TELKOMNIKA, Vol.17, No.4, August 2019, pp.1914~1922
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i4.12615

FPGA-based implementation of speech recognition for robocar control using MFCC

Bayuaji Kurniadhani*[1], Sugondo Hadiyoso[2], Suci Aulia[3], Rita Magdalena[4]
[1,4]School of Electrical Engineering, Telkom University, Terusan Buah Batu, Bandung, Indonesia
[2,3]School of Applied Science, Telkom University, Terusan Buah Batu, Bandung, Indonesia
*Corresponding author, e-mail: kurniadhani.bayuaji@gmail.com[1], sugondo@telkomuniversity.ac.id[2], suciaulia@telkomuniversity.ac.id[3], ritamagdalena@telkomuniversity.ac.id[4]

Abstract

This research proposes a simulation of the logic series of speech recognition on the MFCC (Mel Frequency Spread Spectrum) based FPGA and Euclidean Distance to control the robotic car motion. The speech known would be used as a command to operate the robotic car. MFCC in this study was used in the feature extraction process, while Euclidean distance was applied in the feature classification process of each speech that later would be forwarded to the part of decision to give the control logic in robotic motor. The test that has been conducted showed that the logic series designed was precise here by measuring the Mel Frequency Warping and Power Cepstrum. With the achievement of logic design in this research proven with a comparison between the Matlab computation and Xilinx simulation, it enables to facilitate the researchers to continue its implementation to FPGA hardware.

Keywords*: Euclidean distance, FPGA, MFCC, speech recognition*

1. Introduction

The research on Automatic Speech Recognition (ASR) is still the most interest topic for its very important role in daily life such as in the number of works that have been conducted on SR (Speech Recognition) such as Smart Home [1-4], Artificial Intelligence such as human emotional classification based on speech recognition [5-9] and robotic application [10], in the field of Student Learning [11-13], in medical sector, SR is used to detect the stress level of a person [14]. Therefore, the optimization process in SR is still in the progress for obtaining the better accuracy such as with speech enhancement [15-17] or noise reduction [18-20]. Automatic Speech Recognition (ASR) is a signal recognition process of speech to be a number of word orders. This concept has been developed since 2000 started from the emergence of Hidden Markov Model-based CMU Sphinx-N, Google Voice Recognition in Androin in 2010. In 2012 Apple released its application named Siri [21]. Since the end of 2017, there have been many applicable applications of speech recognition along with the increasing use of speech technology in real life environment.

Of some applications of ASR above, speech features that are frequently used in the extraction process are [22, 23] LPC (Linear Predictive Codes), MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Prediction), and PLP-RASTA (PLP-Relative Spectra). A technical review to observe the performance of the speech feature extraction techniques (MFCC, LPC, PLP, GFCC) with the combination of its classification technique (DTW, HMM, MLP, SVM, and DT) have been tested for SR Tamil Spoken words by Vimala [24]. Based upon the result of the test, of 5 (five) varieties of the method of feature extraction and classification, GFCC method has been more excellent in comparison to other algorithms [25]. Gurban has also conducted an approach of the MFCC based audio visual SR and conducted the optimization with two methods: CMI (Conditional Mutual Information) algorithm and MIFS (Mutual Information Feature Selection) algorithm. The result was found at best in very noisy conditions. Another study on SR was conducted by Gupta [26] using LPC and LPCC as the feature extraction techniques with a result showing that in terms of speech identification, the LPC parameter was more precise compared to LPCC. On the other hand, in reliability and robustness, LPCC was more excellent. Research study group on Speech and Language Technology for three years successfully did a research on how the effect of stress was on

Received July 1, 2018; Revised February 23, 2019; Accepted March 27, 2019

the speech production with the approaches of HMM and MFCC [27]. Other SR was with hidden Marcov model that is the experiment conducted by Farsi [28], the result showed that the HMM process needed to still be done with GA (Genetic Algorithm) for a good result. To compare [27], another study on MFCC conducted by Mohan [29] was done by combining MFCC with DTW algorithm showed the good results in the process of SR with ten features for each word in training phase.

