Journal of Engineering Science and Technology
Vol. 13, No. 9 (2018) 2964 - 2977
© School of Engineering, Taylor's University

COMPARISON OF FIVE CLASSIFIERS FOR CLASSIFICATION OF SYLLABLES SOUND USING TIME-FREQUENCY FEATURES

DOMY KRISTOMO[1,2,*], RISANURI HIDAYAT[1], INDAH SOESANTI[1]

[1]Department of Electrical Engineering and Information Technology of Engineering,
Faculty of Engineering, Universitas Gadjah Mada,
Jalan Grafika, No. 2, Yogyakarta, 55281 Indonesia
[2]Department of Computer Engineering, STMIK Akakom Yogyakarta
Jalan Raya Janti 143 Karang Jambe Yogyakarta, 55198 Indonesia
*Corresponding Author: domy.kristomo@mail.ugm.ac.id

Abstract

In a speech recognition and classification system, the step of determining the suitable and reliable classifier is essential in order to obtain optimal classification result. This paper presents Indonesian syllables sound classification by a C4.5 decision tree, a Naive Bayes classifier, a Sequential Minimal Optimization (SMO) algorithm, a Random Forest decision tree, and a Multi-Layer Perceptron (MLP) for classifying twelve classes of syllables. This research applies five different features set, those are combination features of Discrete Wavelet Transform (DWT) with statistical denoted as WS, the Renyi Entropy (RE) features, the combination of Autoregressive Power Spectral Density (AR-PSD) and Statistical denoted as PSDS, the combination of PSDS and the selected features of RE by using Correlation-Based Feature Selection (CFS) denoted as RPSDS, and the combination of DWT, RE, and AR-PSD denoted as WRPSDS. The results show that the classifier of MLP has the highest performance when it is combined with WRPSDS.

Keywords: C4.5, Feature extraction, Multi-layer perceptron, Naive bayes, Random forest, SMO.

1. Introduction

Speech signal is the most natural and the fastest method of communication between humans. Speech recognition technology allows computer to recognize and comprehend human languages. Research in speech recognition was started in the 1950s [1]. According to Davis et al. [1], research related to vowel was conducted. Based on research by Olson and Belar [2], the first study for recognizing syllables was conducted. The study tried to recognize ten distinct syllables from a single talker, which was characterized by its spectral, amplitude, and frequency band. The system still depends on word spectral measurements mainly during vowel regions.

In the 1980s, the classifier system began to be applied to solve several speech recognition system problems. Several popular methods such as Support Vector Machines (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Multi-Layer Perceptron (MLP) or Artificial Neural Network (ANN) are commonly applied in recognizing and classifying speech signal [3-14].

Several studies on speech signal classification and recognition that utilize various types of classifier such as MLP [3, 8, 12, 15-18], HMM [11, 19-21], LDA [22-24], GMM [6], NB [13] and the other classifier have been conducted [25]. In several previous studies, MLP was used to classify the speech signal using the Mel-Frequency Cepstral Coefficient (MFCC) [17], Linear Prediction Coefficient (LPC) [8], Discrete Wavelet Transform (DWT) [3, 12, 15, 16, 18] and Wavelet Packet (WP) [18] features. According to Kristomo et al. [3], the MLP was utilized to classify the sound signal of syllables through Autoregressive Power Spectral Density (AR-PSD), Renyi entropy and DWT features. Based on Abriyono and Harjoko [8] research, the MLP was used to classify Indonesian syllables through MFCC and LPC features. The accuracy result was 49% for MFCC and 47% for LPC by using testing dataset. According to Kristomo et al. [12], the MLP was utilized to classify syllables sound by utilizing the features fusion-derived by DWT and statistics technique with the performance comparison of three wavelet mother functions namely Daubechies, Haar and Coiflet. The research result shows that Daubechies as wavelet mother functions is more effective than Coif and Haar. Farooq and Datta [15] proposed that the performance of MLP was compared with LDA for classifying Hindi phonemes by using DWT features. The research results indicated that MLP improves the classification performance significantly compared to LDA. According to Trivedi et al. [16], the MLP was utilized to classify the words by applying DWT features. Dede and Sazli [17] explained that three types of MLP were used to classify Turkish digits. Based by Daqrouq and Al-Azzawi [18], the PNN was used to classify Arabic vowel by using a combination of WT and LPC features.

