TELKOMNIKA, Vol.16, No.3, June 2018, pp. 925~933
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v16i3.5014

Wavelet Based Feature Extraction for the Indonesian CV Syllables Sound

Domy Kristomo*[1], Risanuri Hidayat[2], Indah Soesanti[3]
[1,2,3] Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada,
Jalan Grafika No.2, Yogyakarta 55281 Indonesia, telp/fax (0274) 547506
[1] Study Program of Computer Engineering, STMIK AKAKOM Yogyakarta,
Jalan Raya Janti 143 Karang Jambe Yogyakarta, 55198 Indonesia
*Corresponding author, e-mail: domykristomo@yahoo.com[1], risanuri@ugm.ac.id[2], indah@ugm.ac.id[3]

Abstract

This paper proposes the combined methods of Wavelet Transform (WT) and Euclidean Distance (ED) to estimate the expected value of the possibly feature vector of Indonesian syllables. This research aims to find the best properties in effectiveness and efficiency on performing feature extraction of each syllable sound to be applied in the speech recognition systems. This proposed approach which is the state-of-the-art of the previous study consist of three main phase. In the first phase, the speech signal is segmented and normalized. In the second phase, the signal is transformed into frequency domain by using the WT. In the third phase, to estimate the expected feature vector, the ED algorithm is used. The result shows the list of features of each syllables can be used for the next research, and some recommendations on the most effective and efficient WT to be used in performing syllable sound recognition.

Keywords*: Wavelet, euclidean distance, feature extraction*

1. Introduction

The attempt to realize an intelligent pattern recognition system requires the ancillary systems development that are effective, reliable, and efficient to be integrated well in an intuitive interaction system [1]. One of the pattern recognition support system that has been so much developed is a speech recognition system [1]-[3]. Challenges in creating a good speech recognition systems include feature extraction [3]-[4], namely how to find the unique features of a speech sound signal that distinguishes it from other speech signals so that a collection of these unique features which will constructing a reference database to identify each certain speech signal as an input command from the user. One of the feature extraction method is the Wavelet Transform (WT) which is the suitable method for exploring the frequency component of speech signals [3]. Computing the Euclidean Distance (ED) is a key part in many machine learning and template matching methods to find the closest members of the training set as well as to estimate the possibly feature vector. A speech recognition is the one of the pattern recognition support system that has been so much developed [2]-[4]. Feature extraction become a challenges in creating a good speech recognition systems [3]. There are many feature extraction method, so the most suitable method must be found to be used on specific types of sound signals [3].

There are several previous studies that combining WT and ED for analysing both one-dimensional (1-D) and two-dimensional (2-D) signal [5]-[12]. In [5], a new approach that combines the WT, Phase Space Recontruction (PSR), and ED, was proposed to classify the normal and the epileptic seizure EEG signals. The Daubechies 4 was used as a coefficient at the 1 through 5 level of decomposition. In [6], the ED and WT were used for analysis of the blood flow signal. The decomposition was done at one and three decomposition level by using coefficient of Daubechies 2, Morlet, Symlet 2, and Symlet 4. In [7], The ED was used to estimate the expected value of the possibly incomplete feature vectors. In [12], Hidayat et al. used the Discrete Wavelet Transform (DWT) on the 7th level decomposition to extract Indonesian vowels. The Daubechies 2, Coiflet 2, Symlet 5, Haar, Bioorthogonal 2.2, and The discrete Meyer wavelet were used as mother wavelet. The result of the study shows that the Haar wavelet is the best wavelet type used in the speech recognition process for all Indonesian

Received October 25, 2016; Revised March 2, 2018; Accepted April 7, 2018

vowel sounds. Recently [11], the study was done for extracting the Indonesian phonemes by using DWT and Wavelet Packet Transform (WPT) at 2nd through 4th level of decomposition by using Haar as mother wavelet. The result of this study showed that the DWT is a method that is more efficient and effective in extracting the Indonesian phonemes compared with the WPT as shown by the effectiveness ratio of 60% versus 40% and efficiency ratio of 57% versus 43% [11]. In the context of speech classification, there are several previous studies that combining feature extraction methods based on Wavelet [13]-[17], MFCC [16]-[18], LPC [16]-[17], LPCC [16], and the classifier methods such as MLP [13,18,19], HMM [20], GMM [19], and LDA [14,17].