Based on the comparison from a number of studies above (MFCC, LPC, PLP, GFCC) to this study, MFCC was used for the process of feature extraction on the speech signal as commonly it had the high performance rate and low complexity [30]. In this research, a prototype of SR system was made to FPGA that later would be used as an input for the robotic car (further research). One of examples of word recognition system application can be done in a simple robot car. The robot is able to differentiate the word pronounced by the users. By applying the word recognition system in the robotic car, then the users could control the direction of robot motion without a need to touch the button or being close to the robot.

2. Mel Frequency Cepstrum Coefficients (MFCC)

Feature extraction is a process to determine a value or vector that can be used as an object or individual identifier that subsequently will be used in the classification process [31, 32]. MFCC analysis is a standard method [33] used to represent the parameter of sound signal. The mechanism of MFCC is based upon the difference of frequency that can be captured by human ears commonly stated in the scale of Mel (originated from Melody) in which the sound signal would be filtered in linear in Mel frequency scale that is for the low frequency less than 1 KHz and logarithmically for high frequency more than 1 KHz [34]. The block diagram for the process of feature extraction using MFCC is presented in Figure 1.

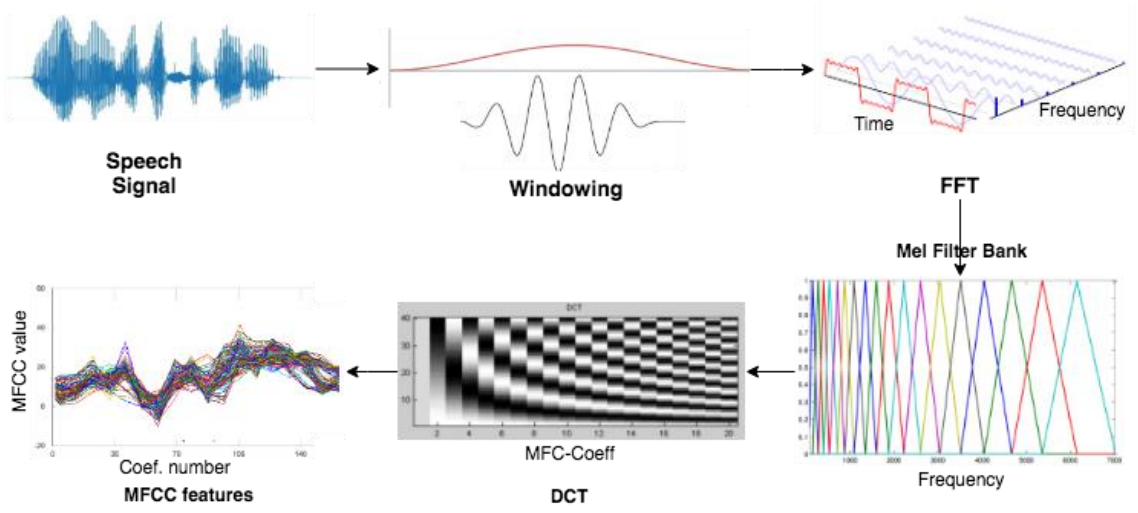

Figure 1. MFCC block diagram

As shown in Figure1 MFCC consists of some following computational steps:

Step 1: Preprocessing

As the initial step, the analog signal is passed through HPF emphasizes or known as pre-emphasis [35] Its purpose is to increase the energy of signal [36] thus, its output becomes the one as in (1) :

$$Y(n) = x(n) - \propto x(n-1) \quad (1)$$

$$where\ 0.9 \leq \propto \leq 1$$

Step 2: Frame Blocking and Windowing

Speech signal enters to the process of short frame or frame blocking with the duration of 10-30 ms on purpose for ADC. However, in its process, aliasing or spectral leakage or discontinue frequently occurs. For this, to cope with this problem, it must firstly start with windowing process before the signal continues to FFT phase. If defined, the window as function of $w(n), 0 \leq n \leq (N-1)$ is dependent upon the N value in which N refers to the number of samples in its frame. The input signal enters to the windowing process with the convolutional concept. The function of window used in this study is Hamming windowing as shown in (2) [37, 38]:

$$w(n) = 0.54 - 0.46\, cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

where $0 \leq n \leq (N-1)$.

The determination of the number of frame length should be in the fold of 2^N to facilitate the FFT process existing in the next block.

