Comparison Euclidean Distance and Manhattan Distance as Classification in Speech Recognition System

Muhammad Ryandy Ghonim Asgar(✉), Risanuri Hidayat, and Agus Bejo

Gadjah Mada University Yogyakarta, Yogyakarta, Indonesia
muhammadryandy@mail.ugm.ac.id

Abstract. One of the uses of a digital system is a speech recognition system. Feature extraction and classification is important step in speech recognition system process. Mel Frequency Cepstrum Coefficient (MFCC) feature extraction is a popular feature extraction used in speech recognition system, while one of the most popular classification technique is K Nearest Neighbour (KNN). There are many KNN classification techniques, but the most commonly used are the Euclidean Distance and Manhattan Distance. Research on speech recognition system in Indonesia and in particular the Indonesian speech recognition system is still very limited, far from the recognition system in English. Therefore, this paper proposes a comparison of the best accuracy generated by the classification between Euclidean distance and Manhattan distance using MFCC as a feature extraction in Indonesian speech recognition system. The model and testing of the proposed system used is 120 data, with 0 to 9 voice signals in Indonesian. By using the 13 coefficients from the MFCC and using 5-fold cross validation to achieve generalized results, the Euclidean distance is able to outperform the accuracy obtained by the Manhattan distance by a value of 88%.

Keywords: Speech Recognition · MFCC · Euclidean Distance · Manhattan Distance · K-Fold Cross Validation

1 Introduction

Technology is something that cannot be avoided in the progress of life, because technological advances will go hand in hand with advances in science. Technology was created to provide positive benefits for human life, providing many conveniences and as a new way of carrying out human activities. Currently, society has enjoyed many benefits that are the impact of the resulting technological innovations. In the past, the world of technology only knew analog system, now digital system have emerged which are expected to make user's work easier and more comfortable. One of the uses of a digital system is a speech recognition system. Speech recognition system is a method for humans and machines or technological tools to be able to establish communication with each other, or in the other words is the process of machine based identification of word and sentences [1].

B. B. Wiyono et al. (Eds.): ICEMT 2022, ASSEHR 727, pp. 454–463, 2023.
https://doi.org/10.2991/978-2-494069-95-4_54

Research on speech recognition began in the 1950s [2]. Although a lot of research has been done on the development of speech recognition system, until now there is still not much use of Indonesian voice data used in research. Research on speech recognition in Indonesia and especially the Indonesiaan speech recognition system is still very limited, far from the recognition system in English. Accuracy is an important point in speech recognition systems. Word size, speaker dependence independent speaker, recognition time, speech type, and recognition environmental conditions, is the things that have been mentioned are parameters that can affect the accuracy of speech recognition systems [3]. In addition, the speech recognition system has an important step in the process, namely feature extraction. Where feature extraction aims to analyze the speech signal and break it down into certain characteristics. There are many methods for feature extraction, therefore it is necessary to select the right feature extraction method for each type of speech signal. Mel Frequency Cepstrum Coefficients (MFCC), Linear Predictive Coding (LPC), and Linear Predictive Cepstral Coefficients (LPCC) are the most widely used methods in speech signcal research [4]. So that the feature extraction has been tested and suitable for many types of speech signals. Beside feature extraction, another important step is classification. Classification aims to identify sound patterns or features that have been processed by feature extraction [5] and previously unknown class labels, which can be predicted using classification technique. One of the most common classification technique is K-Nearest Neighbor (KNN) [6]. Classification commonly used in KNN is a function of distance, such as Euclidean distance, Manhattan distance, Minowski distance, and others [7].

In a study, Punam Mulak et al. [6] conducted a study by analyzing and comparing distances from Euclidean, Chebychev and Manhattan using the KNN classification. The comparisons made in this study are accuracy, sensitivity, and specificity. Manhattan distance provides high performance between euclidean and chebychev using the KDD dataset. Another study, Ranny [8] conducted a speech recognition system using KNN parameters as a classification with Euclidean distance. In this study, Ranny used the alphabet A to K as dataset, and the total dataset was 11. Ranny got a high accuracy score in this study using the double distance method technique. It did not stop there, research on the comparison of classifications between Euclidean, Manhattan, and Chebychev was again carried out by [9]. In this study Euclidean became the best classification between Manhattan and Chebychev.