This research was conducted in accordance with the previous study [3, 12]. In this study, we performed a comparison of five different classifiers of Sequential Minimum Optimization (SMO), NB, RF, C4.5 and MLP. Five different Feature Set (FS) were performed in this study [3]. The first FS was derived by combining the DWT with Statistics (WS) methods. Wavelet mother function used is DB2 with 7-level decomposition. The second FS was derived by the Renyi Entropy (RE). The third FS was derived by combining AR-PSD with statistics features in the domain of frequency and time (PSDS). The fourth FS was derived by combining of AR-PSD with the RE features after selection by utilizing

Correlation-based feature selection method (RPSDS). The fifth FS was derived by combining of WS, RE and PSDS (WRPSDS).

2. Research Method

2.1. Preprocessing

Data used in this study were taken from six male students. The speaker aged between twenty-five and thirty-five years old. All speakers are Indonesian native speakers. The speakers were asked to do a recording in 1 second by repeating each syllable utterance for five times using laptop and microphone as the equipment in an open area. The speech data was sampled at 8 kHz with 16-bits mono/sample. The database used in this step are Indonesian CV syllables, which are formed by consonants /*k, g, l, r*/ and vowels /*a, i, u*/. Each consonant represents different Place of Articulation (POA) in which, */k/* and */g/* is for the velar articulation and part of the stop consonants [22, 26], while */r/* and */l/* is for the alveolar articulation. Since there are 12 syllables for each speaker recorded 5 times, it means, that each speaker recorded 60 syllables, while the whole syllables data accumulated is 360 utterances.

The next step after the recording step is segmentation step. In this step, a rectangular window of the signal is formed. The voice is segmented in order to obtain the speech database by using audio editing software. From the previous acoustic study [22, 27], it was suggested that the length for all relevant acoustic parameters was about 60 ms [22]. Therefore, the duration manually segmented from each syllable to become Consonant-Vowel (CV) in this research was about 60 ms, starting from the initial position (burst) of the associate consonant to the steady state of the following vowel.

2.2. Feature Extraction Methods

Feature extraction or feature generation is a process of taking the characteristics contained in the signals by translating the signal into set parameters called feature vectors. This process is of paramount importance or the key stage in any audio signal classification task. Usually, there are two domains in obtaining the feature, namely time domain and frequency domain. Time domain refers to the variation of the amplitude of the signal in time whereas the frequency domain shows how much signals lie in the frequency ranges. In many cases of signal processing research, transform-based domain (i.e., frequency domain) usually can perform high information packing properties compared to the original input signal (i.e., time domain). However, feature extraction is a problem-dependent task, which is the combination of the 'designer's imagination' can benefit the extraction of informative and discriminative features [28].

The FS we applied in this study was designed to discriminate twelve classes of Indonesian CV syllables. In this study, five different feature set was performed. The FS1 is the combination of features derived by DWT with a statistical technique, which is denoted as WS. The wavelet mother function utilized was DB2 at the 7th level of decomposition. The FS2 is the Renyi Entropy features (RE). The FS3 is the fusion of features derived by AR-PSD and statistics in the domain of frequency and time, which is denoted as PSDS. The FS4 is the fusion of features derived by the RE with AR-PSD features after being selected by using Correlation-based feature selection method or CFS, which is denoted as RPSDS. The FS5 is the fusion of features derived by WS, RE and PSDS, which is denoted as WRPSDS.

2.2.1. Wavelet Transformation

The First FS (FS1) is a Wavelet Transform (WT) based feature. The WT is a method in which, signal is decomposed into several bands through a high-pass filter and a low-pass filter. In this section, extraction of feature applying DWT at the 7-level decomposition was performed. In DWT, the decomposition process is focused only on the lower frequency band or so-called the approximation [6]. By performing 7-level decomposition, it produces the highest frequency band of 2-4 kHz and the lowest frequency band of 0-0.03125 kHz. More decomposition level is not significant to improve classification performance due to a very low frequency will not have discriminatory information [15].

In the DWT, the process of selecting the appropriate wavelet mother function is essential in order to obtain the better classification result. It was stated by Sharma et al. [22] that Daubechies (DB) type of wavelet family was the appropriate wavelet mother function for speech. The class D-2N DB wavelet is given in the following Eq. (1):

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

where $h_0, \ldots, h_{2N-1}$ Є ℝ is the constant filter coefficients satisfying the condition and φ is the scaling function (Daubechies). The WT results in a signal in the domain of frequency. The wavelet-based moving average feature was obtained by calculating every 20 samples of the signal magnitude until 480 samples of the signal magnitude [12]. Additionally, the frequency domain signal was calculated using statistics technique to obtain five additional features.

