ANALYSIS AND VOICE RECOGNITION IN INDONESIAN LANGUAGE USING MFCC AND SVM METHOD

Harvianto[1]; Livia Ashianti[2]; Jupiter[3]; Suhandi Junaedi[4]

[1,2,3,4] Computer Science Department, School of Computer Science, Bina Nusantara University
Jl. K.H. Syahdan No. 9, Palmerah, Jakarta Barat, 11480
[1]harvianto@binus.ac.id [2]liviaashianti@gmail.com,
[3]jupiterc@gmail.com; [4]soe.xoe@gmail.com

ABSTRACT

Voice recognition technology is one of biometric technology. Sound is a unique part of the human being which made an individual can be easily distinguished one from another. Voice can also provide information such as gender, emotion, and identity of the speaker. This research will record human voices that pronounce digits between 0 and 9 with and without noise. Features of this sound recording will be extracted using Mel Frequency Cepstral Coefficient (MFCC). Mean, standard deviation, max, min, and the combination of them will be used to construct the feature vectors. This feature vectors then will be classified using Support Vector Machine (SVM). There will be two classification models. The first one is based on the speaker and the other one based on the digits pronounced. The classification model then will be validated by performing 10-fold cross-validation.The best average accuracy from two classification model is 91.83%. This result achieved using Mean + Standard deviation + Min + Max as features.

Keywords**:** *voice recognition, MFCC, SVM, cross validation*

INTRODUCTION

The human voice contains a lot of information such as gender, emotion, and identity of the speaker Lindasalwa *et al.* (2010). The purpose of the voice recognition is to identify the speaker or the words pronounced by the individual (Yee & Ahmad, 2008). Many techniques have been proposed to reduce the mismatch between testing and training environments. Most of these methods are operated in spectral domain (Lockwood & Boudy, 1992; Rosenberg, Lee, & Soong; 1994) or the cepstral domain.

Gracieth *et al.* (2014) implemented support vector machine (SVM) for automated speech digit recognition. The digit was limited to '0', '1', '2', '3', '4', '5', '6', '7', '8', '9' in Portuguese. The feature was extracted using Mel Frequency Cepstral Coefficients. Discrete Cosine Transform (DCT) was used to produce a two-dimensional matrix that became the input of the SVM. The study produced excellent numerical classification except for the digit '9'. Digit '1' to '8' had the best accuracy. The mean and variance were chosen as the features.

Fokoué and Ma (2013) had demonstrated that the combination of MFCC and SVM produces a great tool in identifying the sex of the speaker. RBF kernel and polynomial kernel give accurate results in cross-validation. MFCC needs more time in the calculation of computing because of the complexity of the calculations.

Putra and Resmawan (2011) wanted to classify gender base on the speech in Bahasa. The researcher is also using MFCC for extraction method and DTW for classification method. They collect

speeches of 27 men and eight women. These people will speak five words and repeat it seven times. For the evaluation, Darma and Adi used the 7-fold cross validation. Based on the result, the best accuracy is 93.254% and the worst accuracy is 59.664%.

This paper will discuss about the voice recognition of digit numeric '0', '1', '2', '3', '4', '5', '6', '7', '8', '9' in Indonesian. The human voice is converted into a digital signal form to produce digital data representing every level of the signal at each different time. Digital sound is then processed using the MFCC for extracting voice features. After that, Support Vector Machine (SVM) is used as classification method to determine the features and combinations of features that generate the most minimal error. The validation process will use 10-fold cross validation. This paper will be separated as follows: background research, the principle voice recognition, the methodology, which will be followed by the results, and conclusions are given.

After taking voice input using a microphone from the speaker, the sound will be analyzed. System design involves the manipulation of the audio signal. At some level, the operation is displayed on the input signal is pre-emphasis, framing, windowing, Mel Ceptrum analysis, and recognition of spoken words. Voice recognition algorithm includes two distinct phases. Figure 1 shows the voice algorithm. It can be shown that the first phase is the training phase while the second phase is the testing phase.

