❒   260

Indonesian Vowel Recognition using Artificial Neural Network based on the Wavelet Features

Nadya Amalia, Arfan E. Fahrudi, Amar V. Nasrulloh
Physics Study Program, Universitas Lambung Mangkurat, Banjarbaru, Indonesia

Article Info

ABSTRACT

There are six vowels in Indonesian language, i.e. /a/, /i/, /u/, /e/, /ə/ and /o/. This paper presents Indonesian vowel recognition using artificial neural network (ANN) based on the wavelet features. The wavelet features were the wavelet coefficients of vowel signal which were extracted by using discrete wavelet transform (DWT). Vowel samples were recorded from native Indonesian speakers, 10 males and 10 females. Db4 and sym4 were used as the mother wavelet, and decomposition level 2, 4 and 6 were implemented for each vowel sample. Minimum, maximum, mean and standard deviation value of the wavelet coefficients then were used as input vectors of ANN with 2 hidden layers. Backpropagation algorithm was used to training the ANN. From the experimental results, an overall recognition rate of 70.83% could be achieved. In case of male speakers the highest recognition rate is 90% and in case of female speakers the highest recognition rate is 80%.

Corresponding Author:

Nadya Amalia,
Physics Study Program, UniversitasLambungMangkurat,
Banjarbaru, Indonesia
Email: aydan.amalia@gmail.com

1. INTRODUCTION

Indonesia has a broad linguistic diversity. There are 726 languages in the country; making it the world's second most diverse, after Papua New Guinea which has 823 local languages [1], [2]. Therefore, Indonesia's national language, Indonesian language, policy has been called a "miraculous success" [3]. It has been effective in uniting the nation, creating a strong national identity and promoting education and literacy throughout the nation [3], [4]. However, there is a difference in pronunciation of a word uttered by people of different mother tongues which is mainly due to the variations in vowels and the way they are pronounced [5], [6]. According to Rogers [7] and Stevens [8], vowels are voiced sounds produced by passing air through the mouth

Without any major obstruction in the vocal tract. There are six vowels (excluding the diphthongs) in Indonesian language [9]. They include /a/ (like "a" in "father"), /i/ (like "ee" in "knee"), /u/ (like "oo" in "moon"), /e/ (like "e" in "bed"), /ə/ (like "e" in "lantern"), and /o/ (like "o" in "boss") [10]. The vowel inventory is relatively normal as shown in Table 1.

Table 1. Indonesian vowels [11]

	Front	Unrounded Central	Rounded Back	
High		i		u
Mid		e	ə	o
Low			a	

Vowel recognition has been increasingly developedfor many languages withseveraldifferent methods. Sadeghi and Yaghmaie [12] introduced Farsi vowel recognition based on visual features of lips. Prica and Ili´c [13] presented Serbian vowel recognition by using formants. In this paper, Indonesian vowel recognition using artificial neural network (ANN) based on wavelet features ofvowel is presented.

Each vowel has characteristic features and the process to obtain those features is called feature extraction. Feature extraction is an important step in recognition system because the recognition rate depends on the feature extraction results [14]. The wavelet features in this paper are extracted from vowel signal via discrete wavelet transform (DWT). DWT is the most promising mathematical transformation which provides both the time-frequency information of the signal and is computed by successive low-pass filtering and high-pass filtering to construct a multi resolution time-frequency plane [15]. Therefore, wavelet coefficients can accurately represent the signal components [16]. Artificial neural network (ANN) with multilayer perceptron architecture is usedas the recognition system. Compared to conventional programming, ANN has the capability of solving problems that do not have algorithmic solution and is therefore found suitable to tackle problems that people are good to solve such as this vowel recognition.

2.   MATERIALS AND METHODS

2.1. Discrete wavelet Transform

Vowel signal is essentially a non-stationary signal. The analysis of a non-stationary signal by using Fourier transform (FT) or short time Fourier transform (STFT) does not give satisfactory results. Wavelet analysis is able to reveal signal aspects that other analysis techniques miss, such as trends, breakdown points, discontinuities, etc. It performs a multiresolution analysis which makes it possible to analyze a signal at different frequencies with different resolutions [17], [18].

