MATEC Web of Conferences **154**, 03015 (2018) https://doi.org/10.1051/matecconf/201815403015
ICET4SD 2017

Syllables sound signal classification using multi-layer perceptron in varying number of hidden-layer and hidden-neuron

Domy Kristomo[1,2,*], *Risanuri* Hidayat[1], and *Indah* Soesanti[1]

[1]Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Grafika Street No. 2, Yogyakarta, 55281 Indonesia
[2]Study Program of Computer Engineering, STMIK Akakom Yogyakarta, Raya Janti Street 143 Karang Jambe Yogyakarta, 55198 Indonesia

Abstract. The research on signal processing of syllables sound signal is still the challenging tasks, due to non-stationary, speaker-dependent, variable context, and dynamic nature factor of the signal. In the process of classification using multi-layer perceptron (MLP), the process of selecting a suitable parameter of hidden neuron and hidden layer is crucial for the optimal result of classification. This paper presents a speech signal classification method by using MLP with various numbers of hidden- layer and hidden-neuron for classifying the Indonesian Consonant-Vowel (CV) syllables signal. Five feature sets were generated by using Discrete Wavelet Transform (DWT), Renyi Entropy, Autoregressive Power Spectral Density (AR-PSD) and Statistical methods. Each syllable was segmented at a certain length to form a CV unit. The results show that the average recognition of WRPSDS with 1, 2, and 3 hidden layers were 74.17%, 69.17%, and 63.03%, respectively.

1 Introduction

One of main goal in speech recognition is to obtain the best accuracy for recognizing or classifying the speech signal uttered by speaker. The technology of speech recognition is currently used in many applications, such as smart phones, security systems, etc. However, these systems still have some difficulties in distinguishing syllables or word that sound similar. The common stages in speech recognition system are pre-processing, feature extraction and recognition stage. In the recognition stage, the process of selecting the suitable parameter in the classifier system is crucial for optimal result of classification.

Many classification methods were applied for classifying the speech signal by previous researchers. Several methods such as the Hidden Markov Model (HMM), Support Vector Machines (SVM), Gaussian Mixture Model (GMM), and Multi-layer Perceptron (MLP) or Artificial Neural Network (ANN) as classifiers [1 - 8]. MLP or ANN is a method when learning algorithm is performed and converged. It involves of weights and the ability of the underlying networks to implement desired function using sufficient number of hidden neuron. Generally, hidden layer is not needed in case of small samples number in the data set. However, there is no proved or accepted theory in determining the numbers of neurons in hidden layer for function approximation. The number of neurons in hidden layer influenced the network.

Several previous studies has been done using MLP classifier [2, 3, 8, 9] for speech classification. In [5], three discrete wavelet families (db, sym, and coif) with a different number of coefficients was used and evaluated with two classifiers (GMM and MLP). The result in computation time showed that MLP has better performance than GMM [5]. In [8], MLP was used for classification of Hindi CV syllables. In [3], the MLP was used for classifying syllables by using combination of Discrete Wavelet Transform (DWT) and statistical features with variation of mother wavelet of Haar, Coiflet and Daubechies. The experiment result showed that Daubechies is the most effective mother wavelet compared to Haar and Coif. In [2], The MLP was used for classifying syllables sound by using DWT, Renyi Entropy (RE), and Autoregressive Power Spectral Density (AR-PSD) features.

This paper presents the classification of Indonesian CV syllables sound signal by using the MLP in varying number of hidden neuron, and the signal processing by using DWT, RE, AR-PSD, and statistical for generating features [10]. Five feature set are performed in this study. Feature Set 1 is the combination of the DWT and statistical (WS). The wavelet type used is db2 at the 7th level of decomposition [11]. Feature Set 2 is the RE features. Feature set 3 is the combination of AR-PSD and statistical features in frequency and time domain (PSDS) [2]. Feature Set 4 is the combination of AR-PSD and the RE features after selected by using Correlation-based feature selection method or CFS (RPSDS). Feature Set 5 is the combination of WS, RE, and PSDS (WRPSDS).

[*] Corresponding author: domykristomo@gmail.com

2 Methodology

A database of 360 CV syllable utterances taken from six different speakers was created. It is formed by three Indonesian consonants (/k, g, l, r/) followed by the three vowels (/a, i, u/). The consonant /g/ and /k/ represent articulation place of the velar as well as the part of the stop consonants [1, 12], while /l/ and /r/ represent articulation place of the alveolar. Frequency sampling used was 8 kHz and 16 bits mono per sample.

