# Voice Recognition using k Nearest Neighbor and Double Distance Method

Ranny
Department of Computer Science
University of Multimedia Nusantara
Indonesia
ranny@umn.ac.id

***Abstract*— Voice recognition process is started with voice feature extraction using Mel Frequency Cepstrum Coefficient (MFCC). The purpose of the MFCC method is to get the signal feature that correlate to the human voice. The converted signal from analog to digital is needed in the MFCC method. The digital signal has a time domain and it make the analysis harder. So, the domain time is converted to time domain for make the analysis more accurate. Furthermore, after get the feature, the recognition step is using k Nearest Neighbor (kNN) method with k number is one. Euclidean Distance is used to get the similarity of the data training and data testing. The previous research shows that kNN has a high accuracy if use normal data, but it has lower accurate when using outlier data. Base on this problem, this research develop a new method to handle the outlier data using kNN and double distance measurement. The double distance method is note each distance of each data to the center of the kNN data. The calculation of the distance is used on recognition step. The accuracy of the method is tested on experiment. The experiment is using 11 subjects as data training and data testing. Each voice of subject is recorded three times. The result of the experiment with kNN method with one data center is 84.85% and the experiment result using double distance measurement is 96.97%. The result shows that the double distance method increase the accuracy of voice recognition.**

***Keywords—voice recognition; Mel Frequency Cepstrum Coefficients, k Nearest Neighbor; Euclidean Distance***

## I. INTRODUCTION

The voice recognition is one of the systems that developed based on pattern recognition. The process of voice recognition on computer science is develop using many kind of method, for example is Dynamic Time Wrapping (DTW), Linear Vector Quantization (LVQ), Artificial Neural Network (ANN) and etc.[3][4][7]. Each method has advantages and disadvantages based on the time process and accuracy rate of recognition. This research discuss is about voice recognition using kNN and double distance measurement. This method is compared using data training and data testing to get the accuracy.

The frame work of the voice recognition consists of two steps are training step and testing step. The training step is voice extraction feature using Mel Frequency Cepstrum Coefficients (MFCC). The MFCC method is convert the domain of signal from time domain to frequency domain. The time domain signal is more difficult to processed and analyzed due to the number of data and the complexity of the data. The frequency domain signal is simple to be analyzed because the pattern of the signal is get from the data. Besides, the MFCC method also can extract the feature from the whole data voice and it represents the voice of each subject. There are many kind of MFCC based on the number of the filter that use to get the human signal pattern [2]. The MFCC also has different type based on the number of the coefficient result [2]. This research using the 24 number of triangle filter and the number of the coefficients is 13 for each frame on voice data.

The next step is testing process. The testing step based on the previous research that using kNN method as the recognition method. The determination number of k and method of measuring distance is common topics research. The topic is concern on the method but cannot handle for outlier data. The purpose of the research is to develop a new method that can increase the accuracy using outlier data. The develop method is note the distance for each training data to the center of the center of the data.

The expansion method with record the distance can keep the correlation between data on each class, especially for outlier data. The developing method in this research is called double distance method. The purpose of testing step is to do the recognition process using kNN method, with k=1 that represented the center of class using the average of the data. The kNN method is combined by the doubles distance method. The using of simple algorithm is to get the accuracy of the method

The accuracy of two methods is compared using the training data and non-training data. The framework of the research is showed on Figure 1. The diagram shows that the system is started with recording the audio input. The next step is training step, the training step do the extraction feature. The method of the extraction feature is using MFCC. The feature of audio input is saved as database and used on testing step. The testing is divided into two parts, the first part using kNN method, the second using double distance method.

The experiment of the research is using voice data that recorded by 11 subjects, each subject pronounce a word of ‘computer’ in Bahasa. The voice data is recorded three times by each subject. The condition of the recorded process is on minimum noise. The data is divided into two groups, the first

978-1-5090-1671-6/16/$31.00 ©2016 IEEE

group is used for training step and the second group is used for testing step. The result of the experiment is used for analyze the accuracy of the method.

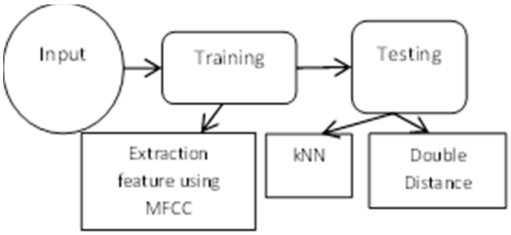


Fig 1. Framework of Voice Recognition

## II. METHODOLOGY

### A. Mel Frequency Cepstrum Coefficients

Human voice is captured as analog signal, to process the analog signal is needed to convert the digital signal with various level of sampling rate, 14400Hz, 16000Hz, and 8000Hz [2][7]. The level of the sampling rate is influence the shape of the digital signal. The digital signal is saved on various formats as wav, mp3, mp4, etc. The difference of the format is on the compressing format data. If the format has higher compressing data it makes quality of the voice is lower.