Discrete wavelet transform (DWT) uses filter banks for the construction of the multiresolution time-frequency plane [17]. A filter bank consists of wavelet filters which separate a signal into frequency bands [17], [19]. In the two channel filter bank as shown in Figure 1, the low-pass filter L(z) and high pass filter H(z) split the signal into two frequency bands.
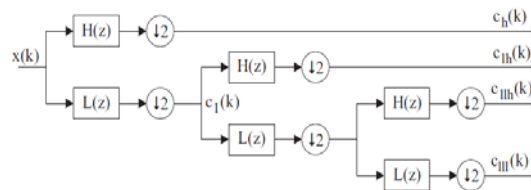

Figure 1. Two channel filter bank with 3 level decomposition [17]

The different output signals of the analysis filter bank are called subbands and the filter bank technique is also called subband decomposition. Decomposition can be expanded to an arbitrary level, depending on the desired resolution [17], [20]. Figure 1 visualized decomposition level 3. The subband with wavelet coefficients c_{III} is called the approximation coefficient cA and contains the lowest frequencies. The other subbands are calleddetails coefficientcD and give the detail information of the signal [17]. For decomposition level p, cA and cD can be obtained by [21]:

$$cA[p] = \sum_n x[k]L[2p - k] \tag{2}$$

$$cD[p] = \sum_n x[k]H[2p - k] \tag{3}$$

2.2. Artificial Neural Network

Artificial neural network (ANN) simulates the neural information processing of human brain [22], [23]. ANN processes inputs in parallel with a large number of processing elements called neurons and uses large interconnected networks of simple and non linear units [24]. Therefore, the ANN function is determined largely by the connections between elements. ANN with multilayer perceptron architecture, as shown in Figure 2, was used in this work. A signal is transmitted in one direction from the inputs to the outputs and therefore this architecture is also called feed forward [25].
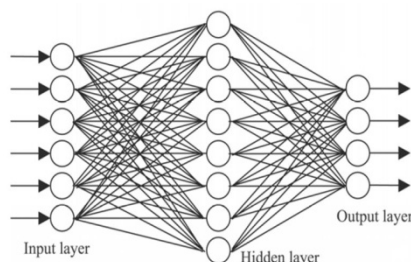
Figure 2. ANN with multilayer perceptron architecture

Each neuron may use any differentiable transfer function f to generate its output. Equation (4) and Equation (5) show the bipolar sigmoid transfer function and purelin transfer function which were used in this work.

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}} \tag{4}$$

$$f(x) = x \tag{5}$$

If the last layer of a multilayer network has sigmoid neurons, then the outputs of the network are limited to a small range –1 to +1. If linear output neurons are used the network outputs can take on any value [26], [27].

2.3. Recording

Vowel samples in this work were recorded from 20 native Indonesian speakers, aged between 19 – 25 years old, 10 males and 10 females. This work restricted only to the six Indonesian monofthong vowels, /a/, /i/, /u/, /ə/, /e/ and /o/, leaving out of consideration the category of diphthongs. Each vowel sound was recorded into 3 seconds *.wav file using a laptop and a microphone. The sampling rate used in recording was 44100 Hz and 16-bit resolution mono. Each vowel was recorded from each speaker for 2 times. The database then was partitioned into two subsets, one for training phase, 75% of the total, and the other for testingphase, 25% of the total.

2.4. Pre-processing

1. Sampling

The effective frequency of human voice is 4000 Hz. Considering Shannon theory [28], sampling frequency of 44100 Hz was used in this work.

2. Pre-emphasize filter

Pre-emphasize filter is used to spectrally flatten the vowel signal. It was implemented by the Equation (6) with a pre-emphasized constant value, a, of 0.96 [29], [30]:

$$x'[n] = x[n] - ax'[n-1] \tag{6}$$

3. Frame Blocking dan Windowing

After being filtered, the signal then was divided into several frames of 20 ms, so one could assume that the vowel signal was stationary within each frame.

$$(\text{Sampling Rate})(\text{Frame Length}) = \text{Number of Samples in a Frame} \tag{7}$$

$$(44100 \text{ samples/s})(0.02 \text{ s}) = 882 \text{ samples} \tag{8}$$

By choosing frames of 20 ms, each frame in this work consists of 882 samples as shown in Equation (7) and (8).