After the phase of recording, the next step was the cropping phase, which was basically a windowing phase to form a rectangular window of the signal. From the acoustic study by the previous researcher (Sharma et al. 2013), it was found that duration for all relevant acoustic parameters was about 60 ms [1]. Therefore, the duration manually taken for each CV unit in this study was about 60 ms, starting from release burst of the associate consonant to steady state of the following vowel. The significant events regions of the CV unit /ka/ are shown in Fig. 1a [13]. The next phase was the peak normalization. By applying peak normalization, the signal magnitude variation which is caused by the differences in the recording condition (such as speaker distance and loudness factor) can be avoided [14].

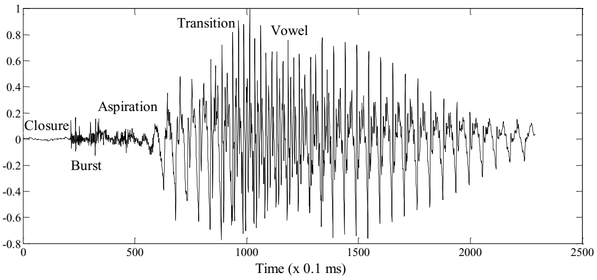

Fig 1a. Example of syllable /ka/ and its significant events regions

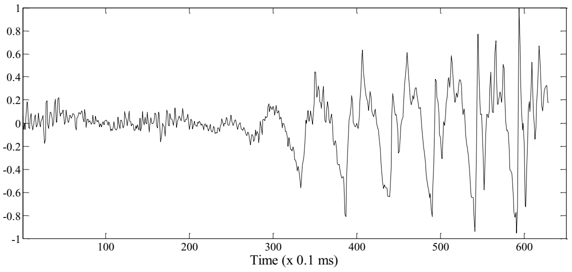

Fig 1b. The syllable /ka/ after windowing phase

2.1. Wavelet

The wavelet transform (WT) is a signal processing method which can decompose a signal into several bands using a low-pass filter and a high-pass filter. In this part, feature extraction using DWT at 7th level of decomposition was conducted. In the decomposition phase of DWT, only at a lower frequency band which is also called as approximation. By decomposing at 7th level, it gives the highest frequency band of 2000-4000 Hz and the lowest frequency band of 0–31.25 Hz. More level decomposition will be insignificant to improve recognition rate because a very low frequency band will not have discriminatory information [15].

In the DWT, the process of selecting the suitable mother wavelet is crucial for optimal result of classification. Based on previous research [1, 3, 11], it was found that Daubechies 2 (db2) was the one of the effective mother wavelet. The Daubechies wavelet of class D-2N can be written as:

$$\psi(x) := \sqrt{2}\sum_{k=0}^{2N-1}(-1)^{k} h_{2N-1-k}\varphi(2x-k) \qquad (1)$$

Where h0,…, h2N-1 ∈ ℝ are the constant filter coefficients satisfying the condition, and φ is the (Daubechies) scaling function. After transformation process by using DWT, then the result was a signal in frequency domain. The moving average feature was calculated of each twenty sample of the signal magnitude until the maximum sample of the signal magnitude. As the additional feature, the signal in frequency domain was calculated using a statistical method [3].

2.2 Renyi Entropy

The Renyi entropy (RE) is a generalization of the Shannon entropy, the collision entropy, the Hartley entropy, and the min entropy. The function of generalized entropy for discrete variable X can be defined in Equation below.

$$H_{\alpha}(X) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} P_{i}^{\alpha}\right) \qquad (2)$$

Where pi is the probability of X belonging to possible outcome, o_1, o_2,..., o_n. The order of entropy, α, has the constraint of $\neq$ 1. In special case of $\alpha = 1$, it converges to Shannon entropy [16], [17].

2.3 Autoregressive Power Spectral Density (AR-PSD)

In this study, PSD using Yule-Walker AR algorithm was performed. The AR model in P order can be defined in equation bellow:

$$x_{pp}(t) = -\sum_{k=1}^{p} a_k x_{pp}(t-k) + e(t) \qquad (3)$$

Where
a_k = AR's Coefficient

Then, using 256 point $x_{pp}(t)$ with Hamming's window, AR-PSD estimation can be formulated as follows:

$$P_{AR}(f) = {T\sigma_W^2}\Big/{|1 + \sum_k^p a_k e^{-2\pi fkT}|^2} \tag{4}$$
$$= T \sum_{m=1}^{C-1} r_{xx} e^{-2\pi fmkT}$$

Where r_{xx} is extrapolation of data series autocorrelation bias estimation data from AR model, T is the period of sampling, and σ_W^2 is the variance of the drive noise input.