2.2.2. **Renyi entropy**

The FS2 is generated by using Renyi Entropy (RE) method. RE is a generalization of the Shannon entropy, the collision entropy, the min-entropy and the Hartley entropy. The generalized entropy function for X variable can be formulated using Eq. (2).

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

where p_i is the probability of X associated with possible outcome, $o_1, o_2, \ldots, o_n$. In a certain case of $\alpha = 1$, it converges to Shannon entropy [29-31]. This method contributes to a total of 20 features.

2.2.3. **Autoregressive power spectral density (AR-PSD)**

In this FS (FS3), the AR-PSD, which utilizes Yule-Walker AR algorithm was conducted. The AR model in P order can be formulated in Eq. (3).

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

where

a_k = AR's Coefficient

Then, applying 256 point $x_{pp}(t)$ with Hamming's window, estimation of AR-PSD can be defined through Eq. (4).

$$\mathrm{P}_{\mathrm{AR}}(\mathrm{f}) = {T\sigma_W^2} \Big/ {|1 + \sum_k^P a_k e^{-2\pi f k T}|^2} \tag{4}$$

$$= T \sum_{m=1}^{C-1} r_{xx} e^{-2\pi f m k T}$$

where r_{xx} is an estimation of data series autocorrelation from *AR* model, *T* is sampling period and σ_W^2 is drive noise variance.

2.2.4. **RE, Correlation-based feature selection and AR-PSD**

In the FS4, we combined RE and AR-PSD, which is denoted as RPSDS. The number of features derived by using RPSDS was nineteen. It uses the Eqs. (2), (3), and (5). To reduce the vector dimension of RE, the CFS method was conducted [32, 33] by using the formula in Eq. (5).

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{5}$$

where $\underline{Ms}$ is the feature subset *S* heuristic (merit), $\overline{r_{ff}}$ is the mean feature inter-correlation, and $\overline{r_{cf}}$ is the average feature-class correlation ($f \in S$).

2.2.5. **WS, RE and AR-PSD**

In the FS5, we combined WS, RE and AR-PSD, which is denoted as WRPSDS. The number of features derived by using WRPSDS was sixty-two. In order to generate the feature, it uses the previous equations (Eqs. (1) to (4)). This feature is expected to have an advantage over the separate (time or frequency) domain generated features due to the better tiling of the time-frequency plane by each feature extraction method.

2.3. Classification Methods

2.3.1. Sequential minimal optimization

Sequential Minimum Optimization (SMO) is an algorithm for solving the optimization problem of Quadratic Programming (QP) at the Support Vector Machine (SVM). SMO has the ability for minimizing QP problem and optimization time. SVM is a machine learning technique based on Structural Risk Minimization (SRM) principal. The method aims to obtain the best hyperplane, which separates two classes at the input space, Fig. 1 illustrates the finding of the best hyperplane.

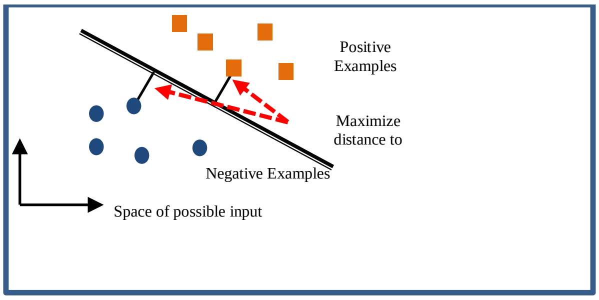

Fig. 1. Linear support vector machine.

2.3.2. Naive Bayes

Naive Bayes (NB) is a classification method based on probability, which has an assumption that each *X* variable has an independent characteristic. In other words, NB assumes that an existence of an attribute or feature is not correlated with the other feature. If *X* is the data sample with the unknown classes, *H* is the hypotheses with *X* is the data with *C* class, *P(H*) is the observed probability sample data, then *P(X*|*H*) is the probability of *X* sample data, assumed that the hypotheses of *H* are valid.

Due to the assumption that feature is not interrelated (conditionally independent), then *P(X*|*Ci*) can be defined through Eq. (6).

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \tag{6}$$

If *P(X*|C_i*)* has been defined, then the class of *X* data sample of syllables can be approached by calculating *P(X*|C_i*)*P(*C_i*)*. Class of C_i where *P(X*|C_i*)*P(Ci)* maximum is class of *X* sample of syllables.

