Jurnal Elektronika dan Telekomunikasi (JET), Vol. 17, No. 2, December 2017, pp. 42-47
Accredited by RISTEKDIKTI, Decree No: **32a/E/KPT/2017**
doi: 10.14203/jet.v17.42-47

Determination of the Optimum Wavelet Basis Function for Indonesian Vowel Voice Recognition

Syahroni Hidayat*, Habib Ratu P.N., Danang Tejo Kumoro

Program Studi Teknik Informatika
STMIK Bumigora Mataram
Jl. Ismail Marzuki No. 32, Cakranegara
Mataram, Indonesia

Abstract

Nowadays, wavelet has been widely applied in extracting features of the signal for automatic speech recognition system. Wavelets have many families that are determined by their mother function and order. The use of different wavelets to analyze the same signal would bring different results. In many cases, a trial and error procedure is used to select the optimal wavelet family. That is because there are no particular wavelet basis functions that can be applied to the entire speech signals. Therefore, it is necessary to analyze the similarity between the speech signal and the wavelet base function. One of the methods that can be used is cross-correlation. In this study, the degree of correlation is determined between wavelet base function and Indonesian vowels. The influence of gender and consistencies of the results are also used in the analysis. The results show that db45 and db44 are most similar to male and female vowels utterance, respectively. For consistencies, only vowel e gives a consistent result. Overall, db44 is most similar to all Indonesian vowels utterance.

Keywords: automatic speech recognition, cross-correlation, Indonesian vowels, wavelet, wavelet basis function determination.

I. INTRODUCTION

Automatic speech recognition (ASR) system aims to allow machine to recognize spoken language then converts the output into readable text in real time. Phoneme-based is widely used paradigm in ASR [1]. A phoneme is a minimum distinctive sound unit, which is identified by extracting a set of its features. In ASR system, a wavelet is generally used for speech signal features extraction. Its performance exceeds FFT on mapping signal into time-frequency domain simultaneously without signal loss [1] - [2]. One of the properties that makes the feature extraction process using wavelet better is its compatibility with voice signals [3].

Wavelet consists of several families where each of them has basis function called mother wavelet. In each family, a wavelet is determined by its order. The order is signed as its coefficient numbers that are developed from dilation and scaling function applied to mother wavelet. These variations make the use of wavelet still like trial and error. That is because there are no particular wavelet basis functions that can be applied to the entire signal [4].

The Daubechies wavelet has been applied in [5] - [6] for phoneme and syllable recognition. It was chosen because its form is non-stationary and compatible with voice signal [3]. The result shows that there are several phonemes and syllables either matched or not with Daubechies wavelet coefficient. The evaluation of the optimal wavelet feature extraction for Indonesian vowels is presented in [7]. To choose the optimal wavelet function, this research only used one of many orders available for each wavelet, which means that there is no comparison between wavelet in the same family. In this case, it is necessary to know the optimal wavelet basis function for speech feature extraction.

Several algorithms have been developed to determine the optimal basis wavelet function, such as entropy value in exact decomposition level [8], the average value of the optimum basis function of wavelet packet cosine-transform (WPCT) [9] and the statistical correlation between all wavelet basis function and biomedical decomposition signal [10] - [11]. In the research mentioned last, the use of statistical correlation gives limitation where the length of the signal and wavelet basis function must be equal.

In this paper, the determination of the optimal wavelet basis function applied for speech signal recognition is explored, especially for Indonesian vowels. Cross-correlation algorithm is used. Compared to statistical correlation process, this method does not require the similarity of length between the two data to be processed. The influence of gender and the consistencies of its result are counted in. This paper is organized as follows: Section 1 provides the introduction to this research. Section 2 describes research methodology to determine the optimum wavelet basis function for Indonesian vowels recognition. In Section 3, the design implementation and analysis of the result is presented, followed by the conclusion in Section 4.