2.4 Single Hidden Layer in Multi-layer Perceptron (MLP)

In the classification process by using MLP, the process of selecting the suitable parameter and architecture is crucial for the optimal result of classification [18], [19]. The architecture used in this section consists of three layer, they are input layer, hidden layer, and output layer. The input layer represents the features of each feature extraction method (WS, RE, PSDS, RPSDS, or WRPSDS). The hidden layer consists of 1-20 hidden neurons. The output layer consist of twelve neurons, it represents the classification result of the syllables. To estimate the reliability of the classification results, the data verification was performed. The verification technique used on the test set was k-fold cross validation or the hold out method [20].

2.5 Hidden Layers in MLP

In this experiment, we used 2 layers in Hidden layer. Based on the previous experiment on single layer of hidden neuron, the optimal result for WRPSD was 55 nodes. Therefore in this part we used 55-55 nodes.

2.6 Hidden Layers in MLP

In this part of experiment, we used 3 layers. As the previous research recommendation [9], the nodes architecture for 3-hidden layer was 20-20-15 nodes.

3 Result and Discussion

In this study, the number of features generated by using WS, RE, PSDS, RPSDS, and WRPSDS were twenty-nine, twenty, thirteen, nineteen, and sixty-two features, respectively. After feature extraction, the next phase was classification that uses MLP-BP. The parameter of the learning rate and the momentum was 0.3 and 0.2, respectively. In the previous study [14], the experiment for the number of hidden neurons 0 to 20 has been done. In this study we continued the experiment for the number of hidden neurons 20-60 as shown in Fig. 2.

Figure 2 shows the percentage accuracy for WS, RE, PSDS RPSDS, and WRPSDS in varying number of hidden neuron. The result showed that WRPSDS has the highest score in average accuracy, but at ninth hidden neuron, the score of WS is higher than WRPSDS. It indicated that the number of hidden neuron can influence the classification result, but the increase of accuracy was not linear.

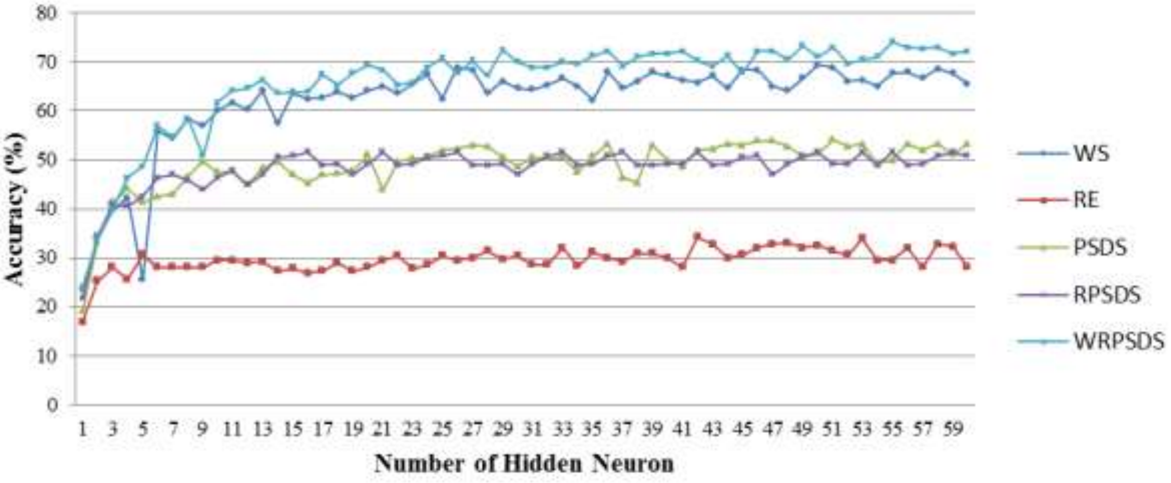

Fig 2. Accuracy for each feature extraction method with various number of hidden neuron in MLP