Research on speech recognition system in Indonesia and in particular the Indonesian speech recognition system is still very limited. Therefore, this paper proposes a comparison of the best accuracy generated by the classification between Euclidean distance and Manhattan distance using MFCC as a feature extraction in Indonesian speech recognition system.

The next section on this paper is organized as follows: in Sect. 2 discusses the methods to be used in this research, Sect. 3 is the result discussion, and the last Sect. 4 is the conclusion.

2 Methods

In this study, a four-step method was proposed. The first is the input of speech signal, second is preprocessing, third is feature extraction process, and the last step is classification (Fig. 1).

In this study, there were 6 male and 6 female speakers, and the total number of speakers is 12 people, were the speakers had an age range of 20–35 years. Each speaker speaks a number from 0 to 9 in Indonesian. And the total dataset used in this study amounted to 120 voice datasets 0 to 9 in Indonesian with.wav format.

In this research, preprocessing to process data from the input signal. In this preprocessing the process of cutting the silent signal. This reduction is intended so that the signal which is not a core input signal does not interfere with the recognition system, therefore the signal must be removed. In this study the signal was uniformly cut with a signal length of 2048 samples or about 256 ms.

Feature extraction is the process of converting sound waves into several types of parametric representations that can be processed, there are several methods of representing further, one of which is the Mel Frequency Cepstrum Coefficient (MFCC) method. MFCC is a coefficient the represents audio, this method was introduced by Davis and Marmelstin in the 1980s [2]. The following in Fig. 2. Are the stages of the process carried out by the MFCC extraction feature, such as preemphasis, framing, windowing, Fast Fourier Transform (FFT) Mel Filter Bank, and Discrete Fourier Transform (DCT).

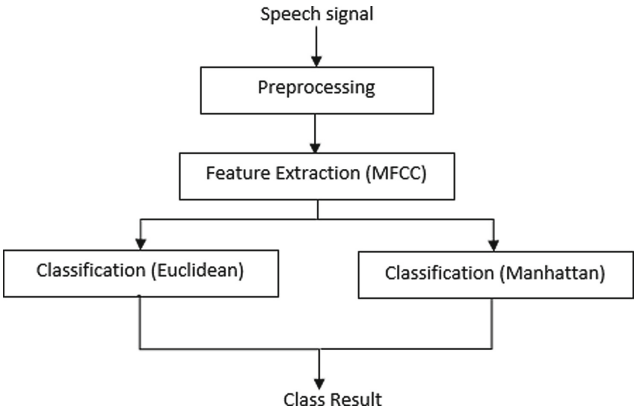

Fig. 1. Proposed Method

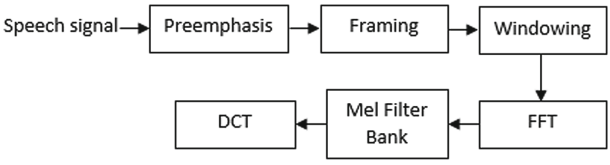

Fig. 2. MFCC Feature Extraction

The first step is preemphasis. In this step, the signal is filtered using a first-order FIR filter to even out the spectral signal, or in the words, pre-emphasis is used to increase the sound at high frequencies, because there is more energy at lower frequencies. The next step is framming. The sound signal is stationary in a short time span, therefore the sound must be cut into short time spans, this is known as short time analysis. The signal is trimed to equal section called frames. The normal duration is 10–30 miliseconds. The framing process will cause the sound signal to be cut off between one frame and another, so this is the task of windowing to reduce noise that appears in that frame, by applying Hamming window to the sound signal [3]. The following Eq. (1) is a Hamming equation, while in Eq. (2) is the output of each frame after the filtering process.