After frame blocking, the next step was windowed each individual frame to minimize the signal discontinuities at the beginning and end of each frame. Hamming window was used in this work, it could be calculated as:

$$W(n) \; = \; 0.54 - 0.46 \cos(2\pi n / 881), 0 \le n \le 881 \tag{9}$$

    4. DWT decomposition

Selection of the suitable mother wavelet and the number of decomposition level play an important role in obtaining good recognition rate in speech recognition. Among the various wavelet bases, db4 and sym4 were used in this work because of their orthogonality property and efficient filter implementation [31]. Decomposition level 2, 4 and 6 were implemented for each vowel.
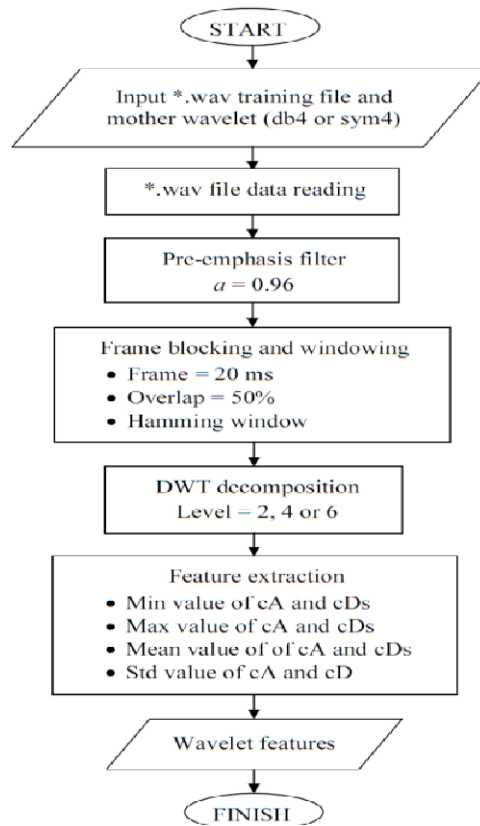

Figure 3. Pre-processing flowchart

2.5. Processing

    1. Building ANN architecture

This work used ANN with multilayer perceptron architecture. The architecture consisted of two hidden layerswith activation function of bipolar sigmoid and output layer with activation function of purelin. The number of neurons in the first hidden layer is 10 neurons and vary in the second hidden layer, 5 and 7 neurons.

    2. Training

To perform a particular function, ANN can be trained by adjusting the values of the connections (weights) between elements. The training process requires a set of examples, training set, of proper network behavior-network inputs and target outputs [26]. The minimum, maximum, mean and standard deviation value of the wavelet features of vowel signal from the first subset were used as the input vectors of the ANN. 1, 3, 5, 7, 9 and 11 were the target value for vowel /a/, /i/, /u/, /ə/, /e/ and /o/, respectively. Backpropagation algorithmwas used to trainthe ANN. The algorithm involved an input vector, compared the network output to the desired output for that vector, and updated each weight by an amount corresponding to the derivative of the error with respect to that weight times some learning rate [32].Learning rate of 0.5, minimum error of 0.0001 and minimum gradient of 0.000001 were determined.
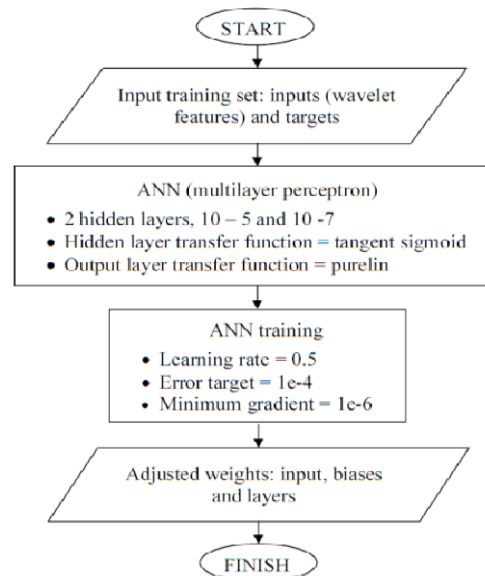
Figure 4. ANN training phase flowchart