The MFCC method is one of the extraction voice feature aims to get the voice feature based on the discrete signal [2][3][6]. The signal discrete based on the time was changed to be the frequency domain is aim to make the analyses process easier. The MFCC method is develop based on the psychophysical study which states that human voice is on linearly [2]. This state is used for do the filtering with scale of mel. Mel scale is linear on frequency with the number is less than 1000 Hz and logarithmic on more than 1000 Hz [2][4]. The next step is reconvert the log mel spectrum to time spectrum using discrete cosine transform (DCT) and the result is called as mel frequency cepstrum coefficients. The result of the MFCC method is voice feature which became the input on next process. Figure 2 show the step on MFCC process.

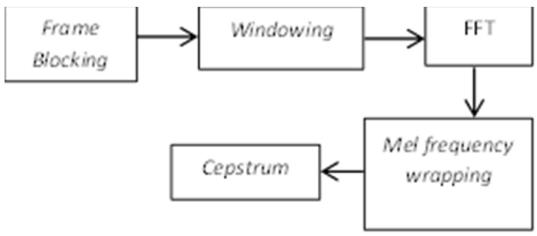


Fig 2 Step of MFCC

The purpose of frame blocking process is to do the segmentation of voice signal [2][4]. The pronunciation of each subject has difference speed, so the segmentation process makes the long of voice data is more consistent. The result of the segmentation process is an overlapping data. The overlapping data is to make the data on each frame is to keep the continuity data.

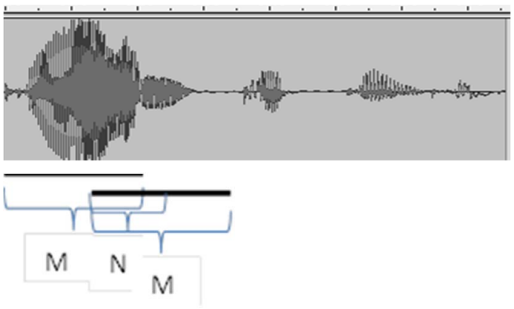


Fig 3 Frame blocking

The Fig 3 showing the process of overlapping data on frame blocking step. The size of each frame is M and the long of overlapping data is N. The purpose of windowing step is to keep the data on each frame in the same range of sampling number. The common method of windowing is Hamming window [2][6]. The next step on MFCC is Fast Fourier Transform (FFT); this step is to change the time domain to the frequency domain. The changing domain make the data easier to be analyzed [2][4][6].

The sampling data is still on random range of frequency. The mel frequency wrapping is to get the frequency of human voice, so it make the data clear or no noise. The mel frequency wrapping is also to get the feature of sampling data [2][4]. The last step of MFCC is cepstrum. The domain of sampling data is changed back to the time domain on cepstrum step. The result of cesptrum is used to be the MFCC coefficients and declare as voice feature extraction.

The size of feature extraction data is n x m, which n is a number of desired data (that is inputted on cepstrum step), while m is the size of data that got from frame blocking when data changed from analog to digital. The result of MFCC became the input on the training and recognition step.

### B. k Nearest Neighbor

The kNN method is a method based on supervised learning. The measurement of similarity distance between the data is the based method of kNN on data classification.

Here is the algorithm of kNN [5][10]:

1-NN Algorithm:

1. Calculate the distance between testing data to each training data.

2. Determine one data label that has the most minimum distance.

3. Classified the testing data to the label data (based on the number 2 step).

k-NN Algorithm:

1. Determine k value

2. Calculate the distance of each training data to the data label.

3. Determine the data label that has the most minimum distance

4. Classified the training data to the data label on 3.

5. Repeat the step 2 to 4 until the number of each class is k.

Based on the both algorithm, the research use 1-NN Algorithm due to this algorithm suitable for the research problem that need only one single input data to be recognized. The Euclidean Distance [9] is used in the algorithm.

## III. IMPROVEMENT AND EXPERIMENT

### A. Purposed Method

The voice recognition system needs a large number of training data to improve the accuracy level. Commonly, the 1-NN Algorithm calculates the average of data training and use the average value to represent the class or the label. The result of recognition is got from the shortest distance between the testing and the average value of the class. This make the process of recognition spend more time hence it need a method with high speed and good accuracy.