* Corresponding Author.
Email: syahroni.hidayat@stmikbumigora.ac.id
Received: October 31, 2017 ; Revised: December 18, 2017
Accepted: December 19, 2017 ; Published: December 30, 2017

II. Methodology

The optimal basis wavelet determination system has been developed using Matlab, with the 2.3Hz Intel-core i3 processor. Overall, there are three stages process that consist of pre-processing, processing and post-processing. In pre-processing, data were recorded, followed by de-noising, endpoint detection and normalization process. The cross-correlation process between signal and wavelet basis function is then employed at processing stage. At post-processing stage, the optimal wavelet basis function is determined for Indonesian vowels recognition. Details of each process are described as follows.

A. Data Selection and Recording

A syllable is the smallest part of the word that cannot be dispersed. In Indonesian language, the syllable always has vowels as the core of a word. Table 1 shows the set of Indonesian syllable pattern [12]. A syllable may be composed of only a single vowel. In this research, the used data are Indonesian vowels a, i, u, e, è, o, and ò.

The data were recorded from 50 adults with 25 of each male and female speakers using recording software Audacity and taken using a headset. The recording process using sampling frequency 16000 Hz, mono, PCM 16 bit. The utterance standard is following International Phonetic Association (IPA) sample utterances [13]. It is done in a closed room to reduce the background noise effect. The properties and variable of the recording are described in Table 2.

Table 1
The frequency of syllable patterns calculated from 36,395 native Indonesian basic words (roots)

No	Syllable Pattern	Frequency
1.	V	3.379
2.	VK	4.064
3.	KV	62.125
4.	KVK	47.925
5.	VKK	4
6.	KVKK	99
7.	KVKKK	5
8.	KKV	879
9.	KKVK	572
10.	KKVKK	13
11.	KKKV	54
12.	KKKVK	44
Total syllables		119.163
No consecutive consonants (pattern 1 to 4) = 117493/119163		98.60 %
With consecutive consonants (pattern 5 to 12) = 1670/119163		1.40 %

Source: Suyanto & Hartati [12]

Table 2
Variable and properties of recording

No	Variable	Properties
1.	Data recorded	Vowels
2.	Sample numbers	50
3.	Repetition	1
4.	Tempo	IPA standard
5.	Sampling Frequency	16000 Hz
6.	Duration/Utterance	1 second
7.	Recording Environment	Closed room
8.	Data format	*.wav

B. Signal de-Noising and End Point Detection

The recorded speech signal usually consists of three elements, i.e. silence, unvoiced and voiced. Silence in speech signal can be categorized as noise. The noise influence neither computation process nor the quality of the signal. Therefore, the noise must be reduced to produce best recording quality and to lighten the computational process. The voiced signal is extracted from noise by removing the silence using endpoint detection algorithm [14] - [16]. The endpoint detection applied Short-Time Energy (STE) and Short-Time Zero Crossing Rate (ZCR) [17]. STE and ZCR formula is described in equation (1) and (2) below, respectively.

$$E_n = \sum_{m=-\infty}^{\infty} x(m)W(n-m) \tag{1}$$

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|W(n-m) \tag{2}$$

These two formulas use windows function W in their process, which applied to framed signal $x(m)$. There are several windows functions can be used, such as rectangle window, Hamming window and Hanning window [18]. Hamming window is applied in this research because it has a small amount of discontinuity and widely used in speech processing. Besides, Hamming window attenuation for all frequencies is approximately constant[18] - [20]. ZCR applied sgn function.

C. Normalization

In the recording process, the recorded voice maximum amplitude fluctuates from each other. This is influenced by the variation of the distance between microphone and signal source. Amplitude normalization process is applied to the signal to make its maximum magnitude homogenous. The normalization process is described in equation (3) below.

$$S_{no}(n) = \frac{S(n)}{max\left(abs\left(S(n)\right)\right)} \tag{3}$$

where S_{no} and $S(n)$ are a normalized signal and a raw signal respectively. This process divides the signal by its absolute maximum value.

D. Cross-Correlation

Correlation is the statistical measurement to determine the degree of linear similarities between two variables, which is described in terms of coefficient between -1 and +1. The closer the coefficient value to either -1 or +1, the correlation between the two variables is powerful [9]. The correlation formula is described in equation (4) and (5) as follows.