3. Testing

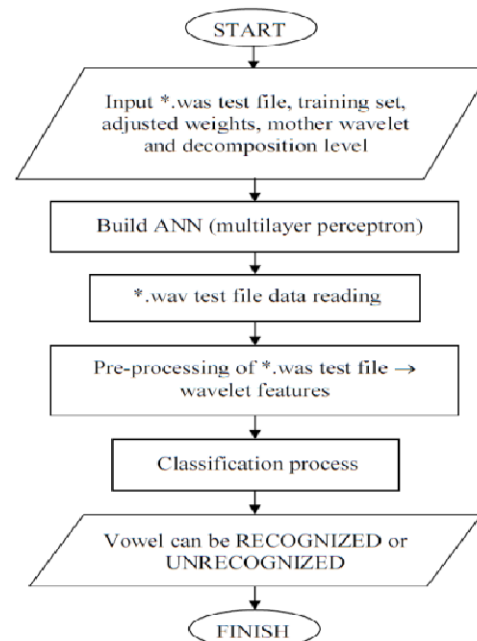Figure 5. ANN testing phase flowchart.

3. RESULTS AND DISCUSSION

A. Vowel Signal

The total of the recorded vowel samples in this work is 120 in case male speakers and 120 in case of female speakers. Figure 6 shows the original Indonesian vowel signals of the first male speaker, Figure 7 shows the spectrally flattened signals and Figure 8 shows the signals which have been frame blocked and windowed.
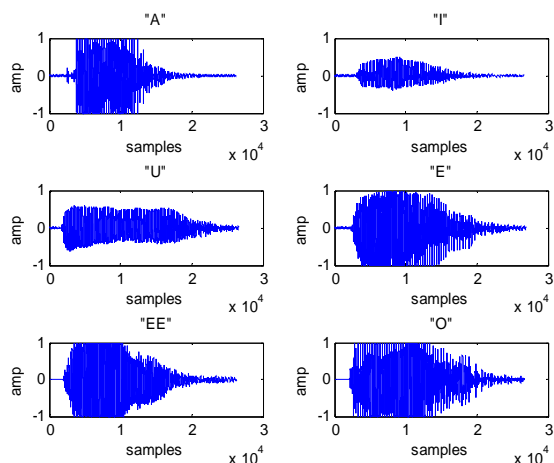
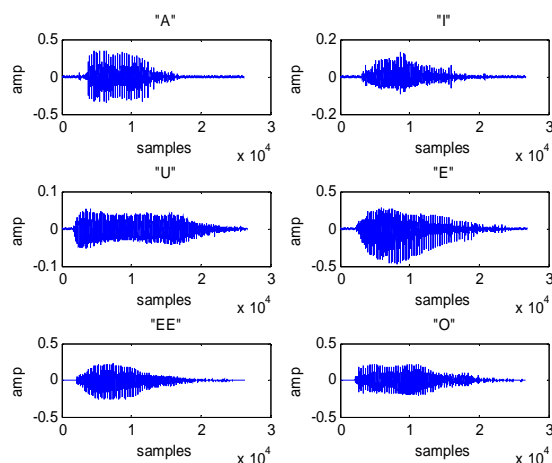Figure 6. Original signals of Indonesian vowels of a male speaker 1
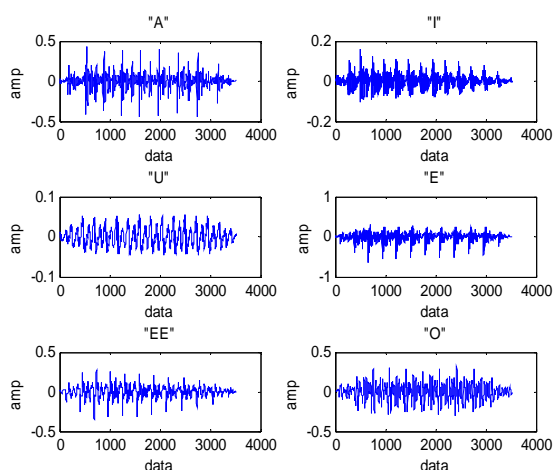
Figure 7. Pre-emphasized signals