2.3.3. C4.5

C45 is a decision tree used for classification with the concept of information entropy. To obtain a C4.5 pruned tree, decision creation is done by splitting each data attribute or syllable feature into a smaller subset to check different entropy and selecting the syllable feature with the highest gain. Splitting process will be stopped when instance subset included into the same class of syllable is found and leaf node will be made. If there is no leaf node found, C4.5 will create a higher destination node based on the expected class value of syllable.

2.3.4. Random forest

Random Forest (RF) is an ensemble classifier, which is the development of a decision tree with random factors. This method is commonly used as an ensemble learning approach for regression and classification. An RF consists of a randomized decision trees set and its prediction output. Generally, there are two forms injected by RF into its learning process; they were random parameterization and random sampling. Training parameters were chosen by Random parameterization during each decision trees training. In the training process, it is possible to use both or either of these two forms of randomness [34]. Several functions generated by RF used an extension of bagging techniques to overcome the over-fitting problem when is faced with small data sets [35].

RF is formed with the following procedure for minimizing correlations and bias between individual trees. For a P-element of the feature vector, p features are choosen at each phase of tree structure randomly, i.e., for each node of any RF tree [36]. The optimal split on these p attributes is applied to split the syllable data in the node. Classification of each syllable is conducted by simple voting of all RF trees.

2.3.5. Multi-layer perceptron (MLP)

MLP is commonly used in data pre-processing, classification application, data mining and speech recognition. It is a supervised algorithm, which defines the error value calculation by using backpropagation algorithm.

The MLP-BP architecture utilized in this study is composed of three layers; those are one input layer, one hidden layer and one output layer. The input layer represents the number of features of each feature generation technique. The optimal hidden neuron numbers in the hidden layer were selected manually. The output layer consists of twelve neurons, which presents the syllables classification result. The parameter of learning rate (η) was set to 0.3; the momentum (μ) was set to 0.2.

Furthermore, data verification was conducted in order to determine the reliability of the classification result. The verification method used in the test set was the holdout or k-fold cross-validation technique. The holdout estimated accuracy is shown in Eq. 7.

$$acc_h = \frac{1}{h} \sum_{\langle v_i, y_i \rangle \in \mathcal{D}_h} \delta(\mathcal{I}(\mathcal{D}_t, v_i), \mathcal{Y}_i \tag{7}$$

where $\delta(i, j)$ = 1 if $i = j$ and 0 otherwise [37].

3. Results and Discussion

Based on the research method, there are three main phase in this study. The first phase is pre-processing, which aims to collect the speech data. The second phase is feature extraction, which aims to convert the speech signal into a set of parameter or feature vector. The third phase is classification by using five different types of the classifier.

3.1. Feature extraction

Feature extraction phase aims to obtain the unique feature of the signal, which allows the recognition system to discriminate one syllable sound with the other from a different speaker. In this research, the features number generated by applying WRPSDS, RPSDS, PSDS, RE and WS were sixty-two, nineteen, twenty and twenty-nine, respectively.

After the feature extraction, the next phase is classification using five different classifiers (SMO, NB, RF, C4.5 and MLP) to obtain each average classification score of the classifier.

3.2. Classification

In this section, the classification results of each syllable by applying five different classifiers and feature generation methods are given in Tables 1 to 4. As a validation technique approach, the *k*-fold method was used. In many studies, the Ten-Fold Cross Validation (Ten-FCV) is often used as standard testing, but in this study, we also used Fifteen FCV as a comparison due to the number class of data is more than ten classes. The recognition score of each consonant or vowel context was computed from the two-dimensional confusion matrices obtained.

Table 1 shows comparative recognition rate of each syllable in following vowel context by using feature extraction methods of WS, RE, PSDS, RPSDS and WRPSDS combined with classification methods of SMO, NB, RF, C4.5 and MLP in Ten-FCV testing. Among five classifiers, it can be noticed that MLP has the best classification performance when it is combined with WRPSDS. The performance of the WRPSDS features shows the effectiveness and significance of utilizing both time and frequency domain features. The rank of classifiers is MLP, RF, SMO, NB and C4.5 as shown by the average recognition of 73.33%, 70.56%, 66.11%, 58.33% and 47.78%, respectively.

In the case of NB and RF classifier, the recognition rate of WS is better than WRPSDS as shown by the recognition score of 58.33% versus 55.56% and 70.56% versus 70.28%. It indicated that the high dimension of features does not always give a better performance in classification. Based on Table 1, it can be seen that WRPSDS features have the best average score in classification for the most of classifier. It indicates that higher dimension of feature gives a better average recognition score.