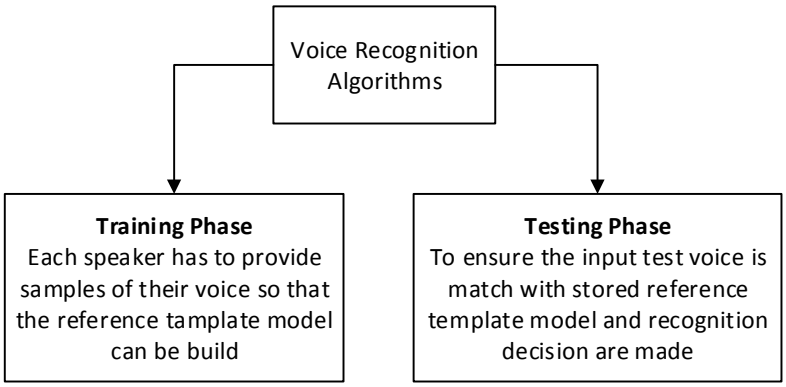

Figure 1 Voice Recognition Algorithms

Mel Frequency Cepstral Coefficients (MFCC) algorithm is a sampling technique. MFCC is one of the most popular feature extraction techniques used in voice recognition based on the frequency domain. MFCC using the Mel scale which is based on the human ear scale. MFCC which is being considered as frequency domain features, are much more accurate than time domain features. The simplicity and ease of the procedures used to implement the method MFCC make this the most favored technique for speech recognition.

MFCC considers the sensitivity of human perception of frequency and this makes the best in voice recognition. Figure 2 shows the following steps used in MFCC.

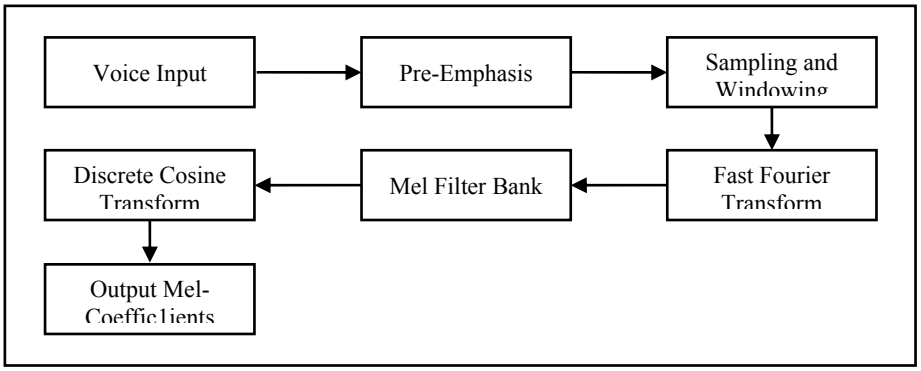

Figure 2 Block Diagram to Get the Coefficients MFCC

When feature extraction using MFCC in pre-emphasis block, voice signal is filtered with high pass filter. Pre-emphasis improves the voice signal and compensates the suppressed part of the signal during voice production. Then, the pre-emphasized signal is segmented into frames with an optional overlap of 1/3 until 1/2 of the frame size. This step is important to create good results because the variation of amplitude is more in larger signals compared to smaller signals. Then, framing signal will be multiplied with a Hamming window to the keep continuity of the first and last points in signal frame. Then, signal will be converted into frequency domain signal using Fast Fourier Transform. The output of Fast Fourier Transform block is multiplied by triangular band pass filters for getting log energies of each filter.

MFCC is defined as follows:

$$F_{\text{mel}} = \frac{c \log\left(1 + \frac{f}{c}\right)}{\log(2)} \tag{1}$$

F_{mel} is a logarithmic scale of normal frequency scale. Mel-cepstral features [2], can be illustrated by MFCCs, which is calculated from the Fast Fourier Transform (FFT) power coefficient. Power coefficient filtered by triangular bandpass filter bank. When c(5) is in the range of 250-350, the number of filters triangles that fall in the frequency range of 200-1200 Hz (the frequency range of audio information that is dominant) is higher than the other values of c. Therefore, it is efficient to set the value of c in the range to calculate MFCCs. The output is shown from the filter bank S_k (k = 1, 2, …., K), then MFCCs are calculated as follows:

$$\mathrm{C_n} = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_k) \cos\left[n(k - 0.5)\frac{\pi}{K}\right], n = 1, 2, \ldots, L \tag{2}$$

Support Vector Machine is a statistical machine learning techniques that are useful and successfully applied in pattern recognition. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory.