Figure 8. Frame blocked dan windowed signals

The vowel database whether from male speakers or female speakers was divided into two subsets. The first subset was used for ANN training phase, 90 vowel samples, and the second subset was used for ANN testing phase, 30 vowel samples.

B. Feature Extraction Results

Wavelet coefficients were the results of vowel feature extraction by using DWT in this work. There are three sets wavelet features obtained from each vowel signal for each used mother wavelet:

1. Decomposition level 2 obtained 3 wavelet features: 1 approximation coefficient cA, and 2 detail coefficients cD.
2. Decomposition level 4 obtained 5 wavelet features: 1 approximation coefficient cA, and 4 detail coefficients cD.
3. Decomposition level 6 obtained 7 wavelet features: 1 approximation coefficient cA, and 6 detail coefficients cD.

C. Training and Testing Results

1. Training

The minimum, maximum, mean and standard deviation value of the wavelet features of vowel signal from the first subset were used as the input vectors of the ANN in this work

Multiple passes of training was required on the entire training set(rather like a person learning a new skill)until minimum error target or gradient could be reached.Each pass is called iteration or epoch.Table 2 and 3 show epoch number of each training set in case of male speakers and female speakers. It is also shown that the ANN which was used input vectors from decomposition level 6 needed less epoch than the ANN

which was used input vectors from decomposition level 4 and furthermore level 2. It because higher decomposition level obtained more wavelet features from vowel signal and therefore gave more input vectors for the ANN. More input vectors means more knowledge for the ANN to learning.

Table 2. Epochs in case of male speakers			
Wavelet	Dec. level	Neurons in 2nd hidden layer	Epoch
db4	2	5	143693
		7	148987
	4	5	20184
		7	12842
	6	5	3209
		7	2975
sym4	2	5	63603
		7	65900
	4	5	14109
		7	8641
	6	5	3722
		7	2535

Table 3. Epochs in case of female speakers			
Wavelet	Dec. level	Neurons in 2nd hidden layer	Epoch
db4	2	5	241849
		7	220597
	4	5	35788
		7	29301
	6	5	11134
		7	12087
sym4	2	5	612667
		7	868331
	4	5	20280
		7	23468
	6	5	9654
		7	9421

Training the ANN means adapting its connections so that the network exhibits the desired computational behavior for the input patterns. The process involved modifying the weights which was random at the initial. Adjusted weights of the ANN were obtained from this phase.Training of the ANN with different number of neurons in the second hidden layer within same inputsresulted different matrix of weights. The ANN with 5 neurons in the second hidden layer obtained first hidden layer weights in matrix 5×10, second hidden layer weights in matrix 1×5 and first hidden layer bias weights in matrix 5×1. The ANN with 7 neurons in the second hidden layer obtained first hidden layer weights in matrix 7×10, second hidden layer weights in matrix 1×7 and first hidden layer bias weights in matrix 7×1.

2. *Testing*

The ANN in this phase used the adjusted weights from the training phase. Threshold of 0.8 was used to determining the output decisions. Table 4 shows ANN outputs classification based on the threshold.