Table 2 shows the confusion matrix derived for the MLP classifier with Ten-FCV. In the confusion matrix, each row represents the classification rate refers to a syllable class, where each cell represents the syllable sound number of that class being classified into the class stated by the column label. The diagonal entries represent the number of syllables when the particular class of syllables is correctly classified. It can be seen from the table that MLP classifier is able to classify the 1st class until the 12th class with the average recognition rate of 73.3%.

Table 3 presents the percentage of classification results obtained utilizing Ten-FCV. For /*a*/, the highest result for the velar consonant */k/* was 86.7% by applying SMO classifier. For the velar consonant /*g*/, the classification score obtained is the lowest among other consonants, especially for C45. In WRPSDS, the highest score for consonant */l/* was 83.3% for SMO.

In the vowel */i/* case, the results for consonant */k/* for SMO, NB, RF, C45 and MLP classifier were 70%, 76.6%, 90%, 70% and 83.3%, respectively. This result indicated that the RF has the best result in accuracy but the result for consonant */l/* showed that MLP and SMO have the highest accuracy score.

In the vowel */u/* case, the average classification results for SMO, NB, RF, C45, and MLP were 72.5%, 61.65%, 72.5%, 49.16 and 79.98%, respectively. In this case, the MLP score was highest than the other classifiers.

Table 1. Average classification of syllables in the following vowel context.

Classification method	Accuracy in following vowel context			Average classification (%)
	/a/	*/i/*	*/u/*	
WS-SMO	56.67	46.67	70.83	58.06
RE-SMO	24.17	20.83	24.17	23.06
PSDS-SMO	46.67	35.83	40.83	41.11
RPSDS-SMO	41.67	45	37.5	41.39
WRPSDS-SMO	**63.33**	**62.5**	**72.5**	**66.11**
WS-NB	**55.83**	**51.67**	**67.7**	**58.33**
RE-NB	23.33	2	19.17	20.83
PSDS-NB	45.83	30.83	37.5	38.06
RPSDS-NB	35.83	39.17	25.83	33.61
WRPSDS-NB	55	50	61.67	55.56
WS-RF	60.83	**75**	**75.83**	**70.56**
RE-RF	39.17	23.33	34.17	32.22
PSDS-RF	51.67	55	45.83	50.83
RPSDS-RF	56.67	48.33	51.67	52.22
WRPSDS-RF	**64.17**	74.17	72.5	70.28
WS-C.45	47.5	49.17	44.17	46.94
RE-C.45	35	17.5	25.83	26.11
PSDS-C.45	**49.17**	33.33	30	37.5
RPSDS-C.45	45.83	32.5	32.5	36.94
WRPSDS-C.45	44.17	**50**	**49.17**	**47.78**
WS-MLP	63.33	60.83	75.83	66.67
RE-MLP	36.67	29.17	30	31.94
PSDS-MLP	55.83	47.5	48.33	50.56
RPSDS-MLP	51.67	55	51.67	52.78
WRPSDS-MLP	**70.83**	**69.17**	**79.98**	**73.33**

Table 2. Confusion matrix using MLP-WRPSDS.

Class	Identified class (number of samples)												Accuracy (%)
	1	2	3	4	5	6	7	8	9	10	11	12	
1	**25**	0	1	0	0	1	1	0	0	2	0	0	**83.3**
2	0	**25**	1	0	4	0	0	0	0	0	0	0	**83.3**
3	0	0	**25**	0	0	1	0	0	2	0	0	2	**83.3**
4	2	2	1	**16**	2	2	2	0	1	0	2	0	**53.3**
5	0	5	0	3	**18**	1	0	3	0	0	0	0	**60**
6	0	0	1	1	1	**25**	0	0	1	1	0	0	**83.3**
7	1	0	0	1	0	0	**23**	0	1	2	1	1	**76.7**
8	0	0	0	0	3	0	0	**23**	3	0	1	0	**76.7**
9	0	0	1	0	0	0	0	4	**25**	0	0	0	**83.3**
10	2	0	0	0	0	2	4	0	0	**21**	0	1	**70**
11	0	1	0	1	0	0	0	3	1	1	**17**	6	**56.7**
12	0	0	1	2	0	1	1	0	2	1	1	**21**	**70**
Total													**73.3**

Table 3. Classification result using WRPSDS features in Ten-Fold Cross Validation (TEN-FCV).