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E\left((x-\mu_x)(y-\mu_y)\right)}{\sigma_x \sigma_y} \tag{4}$$

where $\rho_{x,y}$ defined as correlation coefficient between sequence x and y, μ_x and μ_y are expectation value, and σ_x, σ_y are standard deviation. E is expected value and *cov* defined as covariance. $E(x) = \mu_x$, $E(y) = \mu_y$ and $\sigma^2(x)= E(x^2) - E^2(x)$, $\sigma^2(y)= E(y^2) - E^2(y)$, $\rho_{x,y}$ described as [10], [19]:

$$\rho_{x,y} = \frac{E(xy) - E(x)E(y)}{\sqrt{E(x^2) - E^2(x)}\sqrt{E(y^2) - E^2(y)}} \tag{5}$$

Equation (5) requires equal data length between two variables which will be measured. So, for the speech signal that has a variation of length, there is cross-correlation measurement algorithm. Given two real-valued sequences $x(n)$ and $y(n)$ of finite energy, the cross-correlation of $x(n)$ and $y(n)$ is a sequence of $\rho_{x,y}(l)$ that is defined as [18] - [19]:

$$\rho_{x,y(l)} = \sum_{n=-\infty}^{\infty} x(n)y(n-l) \tag{6}$$

The index *l*in equation (6) is called the shift or lag parameter.

E. Wavelet Transform

Wavelet is defined as the limited duration signal that has zero average value. Not like a sinusoid signal that has the length from -∞ until +∞, theoretically, wavelet has the beginning and the ending. Wavelet concentrates its energy into time and frequency domain simultaneously. This property makes wavelet as an appropriate choice to analyze an instant signal. This was the result of dilation and scaling of limited energy called mother wavelet that acts as high pass filter, where scaling function acts as low pass filter [21]. There are several wavelet families, such as Haar, Daubechies (db), Symlet, Coiflet, Gaussian, Morlet, complex Morlet, Mexican hat, bio-orthogonal, reverse bio-orthogonal, Meyer, a discrete approximation of Meyer, complex Gaussian, Shannon, and frequency B-spline [10], [22].

There are two transformation methods in wavelet, i.e. Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT). On the application, DWT is more efficient in computation, but CWT is optimum and more efficient because of its capabilities to maintain the information without downsampling the process. The CWT formula is described in equation (7) below.

$$\begin{aligned} CWT(t,\omega) &= \left(\frac{\omega}{\omega_0}\right)^{\frac{1}{2}} \int_{-\infty}^{+\infty} s(t')\,\psi \\ &\quad * \left(\frac{\omega}{\omega_0}\right)(t'-t)dt' \\ &= \{s(t), \psi(t)\} \end{aligned} \tag{7}$$

where {} is inner product of signal $\psi \in L^2(\Re)\backslash\{0\}$ that is defined as mother wavelet. A mother wavelet must fulfill this condition:

$$0 \prec c_\psi = 2\pi \int_{-\infty}^{+\infty} \left|\hat{\psi}(\xi)\right|^2 \frac{d\xi}{|\xi|} \prec +\infty \tag{8}$$

and ω/ω_0 is scaling factor [10].

1) Daubechies Wavelet

Daubechies wavelet has been widely applied in signal processing, particularly in speech processing. Its filter coefficient denotes as dbN, which N denotes the order of the filters. As far as the wavelet Daubechies development, its highest coefficient order is 45 [22]. The use of Daubechies in speech processing is still trial and error. The orthonormal basis function for $L^2(\Re)$ of Daubechies functions, for any integer r are determined using this formula [10], [23]:

$$\phi_{r,j,k} = 2^{j/2}\phi_r(2^j - k) \quad j,k \in Z \tag{9}$$

where the function $\varphi_r(x)$ in $L^2(R)$ has the property that $\{\varphi_r\ (x - k)|k \in Z\}$ is an orthonormal sequence in $L^2(R)$. *j*, *k* and *r* are the scaling index, the shifting index and the filter index respectively [10].

In a function $f \in L^2(R)$, f_j at 2^{-j} scale is defined by:

$$f_j(x) = \sum_k \langle f, \phi_{r,j,k}\rangle\, \phi_{r,j,k}(x) \tag{10}$$

The details are defined as [10]:

$$d_j(x) = f_{j+1}(x) - f_j(x) \tag{11}$$

Daubechies' orthonormal basis has the following properties:

- ω_r has the compact support interval $[0, 2r + 1]$
- ω_r has about $r/5$ continuous derivatives
- limited duration signal with an average value of zero.