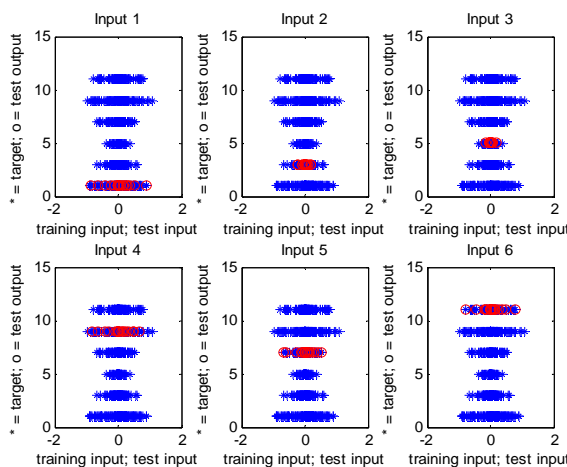| Table 4. ANN outputs classification		
ANN output	Decision	Output
$0.2 \leq$ output ≤ 1.8	recognized as /a/	A
$2.2 \leq$ output ≤ 3.8	recognized as /i/	I
$4.2 \leq$ output ≤ 5.8	recognized as /u/	U
$6.2 \leq$ output ≤ 7.8	recognized as /e/	E
$8.2 \leq$ output ≤ 9.8	recognized as /ə/	EE
$10.2 \leq$ output ≤ 11.8	recognized as /o/	O
others	could not be recognized	X

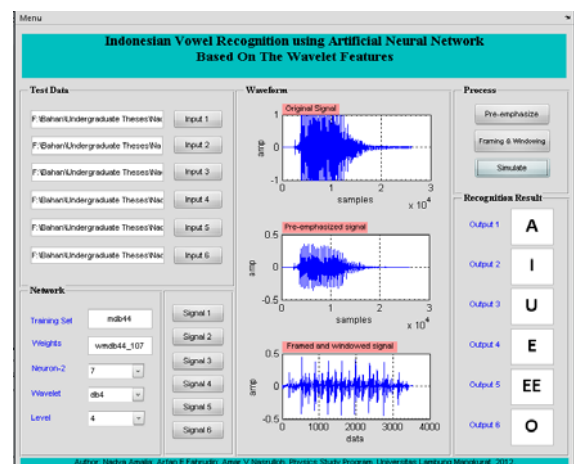Figure 10. Classification graph of the ANN outputs

Figure 11. Indonesian vowel recognition testing module

The value of the threshold depends on the range between each input targets. Bigger input targets range would need bigger threshold value. Figure 10 visualizes the classification graph of the outputs of the ANN and Figure 11 visualizes the testing module.

Based on the testing results, some vowel samples have been recognized wrongly. Some /u/ samples were recognized as /o/ or vice versa. It happened because both vowels has the same inventory, rounded lips and back tongue. Recognition rate shows the effectiveness of the methods in this work:

$$\text{Recognition rate} = \frac{\text{True data}}{\text{Total data}} \times 100\% \qquad (10)$$

30 vowel samples were used for the testing phase. The outputs of the ANN in testing phase can be classified into two. Firstly, true data which are the vowel samples that could have been recognized rightly. Secondly, false data which are the vowel samples that whether were recognized wrongly or could not be recognized at all.

Table 5. Recognition rates in case of male speakers

Wavelet	Dec. level	Neurons in 2nd hidden layer	Recognized	Not Recognized	Recognition rate (%)
db4	2	5	24	6	80
		7	21	9	70
	4	5	24	6	80
		7	22	8	73.33
	6	5	23	7	76.67
		7	24	6	80
sym4	2	5	22	8	73.33
		7	23	7	76.67
	4	5	21	9	70
		7	23	7	76.67
	6	5	22	8	73.33
		7	27	3	90

Table 6. Recognition rates in case of female speakers

Wavelet	Dec. level	Neurons in 2nd hidden layer	Recognized	Not Recognized	Recognition rate (%)
db4	2	5	17	13	56.67
		7	18	12	60
	4	5	24	6	80
		7	17	13	56.67
	6	5	23	7	76.67
		7	21	9	70
sym4	2	5	17	13	56.67
		7	16	14	53.33
	4	5	21	9	70
		7	16	14	53.33
	6	5	20	10	66.67
		7	24	6	80

Table 7. Overall recognition rates

Wavelet	Dec. level	Neurons in 2nd hidden layer	Recognition rate (%)
db4	2	5	68.33
		7	65
	4	5	80
		7	65
	6	5	76.67
		7	75
sym4	2	5	65
		7	65
	4	5	70
		7	65
	6	5	70
		7	85
	Average		70.83

Besides the factors within the ANN, recognition rate was also influenced by physical factors within the vowel signal. Unwanted noise in vowel sound could give different signal representation from the true

signal. The common source was background noise which is always present in any location. Therefore, the obtained approximation coefficients were significantly different. The way the speakers pronounced the vowel was also given an effect. The perfect recordings are each vowel is pronounced naturally, not too low but also not too loud.