Step 3: DFT (Discrete Fourier Transform)

The output signal from Hamming window is in the time domain. To facilitate the measurement in the further process Mel-filter bank, then the signal is transformed mathematically from discrete time domain to the frequency using DFT method [39]. Meanwhile, the algorithm used to do transformation is called as FFT. Mathematically, DFT can be formulated as follows (3) [40, 41]:

$$X[k] = \sum_{n=0}^{N-1} x[n].\, e^{-\frac{j2\pi nk}{N}} \quad (2)$$

$$k = 0,1,2, \ldots., (N-1)$$

By doing FFT process, then it can obtain the value of X[k] as the result of the transformation from FFT representing each value of x(n) from the input signal. From the input signal in which each of the values is the representation of basic frequency from the input signal. X[k] is commonly called as spectrum or periodogram.

Step 4: Mel Frequency Filter Bank

This phase is the filtering process from frequency spectrum of X[k] in each frame using a number of M filter banks. This filter is made by following the perception of mel frequency scale represented to be the function of triangle filter function and mel scale frequency is obtained from the result of the conversion of linear frequency. For the linear frequency (f_{Hz})<1 kHz, it is converted to be f_{Hz} while if f_{Hz}>1 kHz, then it is converted into the (4) presented as follows [29], [35], [42] :

$$Mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \quad (3)$$

warping process to the signal in the frequency domain will result in the value of Mel Spectrum coefficients through the process as shown in (5) as follows:

$$X_i = log_{10}(\sum_{k=0}^{N-1} |X(k)|H_i(k)) \quad (4)$$

X_i is the value of frequency spectrum to *i*, N is the number of coefficients of FFT, and $H_i(f)$ is the filter value to i on the frequency spot *f*.

Step 5: Cepstrum

At this phase, Mel Cepstrum would be converted into the time domain using Discrete Cosine Transform (DCT). The result is called as Mel Frequency Cepstrum Coefficients (MFCC) as shown in (6) [41, 42] in which n is the number of coefficient and M is the number of filter banks. The word cepstrum is originated from the word spectrum that is reversed in its first syllable that is spec to ceps [43]. Cepstrum is the power spectrum obtained mathematically through the logarithmic computation [44, 45].

$$C_n = \sum_{i=1}^{M} (log\ X_i) \left[n \left(M - \frac{1}{2} \right) \frac{\pi}{M} \right] \quad (5)$$

3. Digital Design and Speech Recognition Simulation

This section discusses in detailed about the design of the block of speech recognition (SR) system. The main part consists of audio codec interface, feature extraction using MFCC and classification using Euclidean distance. It also discusses about the simulation of the digital logic design to FPGA using the Xilinx application. In the beginning of condition, the system will wait for the input from the switch to do training or recognizing. In the training mode, the coming sound will face the process of feature extraction using MFCC. The result of the feature extraction will be in the form of cepstral coefficient stored in the database. In the recognizing mode, the sound coming to the system will face the feature extraction process. Then, the coefficient cepstral of input sound would be compared the cepstral coefficient in the database using Euclidean distance method. The lowest value of Euclidean distance would be classified with the corresponding words. Further, the logic on the output pin would be conditioned in accordance with the words recognized. The value of output pin of this FPGA is used as the logic input on the driver motor to control the motion of the robotic car as shown in Table 1 as follows.

Table 1. Logic Output FPGA

Instruction	First Motor			Second Motor			Output
	En	Dir 1	Dir 2	En	Dir 1	Dir 2	Logic of FPGA
"Right"	1	1	1	1	1	1	111111
"Left"	1	0	1	1	1	0	101110
"Forward"	0	0	0	1	1	0	000110
"Backward"	0	0	0	1	0	1	000101
"Stop"	0	0	0	0	0	0	000000

The result of the simulation output then is compared to the result of the MATLAB. The system architecture of the speech recognition block is shown in Figure 2. The system architecture consists of four main blocks: microphone, audio codec, FPGA, driver motor and motor DC. Inside the FPGA, it is added with eight digital series acting as the cores of the system those are codec interface, preprocessing, control unit, clock divider, MFCC, database, Euclidean distance and output logic.