However, the similar study only focused on the sounds of the Indonesian vowels [12] and phonemes [11]. This paper is a development of the previous studies which using the WT and the ED algorithm [11]-[12]. The frequency components of vowel (V) and phoneme were combined to form and estimate the frequency component of CV syllable pattern. This paper aims to explore WT as a tool for extracting feature of Indonesian consonant-vowel syllables and comparative study of the different wavelet coeficient for analysis Indonesian syllables sound signal. The reason of the selection of the mother wavelet refers to studies done by previous study. This work restricts the scope of research as far as the combination of the phonemes */m, n, r, s/* with the following vowels */a, i, u, e/* and also velar consonant */g/* with the following vowels. Although not all, but the selection of the phonemes is expected representing a large portion of existing phonemes. For example, the phoneme */g/* represent a velar sound, the phonemes /m, n/ represent nasal sounds, */r/* represents an alveolar trill sound, */s/* represents an alveolar fricative sound, and */t/* represents an alveolar stop sound. The experiment in effectitiveness and efficiency for velar */g/* with the following vowels was done separately.

2. Research Method

There are three main steps used in this study. The first step is preprocessing, this step aims to select a certain part of the signal that would like to be further processed and to recover speech signal level. Then, the signal is transformed into frequency domain by using the WT algorithm. At this part, the results are the frequency components and the magnitude of each possibly feature vector found. Then, the selection and testing process which uses ED algorithm are conducted to get the most reliable and possible features to be used in the speech recognition process.

2.1. Pre-processing

The speech sound signal was recorded by using a laptop with a microphone from two males and two female speakers in the open area with the minimal noise. Once it was done, then the signal was segmented at the certain length. After segmentation process, the next step was the peak normalization. The purpose of these proces is to to ensure the match volume and the optimal use of media distributed in the recording stage.

In this study, we used the peak normalization, which is slightly different with loudness normalization. The peak normalization is a process where the gain is changed to bring the highest value or peak of Pulse-Code Modulation (PCM) samples of analogue signal to a certain desired level. It is different with loudness normalization which adjusts a signal's gain so that the signal's loudness level equals some desired level. The peak normalization equation can be written:

$$X' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

with X'= output data of peak normalization, X_{max}= maximum value of the input data, X_i= input data that will be normalized, X_{min}= minimum value of the input data.

2.2. Feature Extraction

Feature extraction is a process that is done to find the specific characteristic of a sound signal by converting speech signal into set parameters called feature vectors. This process plays a very important role in the voice recognition process, or the key stage of an overall scheme for pattern recognition and classification [21]. It is a key stage because a better feature is good for the improving recognition rate.

We applied the WT to decompose the signal. Since the speech is a non-stationary signal, it is not suitable to be analyzed using the Fourier Transform (FT) because the FT only provides the frequency information of signal but it does not provide the information about what time which frequency is present. The WT is superior in describing the signal anomaly, pulses, and other events that occur in the short duration time in the signal, e.g. speech signal.

There are several types of WT methods, some of them are DWT and WPT. In the DWT decomposition process, only on the side of approximation is at a lower frequency, whereas WPT is a generalization of the DWT decomposition which gives a wide range of a signal analysis. WPT gives a balanced binary tree structure by decomposing both the lower (approximation) and higher frequency bands (detail) in order to provide more and better frequency resolution features about the speech signal analysis. The basic WT function can be written as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \qquad (2)$$

Where ψ(t) is known as wavelet or prototype function, parameter s and τ are called translation and scaling parameter respectively. The term 1/√s is used for energy normalization in the varying scale. In the wavelet research, the selection of the most suitable mother wavelet is still a relative question mark among researchers [13]. Figure 1 shows the structure of the WT.