Classifier	Following vowels	/k/	/g/	/l/	/r/	Average % classification
SMO	/a/	86.7	33.3	83.3	50	63.33
	/i/	70	56.7	76.6	46.7	62.5
	/u/	83.3	66.7	80	60	72.5
Naive Bayes	/a/	70	46.7	60	43.3	55
	/i/	76.6	26.7	40	56.7	50
	/u/	63.3	53.3	53.3	76.7	61.67
Random Forest	/a/	83.3	43.3	66.7	63.3	64.15
	/i/	90	70	63.3	73.3	74.15
	/u/	86.7	73.3	70	60	72.5
C.4.5	/a/	56.7	20	50	50	44.18
	/i/	70	60	36.7	33.3	50
	/u/	60	53.3	36.7	46.7	49.16
MLP	/a/	83.3	53.3	76.7	70	70.83
	/i/	83.3	60	76.7	56.7	69.17
	/u/	83.3	83.3	83.3	70	79.98

Figure 2 depicts the percentage classification results of CV syllables in the context of its following vowel for five different classifiers. From the graph, it can be noticed that MLP has the best result for */a/* and */u/* vowels. C45 has the lowest score for all vowels. The rank of average recognition rate of each classifier was MLP, RF, SMO, NB and C45 respectively as shown by average classification rate of 73.33%, 70.28%, 66.11%, 55.56% and 47.78%, respectively.

Table 4 lists the percentage classification results utilizing Fifteen-FCV. The average classification scores in the vowel */a/* context, for SMO, NB, RF, C45 and MLP were 59.96%, 55%, 68.33%, 43.3% and 69.98%. In the vowel */i/* context, the average classification results for SMO, NB, RF, C45 and MLP were 63.35%, 48.35%, 68.33%, 50% and 68.3%. While for */u/*, the scores were 71.68%, 61.7%, 76.67%, 51.68% and 74.13%, respectively.

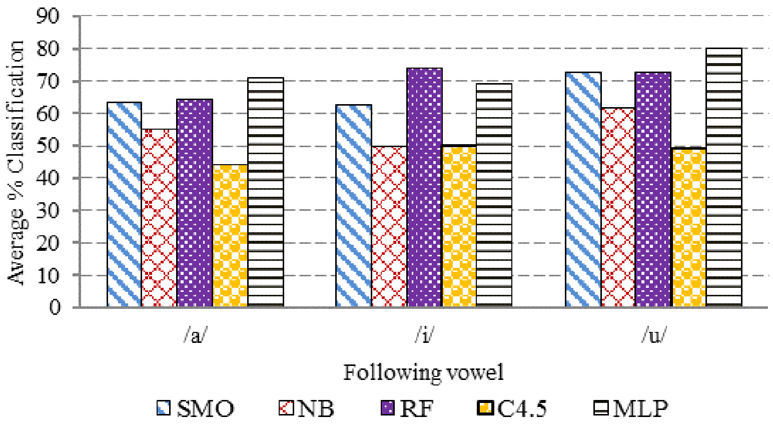

Fig. 2. Graph of the CV syllables classification in following vowel contexts for each classification method.

Table 4. Classification using WRPSDS in Fifteen-FCV.

Classifier	Following vowels	/k/	/g/	/l/	/r/	Average % classification
SMO	/a/	83.3	33.3	80	43.3	59.96
	/i/	66.7	60	70	56.7	63.35
	/u/	76.7	66.7	80	63.3	71.68
Naive Bayes	/a/	63.3	46.7	56.7	53.3	55
	/i/	73.3	26.7	36.7	56.7	48.35
	/u/	56.7	56.7	56.7	76.7	61.7
Random Forest	/a/	83.3	50	70	70	68.33
	/i/	83.3	60	66.7	63.3	68.33
	/u/	90	76.7	73.3	66.7	76.67
C.4.5	/a/	53.3	23.3	53.3	43.3	43.3
	/i/	63.3	66.7	43.3	26.7	50
	/u/	56.7	50	53.3	46.7	51.68
MLP	/a/	83.3	53.3	73.3	70	69.98
	/i/	76.6	63.3	73.3	60	68.3
	/u/	70	76.6	83.3	66.6	74.13

4. Conclusions

This paper presents syllables sound classification by utilizing time-frequency features as well as the performance comparison of five different classifiers namely a C45 decision tree, a Naive Bayes classifier, an SMO, a Random Forest decision tree, and a multi-layer perceptron. According to the result obtained, it can be concluded that the MLP classifier has the higher performance than other classifiers when it is combined with WRPSDS, as shown by accuracy of MLP, C.45, RF, NB, SMO, which are 73.3%, 47.78%, 70.28%, 55.56% and 66.11%, respectively. The performance of the WRPSDS features shows the effectiveness and significance of utilizing both time and frequency domain features. The further work suggested for this research is to utilize bigger syllable database with different ages and gender, applied to the other POA (i.e., dental, labial, etc.) and using a different technique of feature generation.