2) Wavelet determination

The determination of the optimal basis wavelet function process in this research runs following the flowchart shown in Figure 1. This algorithm is developed and modified from [10] that determines the optimal wavelet basis function for the biomedical signal. The modification applied to the length of the segment and the correlation determination method. First is the enhancement of raw speech signal. There are three processes on it, i.e. de-noising, endpoint detection and normalization. The total length of each data is treated alike with 8196 lengths. After these process, the signal is segmented by 2^n length, start from 256 until 4096 data length.

Each segmented signal is then decomposed until the 4th level of decomposition. All scales in this decomposition level then processed using Continuous Wavelet Transform (CWT) to gain its coefficient. The gained coefficient of each scale is then correlated with wavelet basis function (DWBF) using cross-correlation algorithm. Then, the average of the absolute value of cross-correlation coefficient for a segment (RAK) is determined. This process repeated until all segmented raw signals have been evaluated with all of the wavelet basis function and all of the segmentation length accomplished.

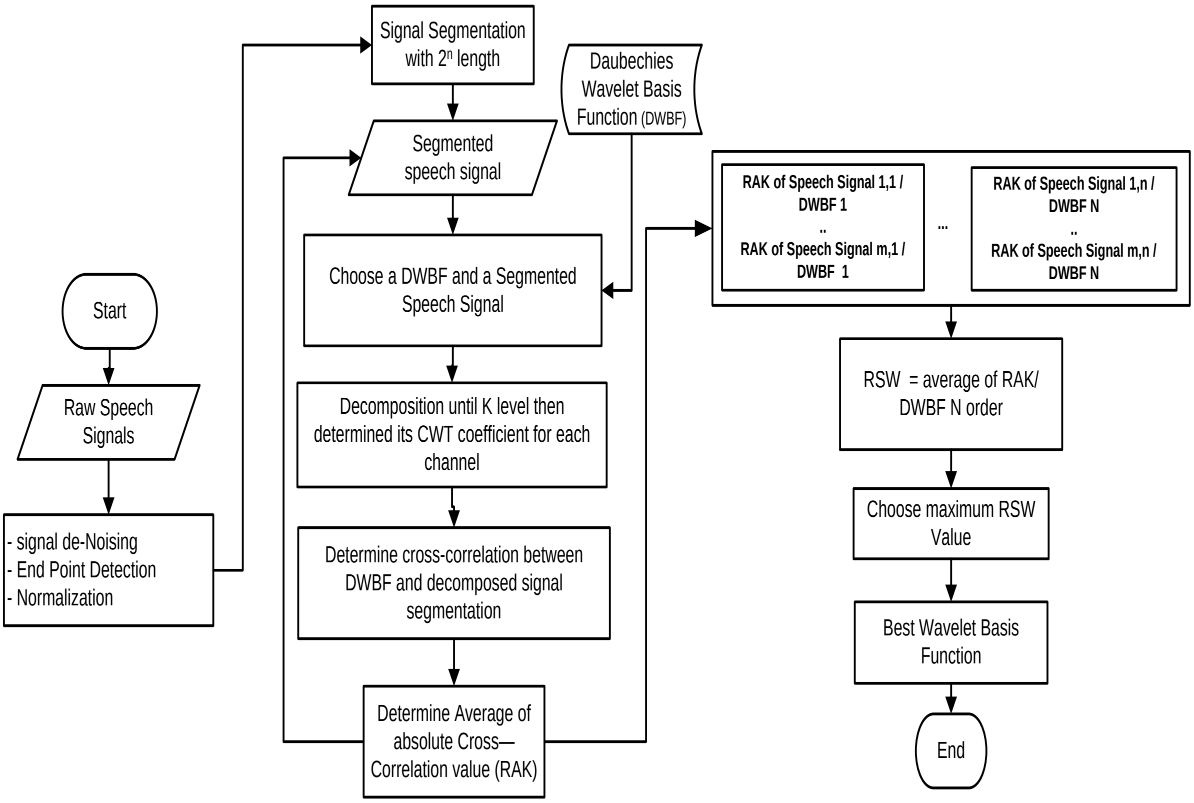
Start
Raw Speech Signals
- signal de-Noising
- End Point Detection
- Normalization
Signal Segmentation with 2^n length
Segmented speech signal
Daubechies Wavelet Basis Function (DWBF)
Choose a DWBF and a Segmented Speech Signal
Decomposition until K level then determined its CWT coefficient for each channel
Determine cross-correlation between DWBF and decomposed signal segmentation
Determine Average of absolute Cross—Correlation value (RAK)
RAK of Speech Signal 1,1 / DWBF 1
..
RAK of Speech Signal m,1 / DWBF 1
...
RAK of Speech Signal 1,n / DWBF N
..
RAK of Speech Signal m,n / DWBF N
RSW = average of RAK/ DWBF N order
Choose maximum RSW Value
Best Wavelet Basis Function
End