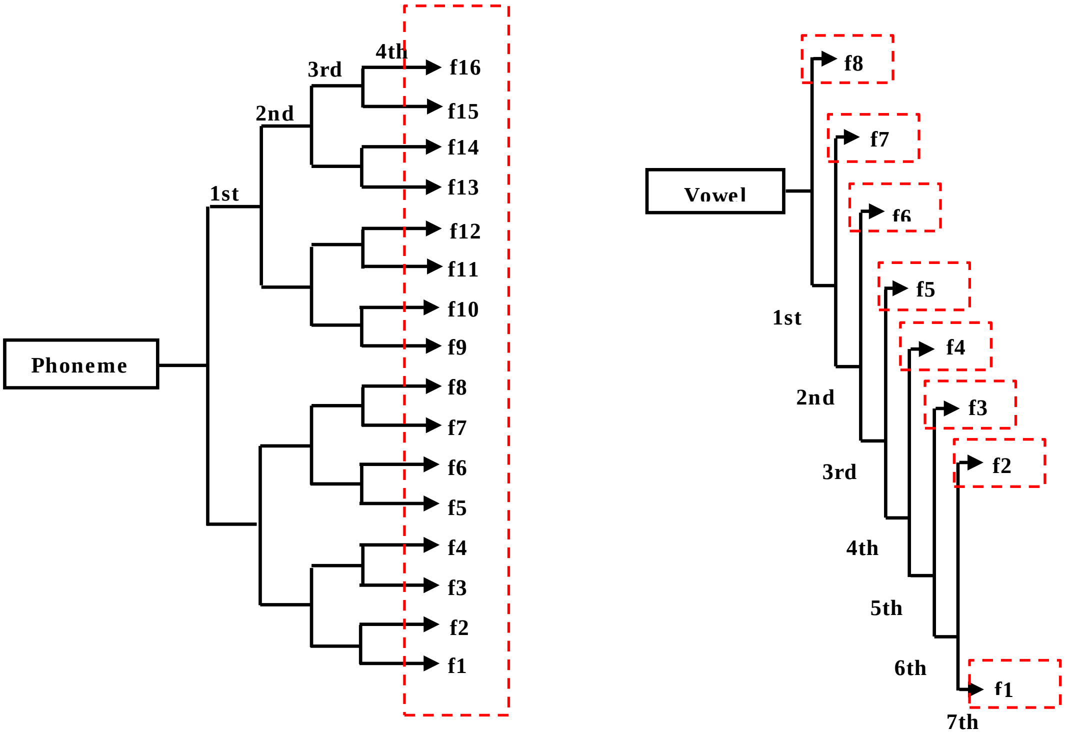

Figure 1. (a) The structure of Wavelet tree for the phoneme signal (b) The structure of Wavelet tree for the vowel signal

In feature extraction using WT, the process of choosing the right mother wavelet is crucial for optimal result of classification [13,22]. The mother wavelets filter used in this study are Haar, Daubechies, and Coiflet.

2.3. Selection of Features

Selection and testing process are conducted simultaneously on the features obtained to get the most reliable features to be used in the speech recognition process. This process is performed to minimize the Euclidean distance value between the matching test data and the features obtained of the respective syllables sound so we can be sure that the features can distinguish a certain syllables sound from the others accurately.The Euclidean distance (d) between two point p and q is given by:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (3)$$

After finding the candidates of the features, they are tested by calculating the Euclidean distance of the syllable features toward another phonemic feature. A feature which is categorized as effective and reliable, for example when a certain feature of the /ga/ syllableis tested with the /ga/ syllable or the syllable itself, it will have a very small ED value, however, when it is tested between the /ga/ syllable and the specific features with other syllables it will have a fairly large ED.

3. Results and Analysis

The first result is the lists of the features of each phoneme of the syllables in Indonesian which obtained by using three kinds of wavelet transform and using ED as its classifier. The second result is a recommendation of the best wavelet types to be used in the Indonesian syllables recognition system.

3.1. List of Features

The vowel and phoneme frequency component which were obtained by using mother wavelet of Haar are shown in Figure 2. Then, the results of vowel and phoneme were combined to estimate the frequency component of syllable.