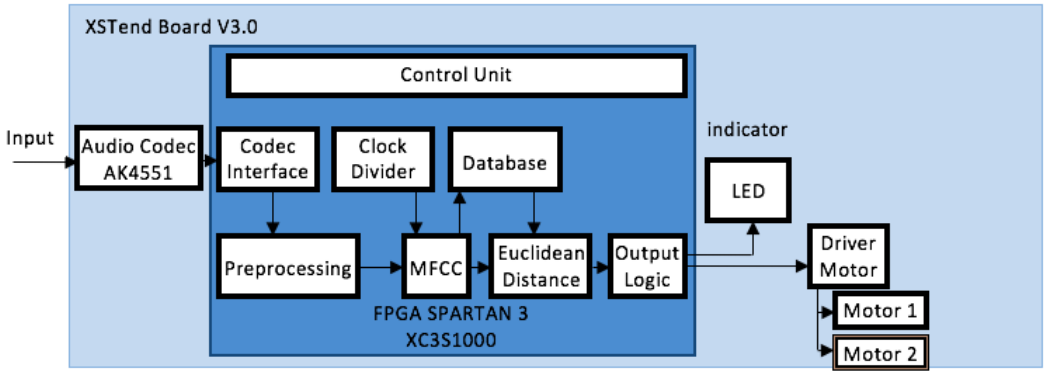

Figure 2. Architecture of proposed speech recognition

4. Results and Analysis

Simulation of the logic series design was conducted in Xilinx ISE Project Navigator. Also, computation simulation was done in MATLAB in order to obtain the comparing data from the designed system. The test was conducted by observing and comparing the data in each

sub-system of speech recognition. The synthesis of the logic circuit in this proposed system can be seen in Figure 3.

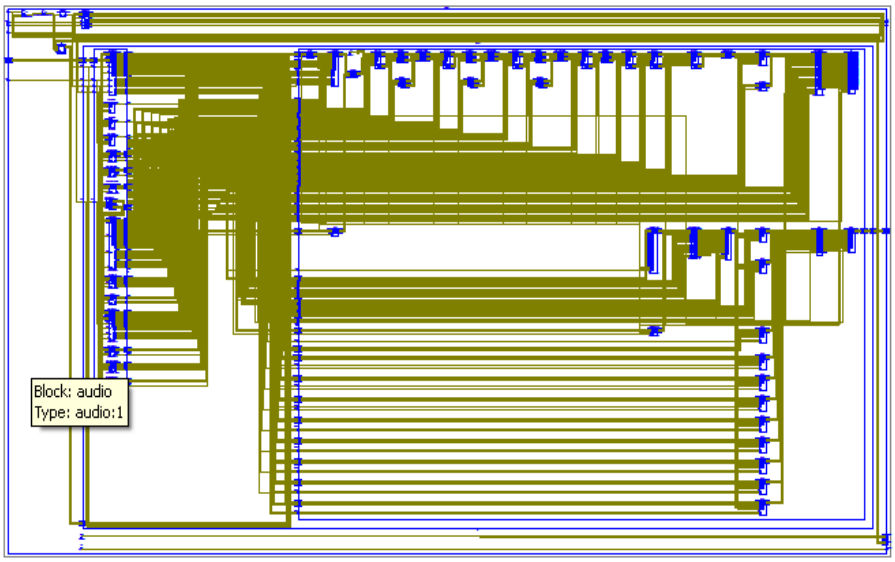

Figure 3. RTL Design schematic

4.1. Pre-emphasis Filter

This series functions to reduce the noise ratio in signal (SNR). This filter maintains the high frequencies on the spectrum eliminated in the process of sound production. From the result of simulation test compared by means of manual calculation, it has been found a similar value in each sample of signal. The results are shown in Table 2.

Table 2. Comparison of the Result of Manual Calculation on the Pre-emphasis Filter and Simulation Result

No	Manual Calculation	Xilinx Simulation
1	380	380
2	-166.25	-167
3	282.875	282
4	-25.1875	-26
5	-436.563	-437
6	431.5625	431
7	23.75	23
8	-519.25	-520
9	-92.1875	-93
10	-42.3125	-43

4.2. FFT

The computation FFT design on FPGA is done separately to result in 2 parts of output. From Figure 4 below can be seen that the result was not much different as the calculation operation in this design did not use the floating point system. The comparison of the results of FFT in the graphic form can be seen in Figure 4.

