$ = input signal $W(n)$ = Hamming window, then the result of windowing signal is shown below:

$$Y[n] = X[n].W[n] \tag{2}$$

$$W(n) = 0.54 - 0.46 * \cos(\frac{2\pi n}{N-1}); 0 \leq n \leq N-1; \tag{3}$$

Step 4: Fast Fourier Transform (FFT)
FFT converts each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse $U[n]$ and the vocal tract impulse response $H[n]$ in the time domain. This statement supports the equation below:

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w) \tag{4}$$

If $X(w)$, $H(w)$ and $Y(w)$ are the Fourier Transform of $X(t)$ and $Y(t)$ respectively.

Step 5: Mel Filter Bank Processing
The frequency range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in Fig. 3 is then performed.

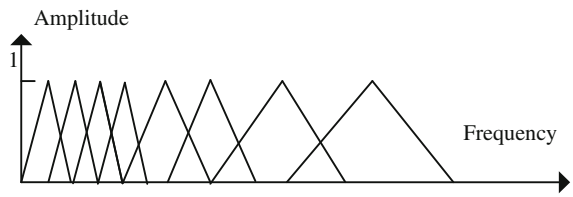

Fig. 3. Mel scale filter bank.

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular

in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [13]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in Hz:

$$Mel(f) = 2595 * \log 10(1 + {}^{f}/_{700}) \tag{5}$$

Step 6: Discrete Cosine Transform
This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Step 7: Delta Energy and Delta Spectrum
The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time axis. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t_1 to time sample t_2, is represented by the below equation:

$$Energy = \sum_{t=t_1}^{t_2} X^2(t) \tag{6}$$

Each of the 13 delta features represents the change between frames in the Eq. (7) corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \tag{7}$$

In [13], the authors presented an improved MFCC algorithm which will be used in this paper.

3 MFCC Hardware Core Design

MFCC architecture presents as shown in Fig. 4. In this work, the MFCC core is designed with 80-point sub-frame and 128-point FFT. Data input is speech signal samples saved in RAM. The controller block is a finite state machine including two states as show Fig. 5.

4 Implementation Results

The MFCC core was implemented with VHDL code and implemented in Xilinx Virtex-VI-LX240T and Spartan 3A DSP 1800 FPGA platforms. The FPGA implementation results are shown in Table 1 in which the MFCC core area can be reduced lower than other designs. Compared to other designs, this work can reduce the number of DSP48E1s significantly and numbers of slice LUTs/Registers are also lower than in [14–16].

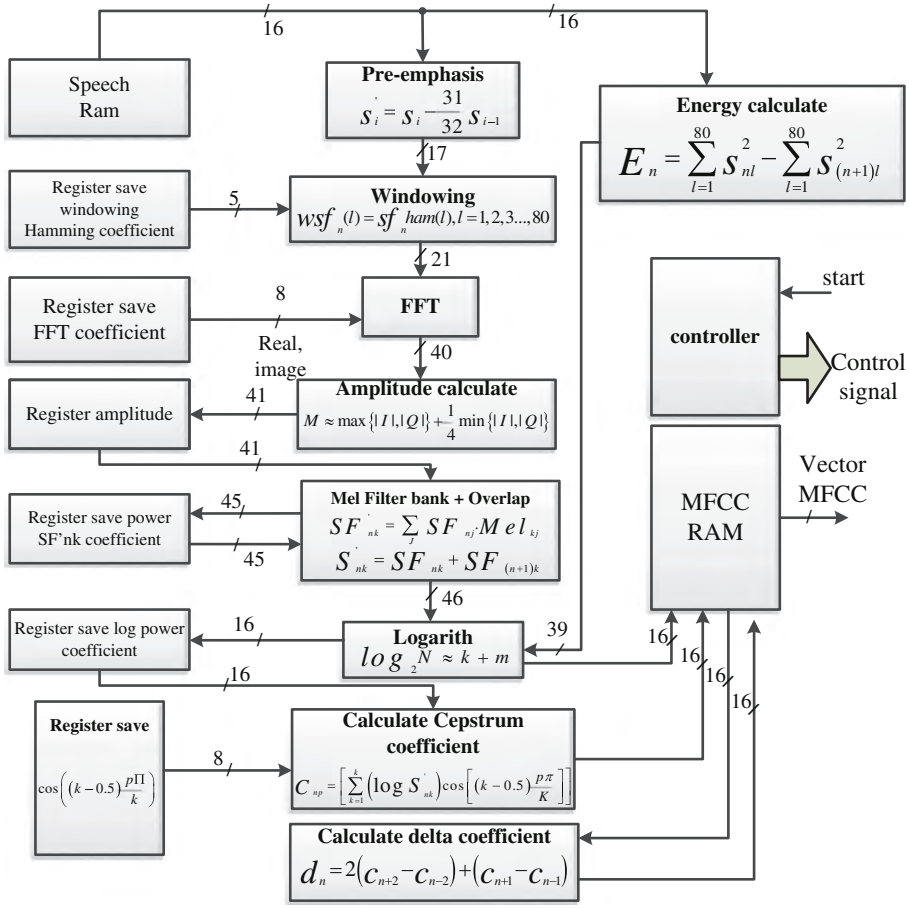

Fig. 4. Mel scale filter bank.