Figure 1. The optimal basis wavelet determination flowchart.

Next is the determination of average value between RAK and DWBF, called RSW. The result of this process is temporarily stored. It is then evaluated to choose the optimal basis wavelet function for Indonesian vowels. The evaluation criterion is using the maximum value of RSW. The result is then divided and analyzed based on all Indonesian vowels, gender influence and the consistencies.

III. Results and Discussion

The pre-processing result of the speech signal is shown in Figure 2. The raw signal that has been enhanced had the noise reduced and its maximum magnitude is equal to -1 or +1. The computation time for all Indonesian vowel signals is influenced by the segmentation length. It increased as the segmentation length decreased. The increase in computation time is approximately twice for each segment. The result of time consumption needed is shown in Table 3.

After applying the algorithm, the results show that among Daubechies family, db44 is the most similar function for all Indonesian vowels. The reason is because the Daubechies's function is an orthonormal function. For any integer r, the orthonormal basis function is gained from equation (9). Daubechies wavelets provide convenient results in signal processing techniques due to the characteristics mentioned above.

The influence of gender on Indonesian vowels signal is also shown in Table 4. It is same as all Indonesian vowels, db44 is most suitable wavelet basis function too for the female speaker. As for the male speaker, the most suitable wavelet basis function is db45. It can be concluded that in this research, the female voice is more dominant than the male voice.

The analysis result of the algorithm are shown in Table 5 - Table 7. There are inconsistencies in determining the optimum wavelet basis function for each Indonesian vowels. Only vowel *e* gives a consistent result on every segmentation length, where db44 is more suitable for vowel *e*. And according to Table 4, the inconsistencies start at 2048 segment length for male and female speakers.

Table 3
Computation time

No	Length of Segment	Time (s)
1	4096	98
2	2048	179
3	1024	373
4	512	785
5	256	1678

Table 4
The optimum basis function for Indonesian vowels

Speaker	Length of Segment					
	4096	2048	1024	512	256	Average
M + F	db44	db45	db44	db45	db45	db44
Male	db44	db44	db45	db45	db45	db45
Female	db44	db45	db44	db44	db44	db44