The voice recognition system also needs a method that can handle the outlier data testing to increase the accuracy. The 1-NN is not strong enough to handle it. Here is the simulation of the outlier data. The simulation is using two data series as the training data on database. As can be seen on the Fig 4, one of the members of Series1 is an outlier and the outlier value is close to the Series2 area. This phenomena makes the system has a lower accuracy. The purpose of this research is to handle the outlier data by makes the distance of the outlier data is getting closer to the correct class.

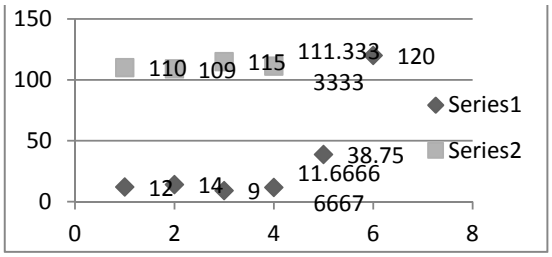


Fig 4 Simulation of outlier data

The main idea of the method is modify the calculation of Euclidean distance. The variable that notes the average of the distance value between each data on the class is added to the Euclidean distance. This innovation is explained on the simulation below:

Training data $X_1$and $X_2$ set

$$X_1 = \{x_1^1, x_1^2, x_1^3 x_1^4\}$$

$$X_2 = \{x_2^1, x_2^2, x_2^3 x_2^4\}$$

$\bar{X} = \{\bar{x}^1, \bar{x}^2, \bar{x}^3, \bar{x}^4\}$

$\bar{X}$ is the average of the training data X.

$$p_X^1 = \sqrt{(\bar{x}^1 - x_1^1)^2 + \cdots + (\bar{x}^4 - x_1^4)^2}$$

Training data $Y_1$ and $Y_2$ set

$$Y_1 = \{y_1^1, y_1^2, y_1^3 y_1^4\}$$

$$Y_2 = \{y_2^1, y_2^2, y_2^3 y_2^4\}$$

$\bar{Y} = \{\bar{y}^1, \bar{y}^2, \bar{y}^3, \bar{y}^4\}$

$\bar{Y}$ is the average of the training data Y

$$p_y^1 = \sqrt{(\bar{y}^1 - y_1^1)^2 + \cdots + (\bar{y}^4 - y_1^4)^2}$$

Calculate the value of p for each training data, so we get:

$$P_X = \{p_X^1\ p_X^2\ p_X^3\ p_X^4\} \text{ and } P_Y = \{p_Y^1\ p_Y^2\ p_Y^3\ p_Y^4\}$$

So, if we have testing data Z with voice features:

$$Z_1 = \{z_1^1, z_1^2, z_1^3, z_1^4\}$$

Then here is the formula to calculate the similarity distance:

$$D_{Z-X} = \sqrt{p_X^1(\bar{x}^1 - z_1^1)^2 + \cdots + p_X^4(\bar{x}^4 - z_1^4)^2}$$

and

$$D_{Z-Y} = \sqrt{p_Y^1(\bar{y}^1 - z_1^1)^2 + \cdots + p_Y^4(\bar{y}^4 - z_1^4)^2}$$

So, the result of the recognition is $min(D_{Z-X}, D_{Z-Y})$

### B. Experiment

The purpose of the experiment is to get the accuracy of the kNN and double distance method. The data of the experiment is got from 11 (eleven) voice recording subject. Each subject voice is recorded three times by saying "computer". The process of the recording is on quite condition and free from noise. The speed of pronunciation is constant for each subject and consistent without intonation variation. The sampling rate that used in the experiment is 8000Hz. The Audacity Program is used to record the voice data and saved on .wav format. The Fig 5 shows the frame rate that been used is 8000Hz with mono condition.

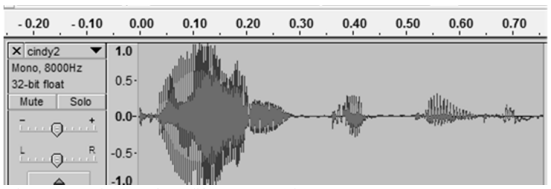


Fig 5 Experiment data file on .wav format

The experiment consist of two parts; the first part is experiment using kNN method with k = 1. The testing voice data is compared to the average of each subject's voice

feature. The minimum distance is known as the recognition result.

The double distance method is used on the second experiment. The first experiment data is also used on the second experiment. The recognition result is got by the minimum distance.

The total number of positive result of recognition is divided by the total number of experiment and multiplied by the 100 is used as the accuracy percentage. The both of the experiment are compared to get the conclusion of the research.