4.   CONCLUSION

Indonesian vowel recognition using artificial neural network based on the wavelet features has been presented in this paper. Different mother wavelets and decomposition levels were implemented to analyze each vowel signal. It is shown that decompositionlevel 2 in DWT obtained 3wavelet features, level 4 obtained 5 wavelet features and level 6 obtained 7 wavelet features. Artificial neural network were used as the recognition system. From the experimental results, it can be concluded that the presented method can recognize the Indonesian vowel well. An overall recognition rate of 70.83% could be achieved. In case of male speakers the highest recognition rate is 90% and in case of female speakers the highest recognition rate is 80%. Both were achieved by using sym4 as the mother wavelet, decomposition level 6 and ANN with 7 neurons in the second hidden layer. For further work, in order to get better recognition rate, more vowel samples for training phase is needed. The vowel samples should be recorded in silent situation. Another mother wavelet and decomposition level can also be implemented.

REFERENCES

[1]   Martí, Fèlix, et.al. *Words and Worlds: World Languages Review*. Bilingual Education andBilingualism. Clevedon, Buffalo [N.Y.]: Multilingual. 2005; 52.
[2]   Riza, Hammam. Resources Report on Languages of Indonesia. *The 6th Workshop on Asian Languae Resources.* 2008; 93-94.
[3]   Woolard, Kathryn A. Review of Shifting Languages: Interaction and Identity in Javanese Indonesian (J. Joseph Errington) *in Language and Society.* 2000.
[4]   Paauw Scott. One land, one nation, one language: An analysis of Indonesia's national language policy. In H. Lehnert-LeHouillier and A.B. Fine (Eds.), *University of Rochester Working Papers in the Language Sciences*, 5(1), 2-16.
[5]   Yana, E. Pengantar Jaringan Saraf Tiruan. *Artikelkuliah*. http://trirezqiariantoro._files._wordpress.com_/2007/05/ jaringan__syaraf_tiruan.pdf
[6]   Puspitaningrum D. *Pengantar Jaringan Saraf Tiruan*. Penerbit Andi: Jogjakarta. 2006.
[7]   Rogers, Henry. *The sounds of Language: An Introduction to Phonetics*. Pearson: UK. 2000.
[8]   Stevens, Kenneth N. Acoustic Phonetics, *Current Studies in Lingistics*, No. 30. Cambridge MA: MIT. 1998
[9]   Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan Indonesia. *Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan dan Pedoman Umum Pembentukan Istilah.* CV Pustaka Setia: Bandung. 1996.
[10] Sani, Auliya, S Sakti, G Neubig, T Toda, A Mulyanto, S Nakamura. *Towards Language Preservation: Preliminary Collection And Vowel Analysis of Indonesian Ethnic Speech Data*.
[11] Lalipola H. A Generative Approach to The Phonology of Bahasa Indonesia. *Pasific Linguistics* 1981; 34.
[12] Sadeghi, Vahideh Sadat, Khashayar Yaghmaie. Vowel Recognition using Neural Networks, *IJCSNS International Journal of Computer Science and Network Security*. 2006; 6(12): 154-158.
[13] Prica, Biljana and Siniˇsa Iliˊc. Recognition of Vowels in Continuous Speech by Using Formants, *Facta Universitatis (NIˇS), Ser.: Elec. EnerG*. 2010; 23(3): 379-393.
[14] Sunny S, David PS, KP Jacob. Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words. *International Journal of Computer Applications*. 2012; 38(9).
[15] Mallat SG. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1989; 11(7).
[16] Rao Nikhil. Speech Compression Using Wavelets. *Undergraduate Thesis*. School of Information Technology and Electrical Engineering. The University of Queensland. 2001.
[17] Merry RJE, M Steinbuch, MJG van de Molengraft. Wavelet Theory and Applications. *Thesis*. Eindhoven University of Technology: Eindhoven. 2005.
[18] Mallat SG. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1989; 11(7).
[19] Strang G, TNguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Second Edition. 1997; 7-1.