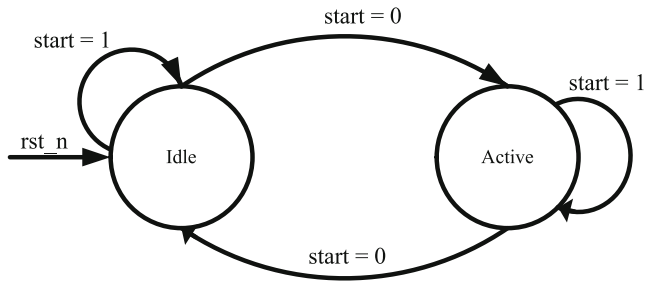

Fig. 5. Finite state machine for controller block.

Table 1. Implementation results of MMFC core on Xilinx FPGA.

FPGA device	Virtex 6-lx240T [14]	Virtex 6-lx240T [15]	Spartan 3A DSP 1800 [16]	Virtex 6-lx240T (Our design)	Spartan 3A 1800 (Our design)
Number of Slice Registers	11577	13726	23251	11567	11335
Number of Slice LUTs	8193	14837	35104	7353	11581
Number of fully used LUT-FF pairs	-	-	39452	1228	10700
Number of Block RAM/FIFO	44	4	14	8	15
Number of DSP48E1s	82	57	33	10	11
Frequency (MHz)	155	65.5	-	139	52

5 Conclusion

This paper has presented an area efficient MFCC core for speech/speaker recognition applications. The implementation results in FPGA show the MFCC core area can be reduced significantly. In the future, we will optimize the power consumption and area for the MFCC core and apply it for speech/speaker recognition applications.

References

1. Boves, L., den Os, E.: Speaker recognition in telecom applications. In: Proceedings of the Interactive Voice Technology for Telecommunications Applications, Torino, pp. 203–208 (1998)
2. McLoughlin, I.V., Sharifzadeh, H.R.: Speech recognition engine adaptions for smart home dialogues. In: Proceedings of the International Conference on Information, Communications and Signal Processing, Singapore, pp. 1–5 (2007)
3. Marchetto, E., Avanzini, F., Flego, F.: An automatic speaker recognition system for intelligence applications. In: Proceedings of the European Signal Processing, Glasgow, pp. 1612–1616 (2009)
4. Selvan, K., Joseph, A., Anish Babu, K.K.: Speaker recognition system for security applications. In: IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 26–30 (2013)
5. Ajgou, R., Sbaa, S., Ghendir, S., Chamsa, A., Taleb-Ahmed, A.: Robust remote speaker recognition system based on AR-MFCC features and efficient speech activity detection algorithm. In: International Symposium on Wireless Communications Systems (ISWCS), Barcelona, pp. 722–727 (2014)
6. Malode, A.A., Sahare, S.L.: An improved speaker recognition by using VQ and HMM. In: Proceedings of the International on Sustainable Energy and Intelligent Systems (SEISCON 2012), Tiruchengode, pp. 1–7 (2012)
7. Lung, V.D., Truong, V.N.: Vietnamese speech recognition using dynamic time warping and coefficient of correlation. In: Proceedings of the International Conference on Control, Automation and Information Sciences (ICCAIS), Nha Trang, pp. 64–67 (2013)
8. Tuzun, O.B., Demirekler, M., Nakiboglu, K.B.: Comparison of parametric and non-parametric representations of speech for recognition. In: Proceedings, pp. 65–68 (1994)

9. Openshaw, J.P., Sun, Z.P., Mason, J.S.: A comparison of composite features under degraded speech in speaker recognition. In: Proceedings on Acoustics, Speech, and Signal Processing, vol. 2, Minneapolis, USA, pp. 371–374 (1993)
10. Vergin, R., O'Shaughnessy, D., Gupta, V.: Compensated mel frequency cepstrum coefficients. In: Proceedings on Acoustics, Speech, and Signal Processing, Minneapolis, USA, pp. 323–326 (1996)
11. Ibrahim, N.J., et al.: Quranic verse recitation feature extraction using Mel-frequency cepstral coefficients (MFCC). In: Proceedings of the International Colloquium on Signal Processing and Its Applications (CSPA), Kuala Lumpur, Malaysia (2008)
12. Price, J., Sophomore Student: Design an automatic speech recognition system using maltab. University of Maryland Estern Shore Princess Anne
13. Wang, J.-C., Wang, J.-F., Weng, Y.-S.: Chip design of mel frequency cepstral coefficients for speech recognition. In: Proceedings of the Advanced IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, vol. 6, pp. 3658–3661 (2000)
14. Wassi, G., Iloga, S., Romain, O., Granado, B.: FPGA-based real-time MFCC extraction for automatic audio indexing on FM broadcast data. In: Proceedings on Design and Architectures for Signal and Image Processing (DASIP), Krakow, pp. 1–6 (2015)
15. Bahoura, M., Ezzaidi, H.: Hardware implementation of MFCC feature extraction for respiratory sounds analysis. In: Proceedings of the International Workshop on Systems, Signal Processing and their Applications (WoSSPA), Algiers, pp. 226–229 (2013)
16. Ehkan, P., Zakaria, F.F., Warip, M.N.M., Sauli, Z., Elshaikh, M.: Hardware implementation of MFCC-based feature extraction for speaker recognition. In: Sulaiman, H.A., Othman, M.A., Othman, M.F.I., Rahim, Y.A., Pee, N.C. (eds.) Advanced Computer and Communication Engineering Technology. LNEE, vol. 315, pp. 471–480. Springer, Heidelberg (2015). doi:10.1007/978-3-319-07674-4_46