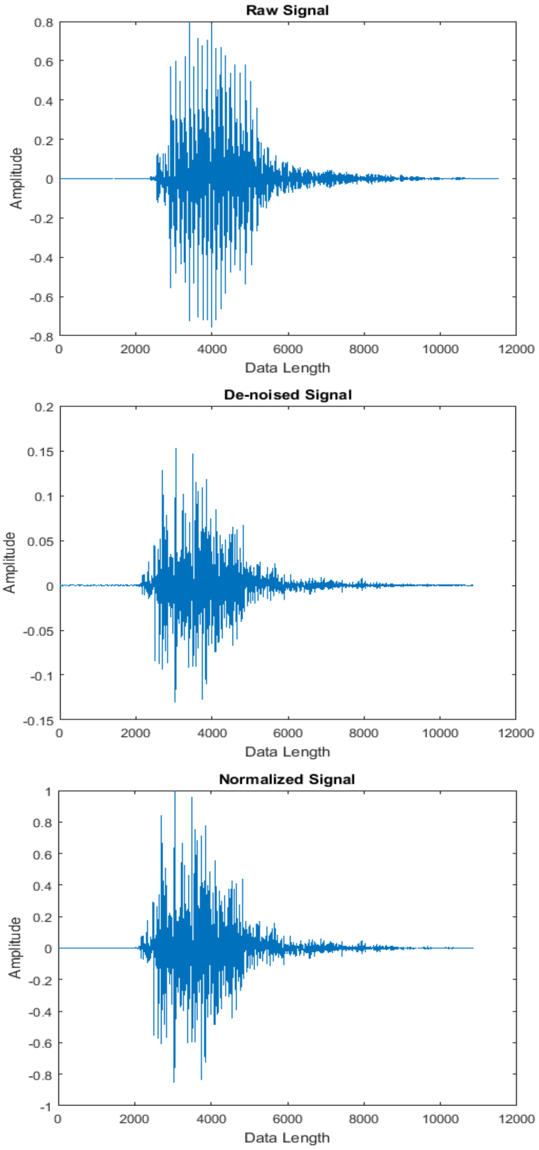

Figure 2. Pre-processing result of vowel '*a*'.

TABLE 5
THE OPTIMUM BASIS FUNCTION FOR ALL SPEAKERS

Vowels	Length of Segment				
	4096	2048	1024	512	256
a	db44	db44	db44	db45	db44
i	db44	db45	db45	db45	db45
u	db44	db45	db45	db45	db45
e**	db44	db44	db44	db44	db44
è	db44	db44	db44	db45	db45
o	db44	db44	db44	db45	db44
ò	db45	db45	db44	db44	db44

**The optimum wavelet basis function is db44

TABLE 6
THE OPTIMUM BASIS FUNCTION FOR MALE SPEAKERS

Vowels	Length of Segment				
	4096	2048	1024	512	256
a	db42	db44	db44	db45	db44
i	db44	db44	db45	db45	db45
u	db45	db45	db45	db45	db44
e	db45	db44	db45	db45	db45
è*	db45	db45	db45	db45	db45
o*	db45	db45	db45	db45	db45
ò	db44	db44	db44	db44	db42

* The optimum wavelet basis function is db45

TABLE 7
THE OPTIMUM BASIS FUNCTION FOR FEMALE SPEAKERS

Vowels	Length of Segment				
	4096	2048	1024	512	256
a	db44	db44	db44	db45	db44
i*	db45	db45	db45	db45	db45
u	db1	db44	db45	db44	db45
e**	db44	db44	db44	db44	db44
è	db44	db44	db44	db44	db45
o**	db44	db44	db44	db44	db44
ò	db45	db45	db45	db44	db44

* The optimum wavelet basis function is db45
**The optimum wavelet basis function is db44

The analysis result of the algorithm are shown in Table 5 - Table 7. There are inconsistencies in determining the optimum wavelet basis function for each Indonesian vowels. Only vowel *e* gives a consistent result on every segmentation length, where db44 is more suitable for vowel *e*. And according to Table 4, the inconsistencies start at 2048 segment length for male and female speakers.

CONCLUSION

In this research, an algorithm to determine the optimal wavelet basis function for Indonesian vowels recognition using cross-correlation has been developed. It can be concluded that female voices are more dominant than male voices. For female voices, the optimum wavelet basis function is db44 for all segmentations. As for male speakers, db45 is the optimum value. Vowel "e" gives a consistent result on every segmentation length with db44 as a suitable basis function.

Future work for this research will include an analysis of the recognition accuracy of an ASR using db45 and db44 as the feature extraction tools.

ACKNOWLEDGEMENT

We would say many thanks to DRPM of Indonesian Ministry of Research, Technology and Higher Education (RISTEKDIKTI) for funding this research and STMIK Bumigora Mataram for supporting this research.

REFERENCE

[1] M. A. Anusuya and S. K. Katti, "Front end analysis of speech recognition: a review," *Int. J. Speech Technol.*, vol. 14, no. 2, pp. 99–145, 2011.

[2] C. R. Rashmi, "Review of algorithms and applications in speech recognition system," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5258–5262, 2014.