## IV. Result and Discussion

The experiment that using the kNN method with k = 1 is calculate the similarity between the testing data to the center of each label data. The center data is the average of each subject' feature voice. The testing is compiled three times for each subject. So, we got 33 numbers of experiment result. The result of testing is showed on the Table 1. Based on testing result we got 28 positive recognitions. The A, B, and K subject are the successfully recognized by the system; otherwise the G subject only once can be recognized. The C, D, E, F, G, I and J are three times successfully recognized for each subject. Based on the testing, the accuracy level is 84.85% that get from 28 divide to 33 and multiplied by 100%.

Table 1 Testing result with kNN average method

| No | Subject | Total of positive result | Minimum Euclidean distance |
|---|---|---|---|
| 1 | A | 2 | 15.9144 |
| 2 | B | 2 | 28.9086 |
| 3 | C | 3 | 20.5702 |
| 4 | D | 3 | 20.7815 |
| 5 | E | 3 | 33.5138 |
| 6 | F | 3 | 15.3762 |
| 7 | G | 1 | 21.863 |
| 8 | H | 3 | 37.4621 |
| 9 | I | 3 | 22.7772 |
| 10 | J | 3 | 22.7772 |
| 11 | K | 2 | 31.685 |

The testing of double distance method is on the second experiment using the data on the first experiment. The result of the second testing is showed on the Table 2. The total of positive result is 32 voices. Only the B subject has been recognized less than three times, it is only two times. So, the accuracy of the second experiment is 96.97%.

Table 2 Testing result with double distance method

| No | Subject | Total of positive result | Minimum Euclidean distance |
|---|---|---|---|
| 1 | A | 3 | 8.265 |
| 2 | B | 2 | 0 |
| 3 | C | 3 | 11.0303 |
| 4 | D | 3 | 14.3835 |
| 5 | E | 3 | 19.2088 |
| 6 | F | 3 | 9.0439 |
| 7 | G | 3 | 8.2326 |
| 8 | H | 3 | 21.8908 |
| 9 | I | 3 | 11.1467 |
| 10 | J | 3 | 11.7415 |
| 11 | K | 3 | 16.2913 |

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## V. Conclusion and Future works

This paper has described the improvement of kNN method to handle the outlier data by using the double distance method. The testing experiment show that the double distance method is improved the recognize accuracy. The accuracy of the double distance is higher than the average kNN method, especially for outlier data. The difference type of data can be the future work. The data can be on image or other format. Besides, the comparison of the double distance method and other machine learning method such as Hidden Markov Model, Neural Network, Linear Predictive Code, etc. can be the next research topic.


## Acknowledgment

This research is supported by the University of Multimedia Nusantara (www.umn.ac.id). The publication of this paper is supported by Ministry of Research, Technology and Higher Education of Indonesia. The writer would like to say thank you to Pattern Recognition Laboratory of University of Tarumanagara that support the data collection process. Thanks to the colleague of Machine Learning 2011 class in Computer Science, University of Indonesia that also support the idea and experiment.


## References


[1] Fomby, T. "K-Nearest Neighbors Algorithm: Prediction and Classification." (2008).

[2] Ganchev, T.D. Speaker Recognition. Patras, 2005.

[3] Gilke, Mandar, Rohit Kothalikar and Varum Pius Rodrigues. "MFCC-based Vocal Emotion Recognition Using ANN." Singapore: IACSIT Press, 2012.

[4] H. Hermansky and N. Morgan. "RASTA Processing of Speech." IEEE Trans. on Speech and Audio Processing. 2(1994): 578-589.

[5] Kaghyan, Sahak and Hakob Sarukhanyan. "Activity Recognition Using K-Nearest Neighbor Algorithm On Smartphone With Tri-Axial Accelerometer." 1 (2012).

[6] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, "Speaker identification using mel frequency cepstral coefficients" ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.

[7] Muda, Lindasalwa, Mumtaj Begam and I Elamvazuthi. "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques." JOURNAL OF COMPUTING (2010): 138.

[8] Shrawankar, Urmila and Vilas Thakare. "A Hybrid Method For Automatic Speech Recognition Performance Improvement In Real World Noisy Environment." Journal of Computer Science, (2013): 94-104.

[9] Thakur, Akanksha Singh and Namrata Sahayam. "Speech Recognition Using Euclidean Distance." 3.3 (2013).

[10] Tenkomo,K. "K-Nearest Neighbor Tutorial." http://people.revoledu.com/kardi/tutorial/KNN/index.html. (2006).