4.3. Mel Frequency Warping

Mel Frequency Warping functions as a filter from the spectrum of frequency of the output result of FFT. The multiplication process was done in parallel to 20 filter banks to make this process faster. The results from 20 filter bank are in the form of magnitude values as

shown in Figure 5. As seen in Figure 5 above, it can be seen that the feature in 5 - 10 from the result of Xilink simulation approached the MATLAB computation. This shows that the logic design made is precise.

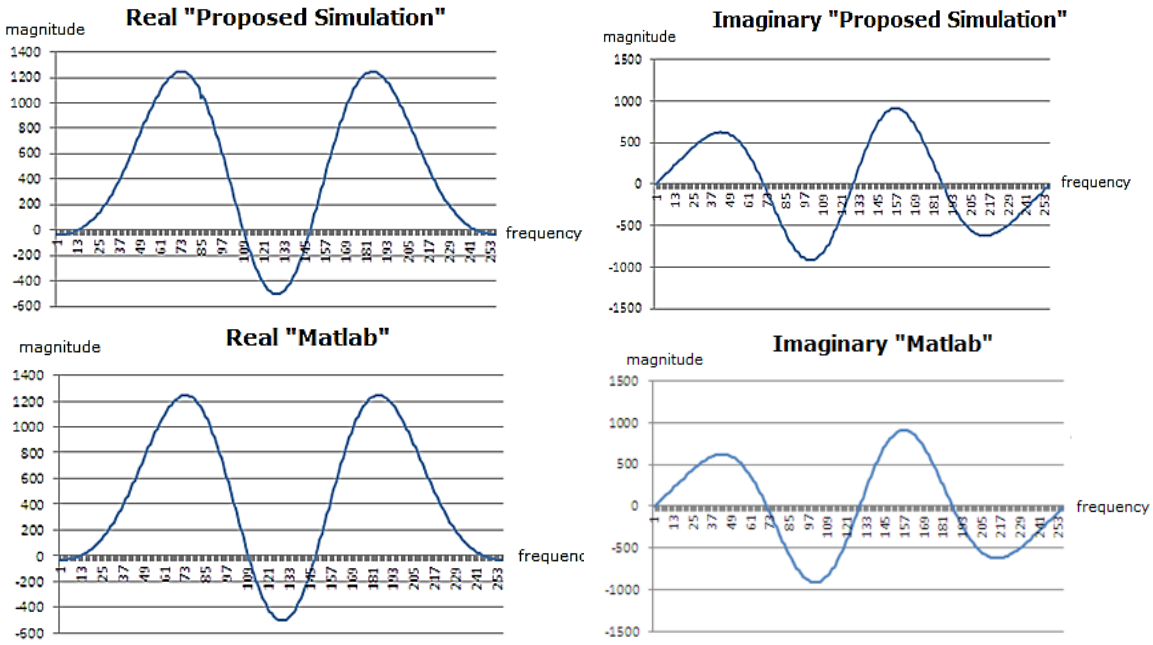

Figure 4. The Graph of the calculation result on 256-point FFT on the MATLAB and Xilinx Simulation

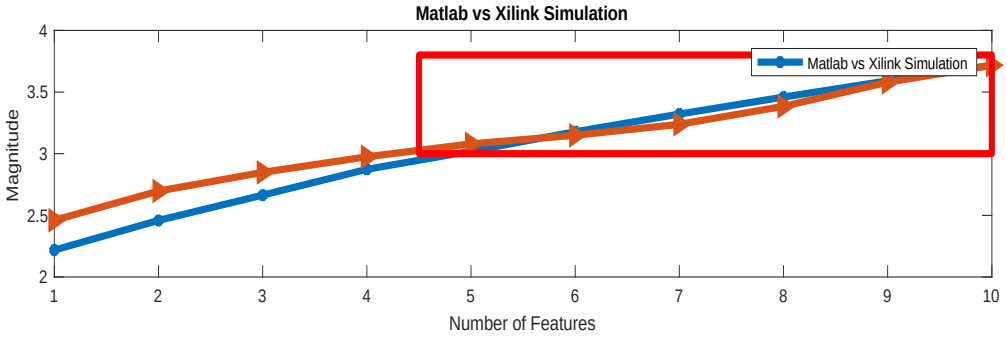

Figure 5. Graph of Mel Frequency Warping result on MATLAB and Simulation on Xilinx

4.4. Cepstrum

This block functions to convert the Mel cepstrum from the result of previous block into the time domain using Discrete Cosine Transform. The coefficient value was then changed into the representation of fixed point 16 bits and stored in ROM. This block also had RAM to store the output from the previous process to facilitate the process. The results of the block, if compared to the result of the calculation on MATLAB as shown in Figure 6, were seen different. This was because the calculation operation in this block involved the numbers that had some digits after the comma. Meanwhile, the representation of the numbers used did not have any accuracy in number; thus, rounding occurred to the representation of numbers closer.