Nomenclatures

acc_h	Hold out estimated accuracy
a_k	AR’s coefficient
C_i	Class
$H_\alpha(X)$	Function of generalized entropy
$\underline{Ms}$	Heuristic “merit” of a feature subset S
$\overline{r_{cf}}$	Mean feature-class correlation (f ∈ S)
$\overline{r_{ff}}$	Average feature-feature inter-correlation
T	Period of sampling, s
x_{pp}	AR model in P order

Greek Symbols

σ_W^2	Variance of the drive noise input
φ	Scaling function of wavelet coefficient
$\psi(x)$	Wavelet function

Abbreviations

ANN	Artificial Neural Network
AR-PSD	Autoregressive Power Spectral Density
C45	Classification Tree 4.5
CFS	Correlation-Based Feature Selection
DB	Daubechie
DWT	Discrete Wavelet Transform
FCV	Fold Cross Validation
FS	Feature Set
MLP	Multi-Layer Perceptron
NB	Naive Bayes
PSDS	Power Spectral Density Statistical
RE	Renyi Entropy
RF	Random Forest
RPSDS	Renyi Power Spectral Density Statistical
SMO	Sequential Minimum Optimization
WPT	Wavelet Packet Transform
WRPSDS	Wavelet Renyi Power Spectral Density Statistical
WS	Wavelet Statistical

References

1. Davis, K.H.; Biddulph, R.; and Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of The Acoustical Society of America*, 24(6), 637-642.
2. Olson, H.F.; and Belar, H. (1956). Phonetic typewriter. *Journal Acoustic and Society America*, 28(6), 1072-1081.
3. Kristomo, D.; Hidayat, R.; and Soesanti, I. (2017). Classification of the syllables sound using wavelet, renyi entropy and AR-PSD features. *Proceedings of the IEEE 13th International Colloquium on Signal Processing & its Application (CSPA)*. Batu Feringghi, Penang, Malaysia, 94-99.
4. Kulkarni, P.; Kulkarni, S.; Mulange, S.; Dand, A.; and Cheeran, A.N. (2014). Speech recognition using wavelet packets, neural networks and support vector machines. *Proceedings of the International Conference on Signal Propagation and Computer Technology (ICSPCT)*. Ajmer, India, 451-455.
5. Kral, P. (2010). Discrete wavelet transform for automatic speaker recognition. *Proceedings of the 3rd International Congress on Image and Signal Processing*. Yantai, China, 3514-3518.
6. Zhao, X.; Wu, Z.; Xu, J.; Wang, K.; and Niu, J. (2011). Speech signal feature extraction based on wavelet transform. *Proceedings of the International Conference Intelligent Computation and Bio-Medical Instrumentation*. Wuhan, Hubei, China, 179-182.
7. El Ayadi, M.; Kamel, M.S.; and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Journal of Pattern Recognition*, 44(3), 572-587.