$$W[n] = 0.54 - 0.4\cos[\frac{2\pi n}{N-1}] \tag{1}$$

$$Y[n] = X[n] \times W[n] \tag{2}$$

With N is the number of samples per frame, W[n] is the nth coefficient of the Hamming window, and Y[n] is the output signal [3]. Then convert the signal from the time domain to the frequency domain using the fast Discrete Fourier Transform (DFT) algorithm [10], or commonly known as the Fast Fourier Transform (FFT) process. Then the Mel Filterbank process is carried out where in this process the bandpass filters overlap each other with Mel, linear, and logarithmic scales below frequency of 1 kHz [3]. The form of the mathematical equation of the mel scale can be seen in Eq. (3).

$$mel = 2595log_{10}(1 + \frac{f}{100}) \tag{3}$$

In Eq. (3) m is the output of the filterbank and f is the input value of the filterbank and mel is the output of the filterbank. The result that will be obtained at this step is the number of mel filter bank. The mel filter bank value shows how much energy in the frequency range there is in each mel filter. Many studies [11, 12] uses the values 2595 and 700 which are fixed values for some studies. Filterbank ilustrated can be seen in Fig. 3.

The final step is to use Discrete Cosine Transform (DCT). DCT works to return the sound signal in the frequency domain back to the time domain so that the cepstrum coefficient is obtained. The mathematical equation of DCT can be seen in Eq. (4).

$$\sum_{k=1}^{N} \log(Y(i)) \times \cos[mx(k-0.5)x\pi \div N] \tag{4}$$

The classification process aims to assess a data or object and enter it into a certain class from several existing classes. The purpose of the classification process is to be able to recognize a data into a certain class. To determine the performance in a classification, generally required performance measurement, classification performance measurement is done with a confusion matrix.

In this study, use the K-Nearest Neighbour (KNN) for the classification process. The K-Nearest Neighbour algorithm that uses a supervised algorithm that classifies the results

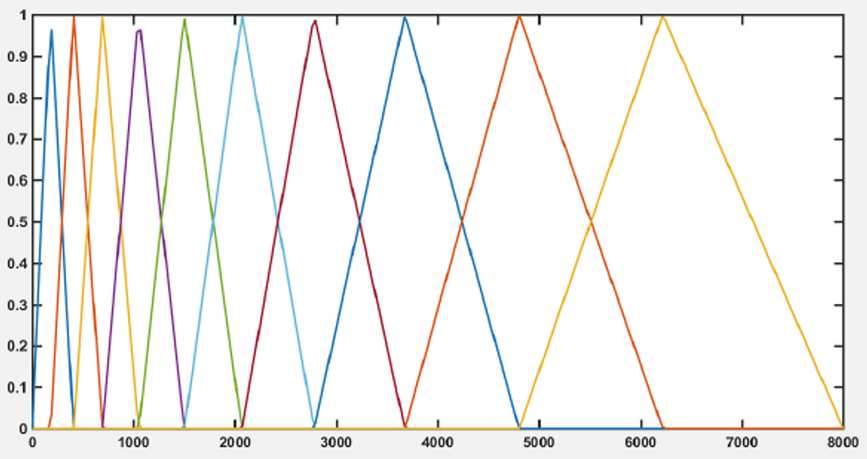

Fig. 3. Mel Spaced Filterbank

of query instances based on most of the KNN categories [7]. Classification uses the most votes in the classification of k objects. The KNN algorithm uses adjacent classifications as predictors for new query instances.

The optimal value of k in the KNN algorithm depends on the data. In general, increasing the value of k reduces the effect of noise on the classifications, but blurs the boundaries between each classification. The KNN algorithm based on the shortest distance from the test data to the training data to determine the KNN value. After determining the similarity value, then the value is grouped into certain classes.

KNN works requires input data in the form of training data, test data and the value of k. Then sort the distance training data based on the distance calculation of the test data with the training data. After that it is taken from the top k training data to determine the dominant class classification class from the k training data. The distance measurement method used in this study is as follows:

Euclidean distance is a distance calculation methods used to measure the distance from two points in euclidean space that covers two or more dimensions.The mathematical form of euclidean distance can be seen in the Eq. (5).

$$D(a, b) = \sqrt{\sum_{k=1}^{d} (a_k - b_k)^2} \tag{5}$$

In the Eq. (5), where D is the number of dimensions of each data, with a and b being two data to be searched for the distance. If the number of results from the formula is large, then the level of similarity between training data and testing data will be not accurate, an the otherwise if the number of results from the formula is small, then the level of similarity between training data and testing data will be accurated.