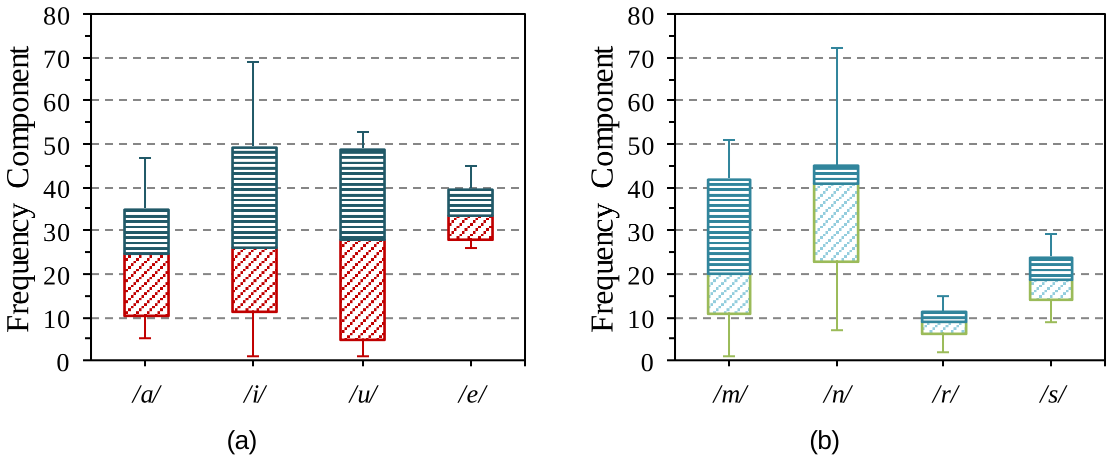

Figure 2. The boxplot of frequency component: (a) vowel, and (b) phoneme

Figure 2 shows the boxplot of frequency component distribution for each vowel and phoneme sound signal using the WT, which is also the exploratory frequency component data chart showing median, central spread of data and position of relative extremes [14]. The lower and higher whisker shows the lowest and highest frequency component to form a certain type of vowel or phoneme, whereas the lower, midle and upper box shows the first quartile, median, and the third quartile, respectively. From the box plot is understood that both vowel and phoneme have different range of frequency component. In Figure 2(a) shows that */i/* has the wider range of frequency component compared to the other vowels, whereas */e/* has the shortest frequency range. In Figure 2(b) shows that phoneme */n/* has wider range of frequency component, and */r/* has the shortest frequency component range.

Figure 3 shows the comparison of the DWT (which is highlighted with purple color) and the WPT (which is highlighted with yellow color) in the 2nd through 4th level of decomposition in term of effectiveness for the phoneme sound signal. From the graph it can clearly be seen at the DWT at 2nd level decomposition has the highest score in effectiveness especially for /m/ and /n/ compared to WPT as well as the other level of decomposition. In the overal observation, the DWT is more effective than the WPT as shown by effectiveness ratio is 9 versus 6.

The results of frequency component distribution of each syllable sound signal using the combination of phoneme and vowel frequency components are shown in Figure 4. Four different types of phoneme followed by four different type of vowel, so there are sixteen different types of CV syllables. From the boxplot, it shows that */n-i/* has the wider frequency component range than the other types of syllables.