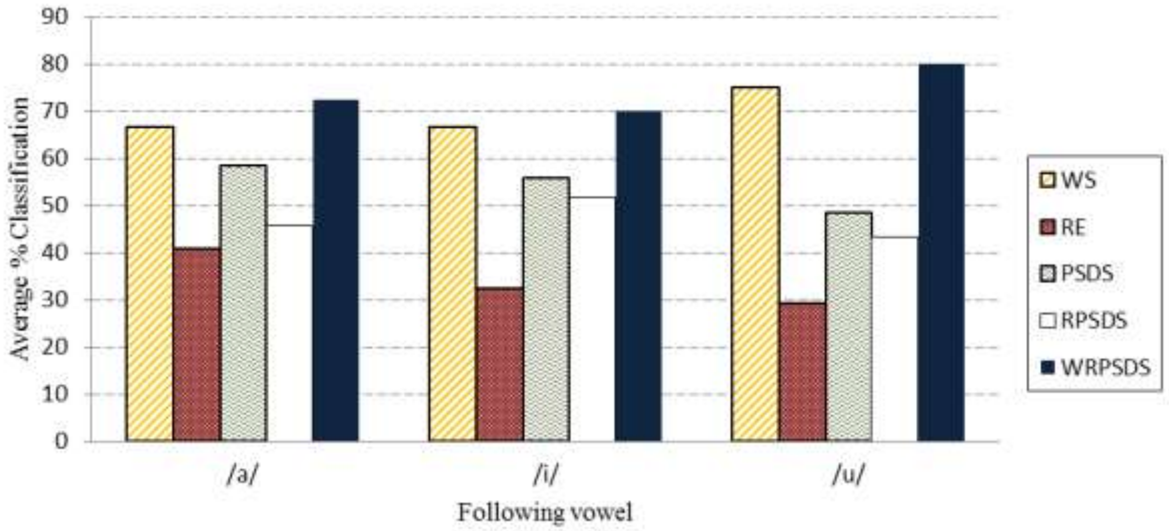

Fig 3. Graph of the average percentage classification at the optimal number of hidden neuron of each feature extraction method for classifying the Indonesian consonant following vowel contexts

Table 1. Accuracy for each feature extraction method with various hidden layer in MLP

Feature extraction Method	1-Hidden Layer (55 nodes)				2-Hidden Layer (55-55 nodes)				3-Hidden Layer (20-20-15 nodes)			
	Following vowel			Average % recognition	Following vowel			Average % recognition	Following vowel			Average % recognition
	/a/	/i/	/u/		/a/	/i/	/u/		/a/	/i/	/u/	
WS	65	61.67	76.67	67.78	59.17	60.83	68.33	62.78	55.83	59.17	6	58.33
RE	31.67	28.33	28.33	29.44	34.17	26.67	28.33	29.72	34.17	23.33	17.5	25
PSDS	51.67	52.5	45.83	50	55	56.67	51.67	54.44	45.83	4	39.17	41.67
RPSDS	45.8	51.65	43.32	46.92	49.17	53.33	46.67	49.72	48.33	43.33	38.33	43.33
WRPSDS	**72.5**	**70**	**79.99**	**74.17**	64.17	67.5	75.83	69.17	68.33	58.33	62.5	63.03

Table I shows the percentage classification scores using Ten-fold cross validation (Ten-FCV) for 1, 2 and 3 hidden layer. The experiment result showed that the average recognition for 1, 2, and 3 hidden layers that using WRPSDS were 74.17%, 69.17%, and 63.03% respectively. It indicated that the MLP architecture in 1-hidden Layer (55 nodes) gives better performance of classification compared to 2-Hidden Layer (55-55 nodes) and 3-Hidden Layer (20-20-15) nodes.

In case of /a/, the highest score was 72.5% by using WRPSDS features. (For the vowel /i/ the highest classification score was 70%. In case of /u/ the highest score was 79.99% which was the highest score among the other vowel. Some feature showed better performance in 2-Hidden Layer architecture (RE, PSDS, RPSDS), but overall 1 layer was still better.