Data can be separated linearly. Data provided is denoted as $\overrightarrow{x_n} \in R^d$ whereas each class label is denoted $y_n \in \{+1,-1\}$ for $n = 1,2,\ldots,N$ where n is the number of data. SVM is looking for the best hyperplane that separates all data sets corresponding to the class by measuring margin hyperplane and looking for the biggest margin. Margin is the distance between the nearest hyperplane with the data from each class. The subset of the data set with the nearest distance is called a support vector.

Class -1 and +1 can be completely separated by hyperplane dimension d, which is defined by the following equation

$$\vec{w} \cdot \vec{x} + b = 0 \tag{3}$$

$\overrightarrow{x_n}$ which includes class -1 (negative samples) can be defined as data that meets inequality $\vec{w} \cdot \overrightarrow{x_n} + b \leq -1$ for $y_i = -1$. While $\overrightarrow{x_n}$ which includes class +1 (positive samples) can be defined as data that meets inequality $\vec{w} \cdot \overrightarrow{x_n} + b \geq 1$ for $y_i = 1$. $\vec{w}$ is normal field, and b is the position of the field about the origin. Value-defined margins is

$$M = \frac{2}{||w||} \tag{4}$$

Where

$$||w|| = \sqrt{\sum_{j \in N} w_j^2} \tag{5}$$

Maximum margin is obtained when the value of $||w||$ minimum of *hyperplane* equation is $\vec{w} \cdot \vec{x} + b = 0$. Therefore, to get the biggest margin, it can be formulated as a constrained optimization problem as follows

$$\min \frac{1}{2}||w||^2 \tag{6}$$

subject to $y_n(\vec{w} \cdot \overrightarrow{x_n} + b) - 1) \geq 0$.

One method for the settlement of constraint optimization problems is by multiplying Lagrange. Thus, it can be formulated as follows

$$\min \mathcal{L}_p(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} \alpha_n (y_n(\vec{w} \cdot \overrightarrow{x_n} + b) - 1) \tag{7}$$

subject to $\alpha_n \geq 0$.

Then the formula $\mathcal{L}_p$ (primal problem) was converted into the formula $\mathcal{L}_d$ (dual problem) as follows

$$\max \mathcal{L}_d(\alpha) = \sum_{i=0}^{n} \alpha_n - \frac{1}{2} \sum_{n=0}^{N} \sum_{m=0}^{N} \alpha_n \alpha_m (\overrightarrow{x_n} \cdot \overrightarrow{x_m})\, y_n\, y_m \tag{8}$$

With the above formula, then α_n is obtained with a positive value. The value of *w* then obtained by the formula as follows

$$w = \sum_{n=1}^{N} \alpha_n y_n \overrightarrow{x_n} \tag{9}$$

Data $\overrightarrow{x_n}$ in which the value α_n is more than zero is called a support vector. By knowing the support vector, the value of b can be obtained by using support vector obtained as follows

$$b = \frac{1}{N}\sum_{n=1}^{N}(y_n - \vec{w}.\overrightarrow{x_n}) \tag{10}$$

By recognizing the value *w* and *b*, the hyperplane equation (1) is obtained. After finding the hyperplane equation, the data classification into class $y_n \in \{+1, -1\}$ can be done as follows

$$f(\vec{x}) = \text{sgn}(\vec{w}.\vec{x} + b)\begin{cases} -1 & if\ \vec{w}.\vec{x} + b \leq -1 \\ 1 & if\ \vec{w}.\vec{x} + b \geq 1 \end{cases} \tag{11}$$

SVM formulation for linearly separable data cannot be used for non-linearly separable data. Searching the best hyperplane can be obtained by transforming data from input space ($\overrightarrow{x_n}$) into feature space ($\emptyset(\overrightarrow{x_n})$. Thus, the data can be separated linearly in the feature space.