[3] N. M. Mahesh S. Chavan, "Studies on implementation of Harr and Daubechies Wavelet for denoising of Speech Signal," *Int. J. Circuits, Syst. Signal Process.*, vol. 4, no. 3, pp. 83–96, 2010.

[4] N. Ahuja, S. Lertrattanapanich, and N. K. Bose, "Properties determining choice of mother wavelet," in *IEE Proc.-Vis. Image Signal Process.*, vol. 152, no. 5, 2005.

[5] O. Farooq and S. Datta, "Phoneme recognition using wavelet based features," in *Information Sciences*, vol. 150, no. 1–2, pp. 5–15, 2003.

[6] S. Hidayat, R. Hidayat, and T. B. Adji, “Speech recognition of KV-patterned Indonesian syllable using MFCC, wavelet and HMM,” *J. Ilm. Kursor*, vol. 8, no. 2, pp. 67–78, 2015.

[7] R. Hidayat, Priyatmadi, and W. Ikawijaya, “Wavelet based feature extraction for the vowel sound,” in *Proc. of The 2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2015, pp. 1–4.

[8] R. R. Coifman and M. V. Wickerhauser, “Entropy-based algorithms for best basis selection,” *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[9] J. Galka and M. Ziolko, “Mean best basis algorithm for wavelet speech parameterization,” in *Proc. of Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009, pp. 1110–1113.

[10] J. Rafiee, M. A. Rafiee, N. Prause, and M. P. Schoen, “Wavelet basis functions in biomedical signal processing,” *Expert Syst. Appl.*, vol. 38, no. 5, pp. 6190–6201, 2011.

[11] J. Rafiee and P. W. Tse, “Use of autocorrelation of wavelet coefficients for fault diagnosis,” *Mech. Syst. Signal Process.*, vol. 23, pp. 1554–1572, 2009.

[12] Suyanto and S. Hartati, “Design of Indonesian LVCSR using combined phoneme and syllable models,” in *Proc. of The 7th International Conference on Information & Communication Technology and Systems (ICTS)*, 2013, pp. 191–196.

[13] International Phonetic Association, “IPA Learning Tool.” [Online]. Available: http://www.internationalphoneticalphabet.org/ipa-learning-tool/IPA-Interface/ipa-chart-with-sounds2.html. [Accessed: 01-Jan-2015].

[14] S. Hidayat, U. Hasanah, and A. A. Rizal, “Algoritma penghapus derau / silence dan penentuan endpoint dengan nilai ambang terbobot untuk sinyal suara,” in *Proc. of Seminar Nasional APTIKOM (SEMNASTIKOM)*, 2016, no. October 2016, pp. 320–323.

[15] Abriyono and A. Harjoko, “Pengenalan Ucapan Suku Kata Bahasa Lisan Menggunakan Ciri LPC, MFCC, dan JST,” *IJCCS*, vol. 6, no. 2, pp. 23–34, 2012.

[16] R. Asliyan, “Syllable Based Speech Recognition,” in *Speech Technology*, I. Ipsic, Ed. InTech, 2011, pp. 263–284.

[17] S. Poornima, “Basic Characteristics of Speech Signal Analysis,” *Int. J. Innov. Res. Dev.*, vol. 5, no. 4, pp. 1–5, 2016.

[18] V. K. Ingle and J. G. Proakis, *Digital Signal Processing using MATLAB,* 3rd ed. USA: CENGAGE Learning, 2012.

[19] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[20] I. McLoughlin, *Applied Speech,and Audio Processing : With Matlab Examples*, 1st ed. United Kingdom: Cambridge University Press, 2009.

[21] O. Rioul and M. Vetterli, “Wavelets and Signal Processing,” *IEEE SP Magazine*, Oct-1991.

[22] M. Misiti, Y. Misiti, G. Oppenheim, and J. Poggi, *Wavelet Toolbox ^{TM} 4 User ' s Guide*. 2009.

[23] I. Daubechies, “Orthonormal bases of compactly supported wavelets,” *Commun. Pure Appl. Math.*, vol. 41, no. 7, pp. 909–996, 1988.