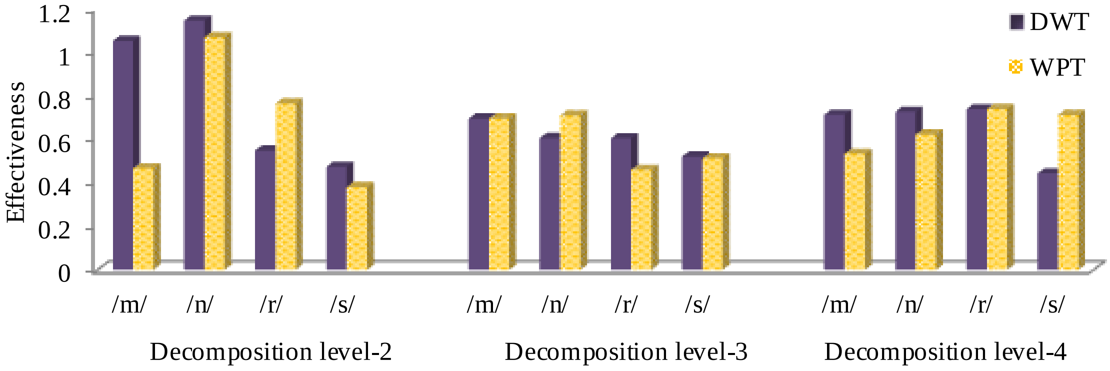

Figure 3. Effectiveness of DWT and WPT in varying decomposition level

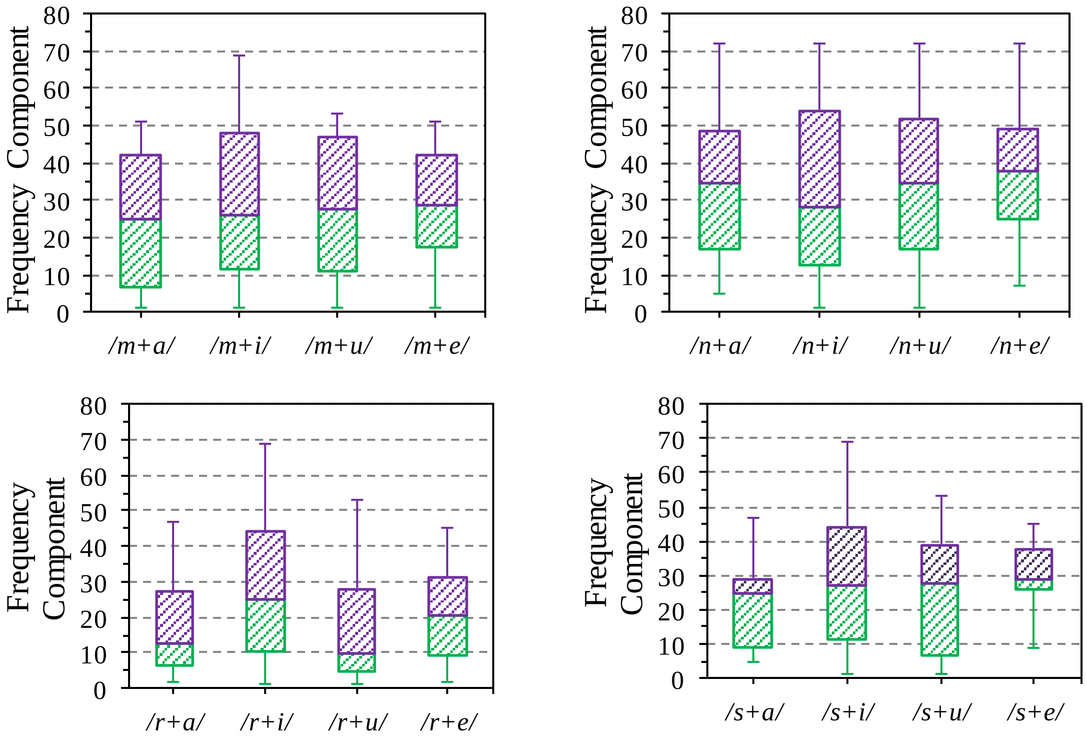

Figure 4. The boxplot of frequency component of each syllable

3.2. Effectiveness Analysis

Effectiveness analysis is conducted to determine the wavelet type that can distinguish a certain syllables sound most accurately and least likely to have errors in recognition due to the good Euclidean distance signature of its features. Effectiveness analysis is performed based on analysis on the Euclidean distance data in the Table 1.