Dimension data in the feature space is higher than dimension data in the input space. This situation can make a very large computation in feature space. These problems can be solved by used kernel trick. By using the kernel trick, transformation functions $\emptyset(\overrightarrow{x_n})$ does not need to be known. Kernel functions that are often used are Linear Kernel, Polinomial Kernel (dimension D), and Radial Basis Function (RBF) Kernel.

The equation of Linear Kernel is as follows

$$k(u, v) = u \cdot v(20) \tag{12}$$

Next, the following is the equation of Polinomial Kernel (dimension D)

$$k(u, v) = (1 + u \cdot v)^D(21) \tag{13}$$

The equation of Radial Basis Function (RBF) Kernel is

$$k(u, v) = \exp(-\gamma|u \cdot v|^2) \text{ where } \gamma > 0(22) \tag{14}$$

$$\gamma = -\frac{1}{2\sigma^2}\ (23)$$

Variable γ is *hyperparameter*.

Cross Validation is a method to assess the accuracy and validation of statistical models. The available dataset is divided into two parts. The first part is used to data modeling (Payam, Lei, & Huan, 2009). The data modeling from first part used to predict the values in the second part. A valid model should show good prediction accuracy.

The procedure of Cross Validation is as follows. First, the data will be divided into three sets; Training, Testing, and Validation

Training	Testing	Validation

Figure 3 First Step of Cross Validation

Second, find the optimal model on the set of training and test sets used to examine the predictive ability.

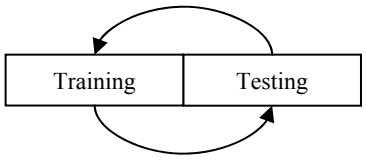

Figure 4 Second Step of Cross Validation

Third, see how well the model can predict the set of testing. Validation errors provide an estimate of the predictive power of the model.

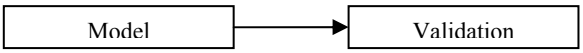

Figure 5 Third Step of Cross Validation

METHODS

This research will be conducted in several stages. To have a better understanding, see the following Figure 6.

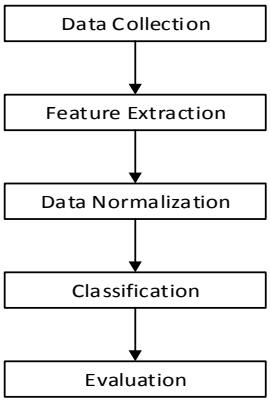

Figure 6 Diagram of Research Methods

Data is collected by recording the voices of 6 participants. Each participant will pronounce numbers between '0' to '9'. The recording process will be repeated up to 5 times; it consists of 3 times recording without noise and two times recording with noise. At the end, there will be 300 voice recording. The voice will be recorded using various devices such as smartphone, PC, and laptop with different types and specifications. The voice recording is saved with frequencies of 44.1 KHz, bit rate 16-bit and Wave Audio (WAV) file format. Each voice recording will be named by numbers pronounced, recording order, and participant name.

The voice recording feature will be extracted using MFCC (Mel-frequency cepstral coefficients). Using MFCC method, each voice recording produced a matrix where is MFCC index and is time frame. For the experiment, MFCC index will be used up to 12. Mean, standard deviation, min, and max of each MFCC index will be calculated. The result will be stored as matrix using format as shown in Figure 5.

Nama file	Label Angka	Label Speaker	$Mean_1$	....	$Mean_a$	Std_1	....	Std_a	Min_1	....	Min_a	Max_1	....	Max_a
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Figure 7 Feature Matrix Format

Support Vector Machine (SVM) with linear kernel function will be used as classification method. There will be two types of classifications to be performed. The first type is classification based on the speaker. The second type is classification based on numbers that the speaker has pronounced. Classification model will be built by using every feature or combination of features from feature matrix.

Classification result is evaluated by using 10 fold cross validation for every experiment.It is used to have better confidence in prediction accuracy. All of voice recording that has been collected will be separated into ten parts. Each part consisted 30 voice recording.

RESULTS AND DISCUSSIONS

For the experiment, there are 300 voice recordings that have been collected. Feature extraction will be applied to every voice recording. From this feature extraction, a new matrix will be obtained. The matrix results then will be classified using SVM method. There are two classification types to be performed. The first type is classification based on the speaker. The second type is classification based on numbers that the speaker has pronounced.