[20] Misiti M, et al. Wavelet Toolbox For Use with Matlab, *User's Guide Version 2*. The MathWorks, Inc.: Natick. 2002.

[21] Schneiders MGE. Wavelets In Control Engineering. *Master's Thesis*, Eindhoven University of Technology. DCT nr. 2001; 38.

[22] Haykin S. *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, Prentice-Hall: New York. 1999.

[23] Agustini K. Biometrik Suara dengan Transformasi Wavelet Berbasis Orthogonal Daubechies. *50 Gematek Jurnal Teknik Komputer,* 2007; 9(2).

[24] Addison PS. *The Illustrated Wavelet Transform Handbook*. IOP Publishing Ltd. 2002.

[25] Christopher, Bishop. *Neural networks for Pattern Recognition*. Oxford University Press. 1995.

[26] Demuth H, Beale M. Neural Network for Use with Matlab, *User's Guide Version 3.0*. The MathWorks, Inc.

[27] MathWorks. *Matlab Wavelet Toolbox 3.0*. The MathWorks Inc., 24 Prime Park Way, Natick, MA: USA. 2006.

[28] Syukron H, Haikal. *Audio Sampling*. http://www.eletroindonesia.com/elektro/elek35a.html. 2001.

[29] Al-Haddad S, S Samad, AHussain, K Ishak, A Noor. Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering. *American Journal of Applied Sciences*. 2009; 6; 290-295.

[30] Shahrul Azmi MY, Sazali Yaacob, Paulraj MP. Vowel recognition using First Formant Feature. *International Journal of Advancements in Computing Technology*. 2009; 1(1).

[31] Hu Dingyin, Li Wei, Chen Xi. *Feature Extraction of Motor Imagery EEG Signals Based On Wavelet Packet Decomposition*. Proceedings of IEEEIICME International Conference on Complex Medical Engineering. 2001.

[32] Chaudhari, Jyoti C. Design of Artificial Back Propagation Neural Network for Drug Pattern Recognition, *International Journal on Computer Science and Engineering (IJCSE)*. 2010.

BIOGRAPHIES OF AUTHORS

Nadya Amalia was born in Wasah Hulu, Kalimantan Selatan, in 1991. Shereceived an undergraduate degree in Physics (2012) from Universitas Lambung Mangkurat (UNLAM), Banjarbaru Indonesia. Her current researchinterests are Digital Signal Processing, Pattern Recognition and Energy.

Arfan Eko Fahrudin was born in Kediri, in 1979. He received anundergraduate degree in Physics (2004) from InstitutTeknologiSepuluhNopember (ITS) and a graduate degree in Electrical Engineering (2010) fromUniversitasGadjahMada (UGM), Indonesia. In 2005, he joined the PhysicsStudy Program of Universitas Lambung Mangkurat as a lecturer. He receivedthe Academic Achievement Directorate General Higher Education (DIKTI) Scholarship (BPPS) for his graduate degree. Currently he is member of HFI (Indonesian Physical Society). His current research interests are ImageProcessing, Pattern Recognition and Medical Instrumentation.

Amar Vijai Nasrulloh was born in Surabaya, in 1978. He received anundergraduate degree in Physics (2004) from Institut Teknologi Sepuluh Nopember (ITS) and a graduate degree in Electrical Engineering (2010) specialty in Biomedical Engineering Program from Institut Teknologi Bandung (ITB), Indonesia. He received the Academic Achievement Directorate General Higher Education (DIKTI) Scholarship (BPPS) for his graduate degree. In 2005, he joined the Physics Study Program of Universitas Lambung Mangkurat as a lecturer. Currently he is member of HFI (Indonesian Physical Society). He became a member of IEEE in 2011. His current researchinterests are Video Processing, Imaging and Image Processing, Biomechanics and Medical Instrumentation.