4.5. Decision

12 coefficients from the result of feature extraction stored in the database at the testing phase will be compared to the coefficient of the sound input feature and then will be cut to do

the instruction as determined. The decision was taken by calculating the closest feature coefficient value using euclidean distance. The block of database made was used to store one sample of each instruction word. Five dataset of feature were saved in the database. Once obtaining the closest number, then decision was taken to regulate the control in the motor in the format of 4 bit data as shown in the following Figure 7.

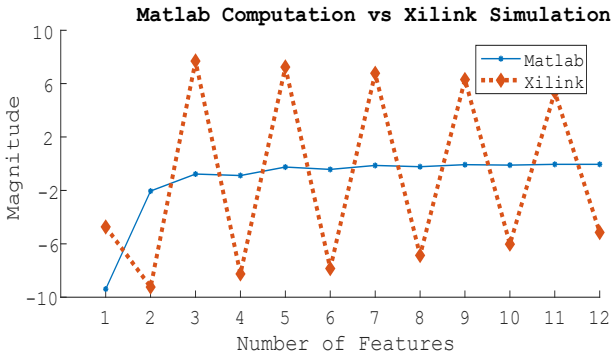

Figure 6. Graph of on the Calculation Result on MATLAB and Xilinx Simulation

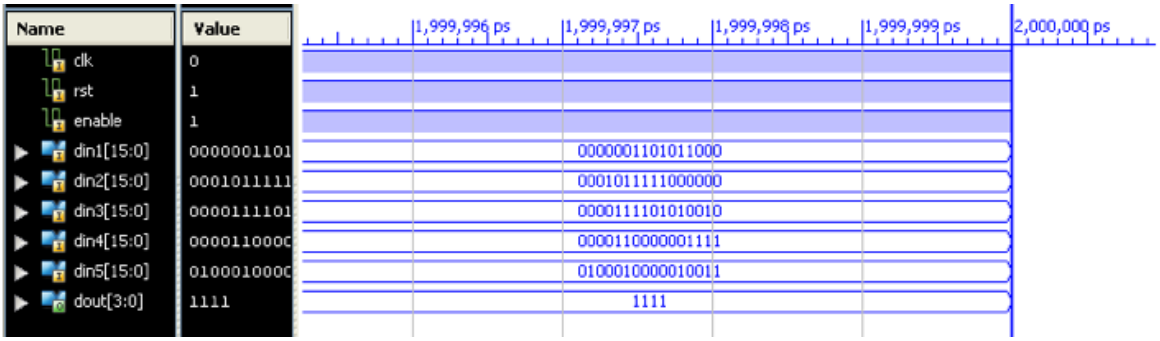

Figure 7. Result of the simulation on the output logic block on Xilinx

5. Conclusion

This research has successfully made a design and simulation of logic series in a speech recognition system using the MFCC method and euclidean distance to control the rate of robotic car. The MFCC method was used to obtain the feature from the command input in the form of sound consisting of right, left, forward, backward and stop. The result of the signal feature from MFCC furthermore was calculated for its similarity and compared with the signal feature in the database using euclidean distance to give the control logic in motor.

Process simplification was also done to obtain the resource of the FPGA memory as minimum as possible but still had a good performance. Validation has been conducted to test the excellence of the system made. This test was done by comparing the output value of logic design that has been made in each part of speech recognition components\with the calculation simulation on MATLAB. There was a difference in the value between the result of the computation of the logic series and the MATLAB as the calculating operation involved the numbers that had some digits after the comma. Meanwhile, the representation of the numbers used did not have any sufficient number accuracy; thus making the rounding occurred in the closest representative numbers. However, this difference value was relative insignificant and the system then still had a good performance to compare the signal features from one class to other as proven in the results. The next research will be done through the implementation to the FPGA board and analysis on the synthesis of logic series to determine the parameter of the performance in IC design.