For each classification, 15 experiments have been performed by using every feature or combination of features from feature extraction. Detail of experiment is listed in Table 1. Each experiment is evaluated by using 10-fold cross-validation. The accuracy of each classification type and its features combination are shown in Table 2.

Table 1 Experiment

Experiment	Feature
1	Mean
2	Standard Deviation
3	Min
4	Max
5	Mean + Standard Deviation
6	Mean + Min
7	Mean + Max
8	Standard Deviation + Min
9	Standard Deviation + Max

Table 1 Experiment (continued)

Experiment	Feature
10	Min + Max
11	Mean + Standard Deviation + Min
12	Mean + Standard Deviation + Max
13	Mean + Min + Max
14	Standard Deviation + Min + Max
15	Mean + Standard Deviation + Min + Max

Table 2 Experiment Result

Experiment	Feature	Accuracy (%)		
		Speaker	Pronounced Number	Average
1	Mean	59.00	90.33	74.67
2	Standard deviation	72.67	60.67	66.67
3	Min	58.33	85.33	71.83
4	Max	52.67	67.67	60.17
5	Mean + Standard deviation	86.67	95.67	91.17
6	Mean + Min	69.33	95.33	82.33
7	Mean + Max	73.00	92.67	82.83
8	Standard deviation + Min	80.00	92.67	86.33
9	Standard deviation + Max	82.33	89.67	86.00
10	Min + Max	74.67	92.00	83.33
11	Mean + Standard deviation + Min	87.00	96.00	91.50
12	Mean + Standard deviation + Max	85.33	96.00	90.67
13	Mean + Min + Max	79.00	95.33	87.17
14	Standard deviation + Min + Max	82.33	95.67	89.00
15	Mean + Standard deviation + Min + Max	86.67	97.00	91.83

The best accuracy for classification based on the speaker is 87.00%. This result is achieved using Mean + Standard deviation + Min as features. The worst accuracy for classification based on the speaker is 52.67%. This result is achieved using Max as features. The best accuracy for classification based on the number pronounced is 97.00%. This result is achieved using Mean + Standard deviation + Min + Max as features. The worst accuracy for classification based on the number pronounced is 60.67%. This result achieved using Standard deviation as features. The best average accuracy from both classifications is 91.83%. This result is achieved using Mean + Standard deviation + Min + Max as features. The worst average accuracy from both classifications is 60.17%. This result is achieved using Max as features.

CONCLUSIONS

Experimental results showed interesting results, feature or combination of features which have the highest accuracy in classification based on the numbers spoken is Mean + Standard Deviation + Min (87%). Feature or combination of features which have the highest accuracy in classification based on the speaker is Mean + Standard Deviation + Min + Max (97%). The best result is obtained by using combination of Mean + Standard Deviation + Min + Max.

REFERENCES

Fokoue, E., & Ma, Z. (2013). Speaker Gender Recognition via MFCCs and SVMs. *RIT Scholar Works*.

Gracieth, B., Washington, S., & Filho, O. (2014). Classification of Pattern using Support Vector Machines: An Application for Automatic. *The Eighth International Conference on Advanced Engineering Computing and Applications in Sciences.* Rome, Italy: IARIA.

Lindasalwa, M., Mumtaj, B., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing, 2*, 138-143.

Lockwood, P., & Boudy, J. (1992). Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. *Journal Speech Communication - Eurospeech '91, 11*(2-3), 215-228.

Payam, R., Lei, T., & Huan, L. (2009). Cross-Validation. *In Encyclopedia of Database Systems*.

Putra, D., & Resmawan, A. (2011). Verifikasi Biometrika Suara Menggunakan Metode MFCC dan DTW. *Lontar Komputer, 2*, 8-21.

Rosenberg, A., Lee, C. H., & Soong, F. (1994). Cepstral channel normalization techniques for HMM-based speaker verification. *Proc. Int. Conf. on Spoken Language Processing*, 1835-1838.

Yee, C. S., & Ahmad, A. M. (2008). Malay language text-independent speaker verification using NN-MLP classifier with MFCC. *Electronic Design, 2008. ICED 2008. International Conference.* Penang: IEEE.