Table 1. Euclidean Distance of the Haar Wavelet

				HAAR			
Speech Sound				FEATURE DATABASE			
Signal		*/ga/*	*/gi/*	*/gu/*	*/ge/*	*/gè/*	*/go/*
/ga/	ga1	1.739336	4.483127	4.487586	2.474394	4.966402	4.619309
	ga2	1.618870	4.339712	4.412261	1.973859	4.626011	4.610792
	ga3	2.011912	6.443900	7.515669	4.791306	7.322466	8.414237
	ga4	1.644084	3.680109	4.710546	1.961331	6.047641	4.624038
/gi/	gi1	3.332901	0.651473	3.411675	1.986983	5.369539	3.105632
	gi2	3.330979	0.451466	3.160718	3.295603	3.252592	3.085233
	gi3	3.117829	1.143073	3.464432	3.555397	3.649972	3.653752
	gi4	2.686213	1.215662	3.476852	4.113113	3.727356	3.621564
/gu/	gu1	2.624604	2.356103	0.769582	2.029690	1.424605	2.322751
	gu2	3.038636	2.927783	0.974952	1.850545	4.225113	2.489054
	gu3	2.331223	2.109750	0.962072	2.095373	3.559679	2.104405
	gu4	2.430047	2.086921	0.881081	2.258877	3.594850	2.159457
/ge/	ge1	2.401993	5.386929	6.739045	0.743418	6.675673	6.713523
	ge2	2.292781	4.199037	5.406717	1.384051	5.399906	5.126524
	ge3	3.549863	5.951152	6.229423	0.433025	6.786071	6.107202
	ge4	3.295424	3.174395	5.704801	0.726690	4.745619	5.510821
/gè/	gè1	3.263488	3.888218	1.655832	2.578474	0.640071	2.619778
	gè2	3.477828	4.114577	2.182904	2.408759	0.945802	2.802802
	gè3	3.197613	2.688976	1.025848	2.497159	0.915809	2.666389
	gè4	2.380542	3.206266	1.685697	2.541283	0.954769	3.294302
/go/	go1	2.386302	2.454211	1.565105	2.622511	1.028232	0.319745
	go2	2.471703	2.459079	1.499455	1.578169	4.410844	0.377439
	go3	3.086000	2.378117	2.039590	2.486216	4.727328	0.653640
	go4	2.636380	2.300141	1.586405	1.765964	4.743941	0.398730
	/a/	2.257869	4.3047	4.863113	1.739255	5.024338	5.471876
	/i/	2.758262	3.403421	2.881118	1.190823	2.474258	3.910252
Vowels	*/u/*	2.85382	4.242534	5.708679	1.840403	5.118138	6.07195
	/e/	2.752748	4.076228	4.865305	2.024756	4.834164	5.50375
	/o/	2.530442	3.935642	4.891535	2.283621	4.765649	4.674425

In Table 1, MEAN X is the average of the value of the data that is highlighted with a diagonal box, for example:

MEAN X = (1.739336 + 1.618870 + 2.011912 + 1.644084) / 4

while the variable ELSE is the average value of the other data, for example:

ELSE = ((MEAN */gi/*) + (MEAN */gu/*) + (MEAN */ge/*) + (MEAN */gè/*) + (MEAN */go/*) + (MEAN */Vowels/*) / 5

while DIFF can be expressed in Equation below:

$$\text{DIFF} = |\text{ MEAN X} - \text{ELSE }| \quad (4)$$

One of the syllables sound (*/ga/*) feature list obtained by using wavelet Daubechies 2 consists of the frequency components that have been tested and can accurately represent each syllables. Average value of the feature magnitude (MEAN X) decreased by the average value of the other types of syllables (ELSE), the decrement result is listed in the Table (DIFF) marked with yellow color for the wavelet type which has the best value (biggest value of Euclidean distance) in performance, as shown in Table 2, Table 3, and Table 4.

Table 2. Daubechies 2

DAUB 2	Syllables					
	/ga/	/gi/	/gu/	/ge/	/gè/	/go/
MEAN X	0.651765	0.44472	0.621208	1.123748	1.158	0.195441
ELSE	3.058578	2.300104	3.196969	2.840083	3.722528	4.754847
DIFF	**2.406814**	1.855384	2.57576	1.716335	2.564528	**4.559406**

Table 3. Haar

HAAR	Syllables					
	/ga/	/gi/	/gu/	/ge/	/gè/	/go/
MEAN X	1.75355	0.865419	0.896922	0.821796	0.864113	0.437389
ELSE	2.544893	3.781848	4.17603	2.149153	4.579792	4.697574
DIFF	0.791343	2.91643	3.279108	1.327357	**3.715679**	4.260185