4 Conclusion

In this paper, a classification of the Indonesian syllables sound using classifier of multi-layer perceptron in varying number hidden layer and hidden neuron fusing with Wavelet, Renyi entropy (RE), AR-PSD features was proposed and implemented. Based on the experimental result presented in this paper, it can be concluded that the MLP architecture in 1-hidden Layer (55 nodes) when fusing with WRPSDS gives a better performance of classification score compared to 2-Hidden Layer (55-55 nodes) and 3-Hidden Layer (20-20-15) nodes as shown by accuracy of 74.17%, 69.17%, and 63.03% respectively. Some feature such as RE, PSDS, RPSDS showed better performance in 2-Hidden Layer architecture, but overall 1 hidden layer architecture was still better. The future work recommended for this re-search is to use bigger syllable dataset, applied to the Indonesian stop consonant or the other place of articulation (such as labial, dental, etc.), to use different combination of feature extraction technique, and to use different testing procedure of classification process.

References

1. R. P. Sharma, O. Farooq, and I. Khan, “Wavelet based sub-band parameters for classification of unaspirated Hindi stop consonants in initial position of CV syllables,” *Int. J. Speech Technol.*, **16**, no. 3, pp. 323–332 (2013)
2. D. Kristomo, R. Hidayat, and I. Soesanti, “Classification of the Syllables Sound Using Wavelet , Renyi Entropy and AR-PSD Features,” in *2017 IEEE 13th International Colloquium on Signal Processing & its Application (CSPA 2017)*, **13**, pp. 97–102 (2017)
3. D. Kristomo, R. Hidayat, and I. Soesanti, “Feature extraction and classification of the Indonesian syllables using Discrete Wavelet Transform and

statistical features," in *2016 2nd International Conference on Science and Technology-Computer (ICST)*, **2**, pp. 88–92 (2016)

4. S. Hidayat, R. Hidayat, and T. B. Adji, "Speech recognition of CV-patterned indonesian syllable using MFCC, wavelet and HMM," *J. Ilm. Kursor*, **8**, no. 2, pp. 67–78 (2015)
5. P. Král, "Discrete Wavelet Transform for automatic speaker recognition," *Image Signal Process. (CISP), 2010 3rd Int. Congr.*, **7**, pp. 3514–3518 (2010)
6. X. Zhao, Z. Wu, J. Xu, K. Wang, and J. Niu, "Speech Signal Feature Extraction Based on Wavelet Transform," *2011 Int. Conf. Intell. Comput. Bio-Medical Instrum.*, no. 1, pp. 179–182 (2011)
7. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, **44**, no. 3, pp. 572–587 (2011)
8. C. Chandra and B. Yegnanarayana, "A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances," *IEEE Trans. Speech Audio Process.*, **10**, no. 7, pp. 472–480, (2002)
9. G. Dede and M. H. Sazlı, "Speech recognition with artificial neural networks," *Digit. Signal Process.*, **20**, no. 3, pp. 763–768 (2010)
10. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. United States ofAmerica (2009)
11. R. Hidayat, Priyatmadi, and W. Ikawijaya, "Wavelet based feature extraction for the vowel sound," in *2015 International Conference on Information Technology Systems and Innovation (ICITSI),* pp. 1–4 (2015)
12. F. L. Hardjono and R. A. Fox, "Stop Consonant Characteristics: VOT and Voicing in American-Born-Indonesian Children's Stop Consonants," The Ohio State University (2011)
13. A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points," *Circuits, Syst. Signal Process.*, **31**, no. 4, pp. 1459–1474 (2012)
14. D. Kristomo, R. Hidayat, I. Soesanti, and A. Kusjani, "Heart sound feature extraction and classification using autoregressive power spectral density (AR-PSD) and statistics features," in *AIP Conference Proceedings*, **1755**, pp. 90007-1-90007–7 (2016)
15. O. Farooq and S. Datta, "Phoneme recognition using wavelet based features," *Elsevier Inf. Sci.*, **150**, pp. 5–15 (2003)
16. A. Rényi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, pp. 547–561 (1961)
17. C. Y. Kee, S. G. Ponnambalam, and C. K. Loo, "Binary and multi-class motor imagery using Renyi entropy for feature extraction," *Neural Comput. Appl.*, pp. 1–12 (2016)
18. N. M. Nawi *et al.*, "The Effect of Pre-Processing Techniques and Optimal Parameters selection on Back Propagation Neural Networks," **7**, no. 3, pp. 770–777 (2017)
19. A. Kuri-Morales, "The Best Neural Network Architecture," *Springer*, (2015)
20. R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artif. Intell.*, **14**, no. 12, pp. 1137–1143 (1995)