Manhattan distance is a distance to calculate the absolute difference between the coordinates of a pair of objects. The mathematical form of manhattan distance can be

seen in Eq. (6).

$$D(a, b) = \sum_{k=1}^{d} |a_k - b_k| \tag{6}$$

In the Eq. (6), where D is the number of dimensions of each data, with a and b being two data to be searched for the distance. If the number of results from the formula is large, then the level of similarity between training data and testing data will be not accurate, an the otherwise if the number of results from the formula is small, then the level of similarity between training data and testing data will be accurated.

The technique to validate the accuracy of a model that is built based on a certain data set is called K-fold cross-validation. With this technique, the dataset is randomly divided into several K partitions. Training data is the data used in the model development process, while testing data is the data used to validate the model. Then a series of collaborative experiments was conducted, then in each experiment, the partition data was used as testing data and the remaining partitions were used as training data.

3 Result and Discussion

In this study used the database consists of 12 speakers, each speaker speaks numbers 0–9 in Indonesian, so the total data set used is 120. In this study using K-fold cross-validation to achieve generalized results, with the number of K used 5. In the preprocessing stage, the silent signal cutting process is carried out so as not to interfere with the recognition system. In this paper, the signal is cut uniformly with a signal length of 2048 samples or 256 ms. After the preprocessing step, each signal is extracted using MFCC. 27 triangular filters are applied to the signal, then 13 coefficients out of 27 are stored as features. Then, after getting 13 features from the MFCC process, the next step is classification. In this study using the KNN classification with a distance function. There are 2 distance functions that will be used, namely the Euclidean distance and the Manhattan distance, these 2 methods will be compared to determine the highest level of accuracy. To get the highest accuracy, need the necessary determining the value of K or the neighborhood in KNN, to affect the accuracy of the system. The use KNN classification is combined with the 5-fold cross-validation by using odd K values in the KNN classification process in order to avoid the same value. In addition to comparing the distance classification method between Euclidean distance and Manhattan distance, the author also adds a frequency comparison to determine the effect of each input frequency on speech recognition accuracy. The frequencies used include 8000 Hz, 16000 Hz, 32000 Hz, and 44100 Hz.

Table 1 shows the accuracy of the Euclidean distance at K = 3 is 87% K = 5 is 69% K = 7 is 52%. While the Manhattan got an accuracy score 86% at K = 3, at K = 5 at 69% and at K = 7 at 53%. This shows that the highest value of K gives the smallest accuracy, because the highest value for K is more inflexible and high of bias. And the otherwise the smallest value of K gives the highest accuracy, because the smallest value for K is more flexible and low of bias. The accuracy obtained using Euclidean is able to outperform with Manhattan, with accuracy 88%. Furthermore Tables 2, 3, 4, and 5 shows the accuracy on each dataset with the other frequency signal, that is 8000 Hz, 16000 Hz, 32000 Hz, and 44100 Hz, using the value K = 3.

Table 1. K Values on Average Accuracy.

K	Classification	
	Euclidean	Manhattan
3	88%	86%
5	69%	69%
7	52%	53%

Table 2. Accuracy of Euclidean Distance and Manhattan Distance Using Frequency 8000 Hz

	Classification	
	Euclidean	Manhattan
Dataset 1	88%	86%
Dataset 2	90%	87%
Dataset 3	85%	85%
Dataset 4	92%	88%
Dataset 5	85%	84%
Average (%)	88%	86%

Table 3. Accuracy of Euclidean Distance and Manhattan Distance Using Frequency 16000 Hz

	Classification	
	Euclidean	Manhattan
Dataset 1	92%	89%
Dataset 2	88%	85%
Dataset 3	84%	84%
Dataset 4	92%	88%
Dataset 5	84%	84%
Average (%)	88%	86%

In Table 6, although using different frequencies, each classification, both Euclidean distance and Manhattan distance, obtains an accuracy value that does not change. Euclidean distance obtained the highest accuracy at each frequency compared to Manhattan distance. The Euclidean distance gets an accuracy value of 88% at each frequency, that is 8000 Hz, 16000 Hz, 32000 Hz, and 44100 Hz, while the Manhattan distance gets an accuracy value of 86% at each frequency.