Table 4. Coiflet 2

COIF	Syllables					
	/ga/	/gi/	/gu/	/ge/	/gè/	/go/
MEAN X	1.469861	0.698773	0.454107	0.536646	0.574475	0.403289
ELSE	3.478831	4.460107	3.830646	2.428409	1.88699	3.86079
DIFF	2.00897	**3.761334**	**3.376539**	**1.891763**	1.312515	3.457501

Table 5. Wavelet Effectiveness Ranking

RANKING	Syllables					
	/ga/	/gi/	/gu/	/ge/	/gè/	/go/
1	DAUB	COIF	COIF	COIF	HAAR	DAUB
2	COIF	HAAR	HAAR	DAUB	DAUB	HAAR
3	HAAR	DAUB	DAUB	COIF	COIF	COIF

3.3. Efficiency Analysis

In addition to an effective method, it needs an efficient method in the speech recognition. It is needed in order to find a quick and precise method when it is used in finding the specific features of the syllable. The efficiency of the method is determined by the features used in the method at each level of the decomposition in finding the specific features of the syllable. Table 6 shows the number of the features used by both feature extraction methods on every syllable and on every level of the decomposition.

Table 6. The Number of Features

SYLLABLES	Wavelet Type		
	Haar	Daub	Coif
/ga/	4	6	4
/gi/	8	15	11
/gu/	7	12	11
/ge/	7	8	7
/gè/	9	8	14
/go/	7	24	9

From the analysis of efficiency then compiled the ranking table of the most efficient types of wavelets to distinguish each of the syllables as shown in Table 7.

Table 7. Wavelet Efficiency Ranking

RANKING	Syllables					
	/ga/	/gi/	/gu/	/ge/	/gè/	/go/
1	HAAR	HAAR	HAAR	COIF	DAUB	HAAR
2	COIF	COIF	COIF	HAAR	HAAR	COIF
3	DAUB	DAUB	DAUB	DAUB	COIF	DAUB

3.4. Choosing the Best Wavelet

Cross ranking process or average ranking is then performed on Table 8 and Table 9 to find the best mother wavelet (the most effective and efficient) to be used to recognize the sound of each syllables. Cross ranking as shown in Table 8, is done by summing each ranking value in Table 5 and Table 7 for each type of wavelet of each syllables. The type of mother wavelet which has the smallest number is the best mother wavelet (the most effective and efficient) to recognize syllables sound. For example, Coif wavelet has the effectiveness rank of 1 and the efficiency rank is 2, then the cross ranking result is 2 + 1 = 3. If another type of wavelet has the value less than 3, that type of wavelet can be said better than the Coif wavelet.

Table 8. The Best Wavelet Ranking

RANKING	Syllables					
	/ga/	*/gi/*	*/gu/*	*/ge/*	*/gè/*	*/go/*
HAAR	4	3	4	2	3	3
COIFLET 2	4	3	3	4	6	5
DAUBECHIES 2	4	6	6	5	3	4

4. Conclusion

In this paper, the combined methods of Wavelet Transform (WT) and Euclidean Distance (ED) to estimate the expected value of the possibly feature vector of Indonesian syllable was proposed. Based on the experimental result presented in this paper, it can be concluded that the combined method of WT and ED are promising to be used for estimating the expected frequency component value of the possibly feature vector of Indonesian syllable. The effectiveness and efficiency ranking of three mother wavelet for the feature extraction of velar consonant */g/* with the following vowels are Haar, Coiflet 2, and Daubechies 2, respectively. The future work recommended for this research is to use bigger syllable dataset, applied to the Indonesian stop consonant or the other place of articulation (such as labial, dental, etc.), and to use the same level of decomposition for estimating the frequency component of vowel and phoneme.

References

[1] S Park, Y Kim, ET Matson, C Lee, W Park. *An Intuitive Interaction System for Fire Safety Using A Speech Recognition Technology.* The 6th International Conference on Automation, Robotics and Applications (ICARA). 2015: 388–392.