References

[1] M Vacher, AFleury, F Portet, J-F Serignat, N Noury. Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living. *New Developments in Biomedical Engineering*. InTech. 2010.

[2] P Putthapipat, C Woralert, P Sirinimnuankul. *Speech recognition gateway for home automation on open platform.* International Conference on Electronics, Information, and Communication (ICEIC) 2018. Hawai, USA. 2018: 1–4.

[3] T Ayres, B Nolan. Voice activated command and control with speech recognition over WiFi. *Sci. Comput. Program.* 2006; 59(1–2): 109–126.

[4] A Mohanta, VK Mittal. *Human Emotional States Classification Based upon Changes in Speech Production Features in Vowel Region*. International Conference on Telecommunication and Networks (TEL-NET) 2017. Noida, India. 2017.

[5] KF Akingbade, OM Umanna, IA Alimi. Voice-Based Door Access Control System Using the Mel Frequency Cepstrum Coefficients and Gaussian Mixture Model. *Int. J. Electr. Comput. Eng.* 2014; 4(5): 643–647.

[6] W Rao *et al. Investigation of fixed-dimensional speech representations for real-time speech emotion recognition system.* International Conference on Orange Technologies (ICOT) 2017: 197–200.

[7] J Yadav, MS Fahad, KS Rao. Epoch detection from emotional speeach signal using zero time windowing. *Speech Commun.* 2017; 96(November): 142–149.

[8] HK Palo, MN Mohanty. Classification of Emotional Speech of Children Using Probabilistic Neural Network. *Int. J. Electr. Comput. Eng.* 2015; 5(2): 311–317.

[9] FE Gunawan, K Idananta. Predicting the Level of Emotion by Means of Indonesian Speech Signal. *TELKOMNIKA*. 2017; 15(2): 665–670.

[10] J Twiefel, X Hinaut, S Wermter, J Twiefel, X Hinaut, S. Wermter. *Syntactic Reanalysis in Language Models for Speech Recognition*. Int. Conf. Dev. Learn. Epigenetic Robot. Lisbon, Portugal. 2017: 215–220.

[11] D Recasens, C Rodríguez. Contextual and syllabic effects in heterosyllabic consonant sequences. An ultrasound study. *Speech Commun.* 2017; 96(December): 150–167.

[12] M Jia, J Sun, C Bao, C Ritz. Separation of multiple speech sources by recovering sparse and non-sparse components from B-format microphone recordings. *Speech Commun.* 2017; 96(May): 184–196.

[13] M Tahon, G Lecorve, D Lolive. Can we Generate Emotional Pronunciations for Expressive Speech Synthesis?. *IEEE Trans. Affect. Comput.* 2018; 3045: 1-1.

[14] K Li, S Mao, X Li, Z Wu, H Meng. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. *Speech Commun.* 2017; 96(September): 28–36.

[15] RK Kandagatla, PV. Subbaiah. Speech enhancement using MMSE estimation of amplitude and complex speech spectral coefficients under phase-uncertainty. *Speech Commun.* 2017; 96(November): 10–27.

[16] A Misra, JHL. Hansen. Modelling and compensation for language mismatch in speaker verification. *Speech Commun.* 2017; 96(January): 58–66.

[17] B Wiem, BM Mohamed anouar, P Mowlaee, B Aicha. Unsupervised single channel speech separation based on optimized subspace separation. *Speech Commun.* 2017; 96 (April): 93–101.

[18] R Soleymani, IW. Selesnick, DM. Landsberger. SEDA: A tunable Q-factor wavelet-based noise reduction algorithm for multi-talker babble. *Speech Commun.* 2017; 96(October): 102–115.

[19] Y Tang, Q Liu, W Wang, TJ Cox. A non-intrusive method for estimating binaural speech intelligibility from noise-corrupted signals captured by a pair of microphones. *Speech Commun.* 2017; 96(June): 116–128.

[20] F Saki, A Bhattacharya, N Kehtarnavaz. Real-time Simulink implementation of noise adaptive speech processing pipeline of cochlear implants. *Speech Commun.* 2017; 96(April): 197–206.