Table 4. Accuracy of Euclidean Distance and Manhattan Distance Using Frequency 32000 Hz

	Classification	
	Euclidean	Manhattan
Dataset 1	92%	91%
Dataset 2	89%	88%
Dataset 3	84%	82%
Dataset 4	90%	85%
Dataset 5	85%	84%
Average (%)	88%	86%

Table 5. Accuracy of Euclidean Distance and Manhattan Distance Using Frequency 44100 Hz

	Classification	
	Euclidean	Manhattan
Dataset 1	88%	86%
Dataset 2	89%	85%
Dataset 3	86%	85%
Dataset 4	92%	90%
Dataset 5	85%	84%
Average (%)	88%	86%

Table 6. Average Accuracy of Euclidean Distance and Manhattan Distance Using Each Frequency

Frecuency (Hz)	Classification	
	Euclidean	Manhattan
8000	88%	86%
16000	88%	86%
32000	88%	86%
44100	88%	86%
8000	88%	86%
16000	88%	86%

4 Conclusion

This paper has presented features extraction method which is MFCC and using KNN as the classifier to identify and distinguish voice signal from number 0–9 in Indonesian.

Signal features used to access is 13 coefficients, using the value of K = 3 on the classification process and 5-fold cross-validation was applied to achieve generalize results. This study also presents a comparison of the results of the classification of Euclidean distance and Manhattan distance at each frequency of 8000 Hz, 16000 Hz, 32000 Hz, and 44100 Hz. The highest results were obtained by the Euclidean distance classification, both at frequencies of 8000 Hz, 16000 Hz, 32000 Hz, and 44100 Hz, Euclidean distance still outperformed the Manhattan distance with an accuracy value of 88%. In this paper, Euclidean distance wasable to outperform the Manhattan distance. This is influenced by the dataset used, Euclidean distance is better if used for cluster datasets, while Manhattan distance is more suitable if using absolute datasets.

References

1. S. Sharma, M. Kumar, and P. K. Das, "A Technique For Dimension Reduction Of Mfcc Spectral Features For," no. Icic, 2015.
2. N. Jain and S. Rastogi, "Speech Recognition Systems - A Comprehensive Study Of Concepts And," vol. 3, no. 1, pp. 1–3, 2019.
3. R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System," *2018 10th Int. Conf. Inf. Technol. Electr. Eng.*, pp. 280–284, 2018.
4. Y. Wang and B. Lawlor, "Speaker Recognition Based on MFCC and BP Neural Networks," pp. 0–3, 2017.
5. P. P. Dahake and K. Shaw, "Speaker Dependent Speech Emotion Recognition using MFCC and Support Vector Machine," pp. 1080–1084, 2016.
6. P. Mulak and N. R. Talhar, "Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset," 2015.
7. A. F. Pulungan, M. Zarlis, and S. Suwilo, "Performance Analysis of Distance Measures in K-Nearest Neighbor," 2020.
8. Ranny, "Voice Recognition using k Nearest Neighbor and Double Distance Method," 2016.
9. Y. Religia and A. S. Sunge, "Comparison Of Distance Methods In K-Means Algorithm For Determining Village Status In Bekasi District," *2019 Int. Conf. Artif. Intell. Inf. Technol.*, pp. 270–276, 2019.
10. A. Vijayan *et al.*, "Throat Microphone Speech Recognition using MFCC," no. July, pp. 397–400, 2017.
11. G. Jhawar, P. Nagraj, and P. Mahalakshmi, "Speech Disorder Recognition using MFCC," pp. 246–250, 2016.
12. S. T. Saste and J. Prof, "Emotion Recognition from Speech Using MFCC and DWT for Security System Sonali," pp. 701–704, 2017.