[2] P Youguo, S Huailin, L Tiancai. *The Frame of Cognitive Pattern Recognition.* 2007 Chinese Control Conference. 2006: 694–696.

[3] K Umapathy, S Krishnan, RK Rao. Audio Signal Feature Extraction and Classification Using Local Discriminant Bases. *IEEE Trans. Audio, Speech Lang. Process.* 2007; 15(4): 1236–1246.

[4] DL Fugal. *Conceptual Wavelets in Digital Signal Processing.* San Diego, California: Space & Signals Technical Publishing. 2009.

[5] SH Lee, JS Lim, JK Kim, J Yang, Y Lee. Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance. *Comput. Methods Programs Biomed.* 2014; 116(1): 10–25.

[6] STS Kumar. Blood Flow Analysis using Euclidean Distance Algorithm and Discrete Wavelet Transform. 2010: 330–335.

[7] DPP Mesquita, JPP Gomes, AHS Junior, JS Nobre. Euclidean distance estimation in incomplete datasets. *Neurocomputing.* 2017; 248: 11–18.

[8] A Boudjella, B Belhaouari, SH Bt, D Raja. License Plate Recognition Part II: Wavelet Transform and Euclidean Distance Method. 2012: 695–700.

[9] I Saulcy. Wavelet-based Euclidean Distance for Image Quality Assessment. 2010: 15–17.

[10] D Liu and DSZ Qiu. Wavelet Decomposition 4-Feature Parallel Fusion by Quaternion Euclidean Product Distance Matching Score for Palmprint Verification. 2008; 1: 2104–2107.

[11] R Hidayat, D Kristomo, I Togarma. *Feature extraction of the Indonesian phonemes using discrete wavelet and wavelet packet transform.* 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE). 2016: 478–483.

[12] R Hidayat, Priyatmadi, W Ikawijaya. *Wavelet based feature extraction for the vowel sound.* 2015 International Conference on Information Technology Systems and Innovation (ICITSI). 2015: 1–4.

[13] N Amalia, AE Fahrudi, AV Nasrulloh, N Amalia. Indonesian Vowel Recognition using Artificial Neural Network based on the Wavelet Features. *International Journal of Electrical and Computer Engineering (IJECE).* 2013; 3(2): 260–269.

[14] O Farooq and S Datta. Phoneme recognition using wavelet based features. *Elsevier Inf. Sci.* 2003; 150: 5–15.

[15] S Ranjan. *A Discrete Wavelet Transform Based Approach to Hindi Speech Recognition*. Signal Acquis. Process. 2010. ICSAP '10. Int. Conf. 2010.

[16] NS Nehe and RS Holambe. DWT and LPC based feature extraction methods for isolated word recognition. *EURASIP J. Audio, Speech, Music Process.* 2012; 2012(1): 7.

[17] RP Sharma, O Farooq, I Khan. Wavelet based sub-band parameters for classification of unaspirated Hindi stop consonants in initial position of CV syllables. *Int. J. Speech Technol.* 2013; 16(3): 323–332.

[18] S Nafisah, O Wahyunggoro, LE Nugroho. An Optimum Database for Isolated Word in Speech Recognition System. *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2016; 14(2): 588–597.

[19] P Král. *Discrete Wavelet Transform for automatic speaker recognition.* Image Signal Process. (CISP), 2010 3rd Int. Congr. 201; 7: 3514–3518.

[20] B Achmad, Faridah, L Fadillah. Lip Motion Pattern Recognition for Indonesian Syllable Pronunciation Utilizing Hidden Markov Model Method. *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2015; 13(1): 173–180.

[21] B Boashash, NA Khan, TB Jabeur. Time-frequency features for pattern recognition using high-resolution TFDs: A tutorial review. *Digit. Signal Process. A Rev. J.* 2015; 40(1): 1–30.

[22] J Saraswathy, M Hariharan, T Nadarajaw, W Khairunizam, S Yaacob. Optimal selection of mother wavelet for accurate infant cry classification. *Australas. Phys. Eng. Sci. Med.* 2014; 37(2): 439–456.