[21] S Naragan, MD Gupta. Speech Feature Extraction Technique: A Review. *Int. J. Comput. Sci. Mob. Comput.* 2015; 4(3): 107–114.

[22] N Dave. Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition. *Int. J. Adv. Res. Eng. Technol.* 2013; 1(VI): 1–5.

[23] Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, and Chungyong Lee. Optimizing feature extraction for speech recognition. *IEEE Trans. Speech Audio Process.* 2003; 11(1): 80–87.

[24] C Vimala and V. Radha. Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words. *Int. J. Comput. Sci. Inf. Technol.* 2014; 5(1): 378–383.

[25] M Gurban, JP Thiran. Information Theoretic Feature Extraction for Audio-Visual Speech Recognition. *IEEE Trans. Signal Process.* 2009; 57(12): 4765–4776.

[26] H Gupta, D Gupta. *LPC and LPCC method of feature extraction in Speech Recognition System*. Proc. 6th Int. Conf. Cloud Syst. Big Data Eng. Conflu. (2016). Noida, India: 498–502.

[27] HJM Steeneken, JHL Hansen. *Speech under stress conditions: overview of the effect on speech production and on system performance*. IEEE International Conference on Acoustics, Speech, and Signal Processing. Phoenix, USA. 1999: 2079–2082.
[28] H Farsi, R Saleh*. Implementation and optimization of a speech recognition system based on hidden Markov model using genetic algorithm*. Iranian Conference on Intelligent Systems (ICIS). Bam, Iran. 2014: 1–5.
[29] Bhadragiri Jagan Mohan and Ramesh Babu N. *Speech recognition using MFCC and DTW*. International Conference on Advances in Electrical Engineering (ICAE*E*). Vellore, India. 2014: 1–4.
[30] I McLoughlin. *Applied Speech and Audio Processing*. Cambridge: Cambridge University Press. 2009; 9780521519.
[31] J Yang, J Yang. Generalized K–L transform based combined feature extraction. *Pattern Recognit.* 2002; 35(1): 295–297.
[32] I Guyon, A Elisseeff. An Introduction to Feature Extraction. *Feature Extraction*. 2006; 207: 1–25.
[33] P Motlicek. Feature Extraction in Speech Coding and Recognition. *Report of PhD research internship in ASP Group*. 2003: 1–50.
[34] GSVS Sivaram, H Hermansky. Sparse Multilayer Perceptron for Phoneme Recognition. *IEEE Trans. Audio. Speech. Lang. Processing*. 2012; 20(1): 23–29.
[35] L Muda, M Begam, I Elamvazuthi. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *J. Comput.* 2010; 2(3): 138–143.
[36] A Přibilová. Preemphasis influence on harmonic speech model with autoregressive parameterization. *Radioengineering*. 2003; 12(3): 32–36.
[37] R Goldberg, L Riek. *A practical handbook of speech coders*, 1st ed. Boca Raton: CRC Press, 2000.
[38] MA Samad. A Novel Window Function Yielding Suppressed Mainlobe Width and Minimum Sidelobe Peak. *Int. J. Comput. Sci. Eng. Inf. Technol.* 2012; 2(2): 91–103.
[39] KR Ghule, RR Deshmukh. Feature-Extraction-Techniques-for-Speech-Recognition-A-Review. *Int. J. Sci. Eng. Res.* 2015; 6(5): 143–147.
[40] SC Joshi, AN Cheeran. MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition. *Int. J. Sci. Eng. Technol. Res.* 2014; 3(6): 1820–1823.
[41] SE Levinson. *Mathematical Models for Speech Technology*. Chichester, UK: John Wiley & Sons, Ltd. 2005.
[42] S Dhingra, G Nijhawan, P Pandit. Isolated speech recognition using MFCC and DTW. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering.* 2013; 2(8): 4085–4092.
[43] S Bedi, R Singh. Hamming Generalized Low Time Liftered Cepstral Analysis. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 2017; 7(6): 157–160.
[44] NA Michael. Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection. *J. Acoust. Soc. Am.* 1964; 36(2): 296–302.
[45] RM Martin, CL Burley. *Power Cepstrum Technique With Application to Model Helicopter Acoustic Data*. NASA Technical Paper 2586. 1986.