8. Abriyono; and Harjoko, A. (2012). Pengenalan ucapan suku kata bahasa lisan menggunakan ciri LPC, MFCC, dan JST. *Indonesian Journal Computing and Cybernetics Systems*, 6(2), 23-34.
9. Sekhar, C.C.; and Yegnanarayana, B. (2002). A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances. *IEEE Transactions on Speech and Audio Processing*, 10(7), 472-480.
10. Vuppala, A.K.; Rao, K.S.; and Chakrabarti, S. (2012). Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points. *Circuits, System and Signal Processing*, 31(4), 1459-1474.
11. Sakti, S.; Hutagaol, P.; Arman, A.A.; and Nakamura, S. (2004). Indonesian speech recognition for hearing and speaking impaired people. *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*. Jeju Island, Korea. 4 pages.
12. Kristomo, D.; Hidayat, R.; and Soesanti, I. (2016). Feature extraction and classification of the Indonesian syllables using discrete wavelet transform and statistical features. *Proceedings of the 2nd International Conference on Science and Technology-Computer (ICST)*. Yogyakarta, Indonesia, 88-92.
13. Arifin N.A; and Tiun, S. (2013). Predicting malay prominent syllable using support vector machine. *Procedia Technology*, 11, 861-869.
14. Anusuya, M.A.; and Katti, S. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, 6(3), 181-205.
15. Farooq, O.; and Datta, S. (2003). Phoneme recognition using wavelet based features. *Information Sciences*, 150(1-2), 5-15.
16. Trivedi, N.; Kumar, V.; Singh, S.; Ahuja, S.; and Chadha, R. (2011). Speech recognition by wavelet analysis. *International Journal of Computer Applications*, 15(8), 27-32.
17. Dede, G.; and Sazlı, M.H. (2010). Speech recognition with artificial neural networks. *Digital Signal Processing*, 20(3), 763-768.
18. Daqrouq, K.; and Al Azzawi, K.Y. (2013). Arabic vowels recognition based on wavelet average framing linear prediction coding and neural network. *Journal of Speech Communication*, 55(5), 641-652.
19. Hidayat, S.; Hidayat, R.; and Adji, T.B. (2015). Speech recognition of CV-patterned indonesian syllable using MFCC, wavelet and HMM. *Journal Ilmiah Kursor*, 8(2), 67-78.
20. Nehe, N.S.; and Holambe, R.S. (2012). DWT and LPC based feature extraction methods for isolated word recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 7 pages.
21. Endah, S.N.; Adhy, S.; and Sutikno, S. (2017). Comparison of feature extraction mel frequency cepstral coefficients and linear predictive coding in automatic speech recognition for Indonesian. *Telecommunication, Computing, Electronics and Control*, 15(1), 292-298.
22. Sharma, R.P.; Farooq, O.; and Khan, I. (2013). Wavelet based sub-band parameters for classification of unaspirated hindi stop consonants in initial position of CV syllables. *International Journal of Speech Technology*, 16(3), 323-332.

23. Farooq, O.; and Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Processing Letter*, 8(7), 196-198.
24. Panwar, M.; Sharma, R.P.; Khan, I.; and Farooq, O. (2011). Design of wavelet based features for recognition of hindi digits. *Proceedings of the International Conference Multimedia, Signal Processing Communication Technologies (IMPACT)*. Aligarh, India, 232-235.
25. Ranjan, S. (2010). A discrete wavelet transform based approach to hindi speech recognition. *Proceedings of the International Conference on Signal Acquisition and Processing*. Bangalore, India, 345-348.
26. Hardjono, F.L. (2011). *Stop consonant characteristics: VOT and voicing in American-born-Indonesian children's stop consonants*. Honor Thesis. The Speech and Hearing Science Department, Ohio State University, Columbus, Ohio, United States of America.
27. Adisasmito-Smith, N. (1998). Degemination in Indonesian phonology and phonetics. *Working Papers of the Cornell Phonetics Laboratory*.
28. Theodoridis, S; and Koutroumbas, K. (2009). *Pattern recognition* (4th ed.). Massachusetts, United States of America: Academic Press.
29. Renyi, A. (1961). On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, California, 547-561.
30. Kee, C.-Y.; Ponnambalam, S.G.; and Loo, C.-K. (2016). Binary and multi-class motor imagery using renyi entropy for feature extraction. *Neural Computing and Applications*, 28(8), 2051-2062.
31. Rizal, A.; Hidayat, R.; and Nugroho, H.A. (2016). Pulmonary crackle feature extraction using tsallis entropy for automatic lung sound classification. *Proceedings of the 1st International Conference on Biomedical Engineering (iBioMed 2016)*. Yogyakarta, Indonesia, 4 pages.
32. Hall, M.A. and Smith, L.A. (1998). Feature subset selection: A correlation based filter approach. *Proceedings of the on Neural Information Processing and Intelligent Information Systems. Progress in Connectionist-Based Information Systems*. Dunedin, New Zealand, 855-858.
33. Amarnath, B.; and Balamurugan, S.A.A. (2016). Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. *Journal of Engineering Science and Technology* (*JESTEC*), 11(11), 1639-1646.
34. Lu, S.; Xia, Y.; Cai, W.; Fulham, M.; and Feng, D.D. (2017). Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and FDG-PET imaging. *Computerized Medical Imaging Graphics*, 60, 35-41.
35. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
36. Wieczorkowska, A.A.; Kursa, M.B.; Kubera, E.; Rudnicki, R.; and Rudnicki, W.R. (2012). Playing in unison in the random forest. *Lecture Notes in Computer Science*, 7053, 226-239.
37. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence*. Montreal, Quebec, Canada, 1137-1143.