Bulletin of Electrical Engineering and Informatics
Vol. 7, No. 3, September 2018, pp. 350~358
ISSN: 2302-9285, DOI: 10.11591/eei.v7i3.957

Speaker Recognition in Content-based Image Retrieval for a High Degree of Accuracy

Suhartono[1], Fresy Nugroho[2], Muhammad Faisal[3], Muhammad Ainul Yaqin[4], Suyanta[5]
[1,2,3,4]Department of Informatics, UIN Maulana Malik Ibrahim, Malang, Indonesia
[5]Department of Mechanical Engineering, Politeknik Negeri Malang, Indonesia

Article Info

Article history:

Received Feb 28, 2018
Revised Jun 06, 2018
Accepted Aug 08, 2018

Keywords:

Fuzzy mamdani
Identification
Manhattan distance
Speaker recognition
Verification

ABSTRACT

The purpose of this research is to measure the speaker recognition accuracy in Content-Based Image Retrieval. To support research in speaker recognition accuracy, we use two approaches for recognition system: identification and verification, an identification using fuzzy Mamdani, a verification using Manhattan distance. The test results in this research. The best of distance mean is size 32x32. The best of the verification for distance rate is 965, and the speaker recognition system has a standard error of 5% and the system accuracy is 95%. From these results, we find that there is an increase in accuracy of almost 2.5%. This is due to a combination of two approaches so the system can add to the accuracy of speaker recognition.

Corresponding Author:

Suhartono,
Department of Informatics, UIN Maulana Malik Ibrahim,
Malang, Indonesia
Email: suhartono@ti.uin-malang.ac.id

1. INTRODUCTION

Speaker recognition is the process of analysis of the speaker identity based on voice characteristics [1]. The accuracy of speaker recognition is needed by system recognition, where the recognition system must be able to perform an accurate response in accordance with data from the human speech [2]. Artificial intelligent approaches have been used to increase accuracy in speaker recognition; one of the methods in the artificial intelligent approach is the fuzzy Mamdani method [3]. The Mamdani method can be used to identify a non-linear system such as speaker recognition [4]. Data derived from human speech has non-linear properties. To increase the accuracy of speaker recognition, we can use a data mining approach. One method in the data mining approach is the Manhattan distance method [5]. The Manhattan distance method is the methods for image matching using distance measurements on two speakers.

Research speaker recognition is studied in relation to the sound. The Mamdani method can be used to identify human speech as a voice recognition system [6]. The Mamdani method can also be used for speaker verification [7], where the Mamdani method can provide 86% accuracy, even though there was the noise level. While the Manhattan distance method can be used for speaker identification. Manhattan distance method can achieve the highest accuracy of 92.5% in sub-image size variations [8]. Both approaches have accuracy level varies, in this research; a combination of two methods can significantly improve the speaker recognition performance by up to 95% compared with no combination. Therefore, the purpose of this research is to build a high accuracy for speaker recognition systems.

In this research, we get the get the higher recognition accuracy by two processes, The first process is to evaluate the performance of speaker identification using the Mamdani method, the second process is to evaluate the performance of verification using Manhattan distance. By the combination of two methods for speaker recognition system, the results are significant in terms of accuracy for speaker recognition. The data

retrieval is carried out using the Content-Base Image Retrieval (CBIR) [9], the sounds signal from the speaker are processed to give a spectrogram. The results were obtained in the digital image from spectrogram. The retrieval of data speaker from the digital image was done by varying sizes. The variations in size are 256x256, 128x128, 64x64, 32x32 and 16x16 as sub-images. The process of retrieving from the spectrogram image is said to be a feature extraction [10]. For each sub-image was processed using kekre-transform. The result of kekre-transform of each sub-image was obtained by the mean value. The process of getting the mean value of each sub-image is said to be a feature vector of the speaker.

2. RESEARCH METHOD

For data processing and analysis were performed in Network Laboratory at BJ Habibie building, faculty of science and technology, State Islamic University of Maulana Malik Ibrahim, Malang city, East Java province, Indonesia. The speaker recognition consists of identification and verification process. In Figure 1, the identification process using the Mamdani method, the identification process determines the most suitable speaker from several speakers based on the sound signal. And then, the verification claim that the sounds match with the sound signal from the speaker, the block diagram of the system can follow in Figure 1. The speaker recognition system is described in the following sections and depicted as a flowchart in Figure 2.

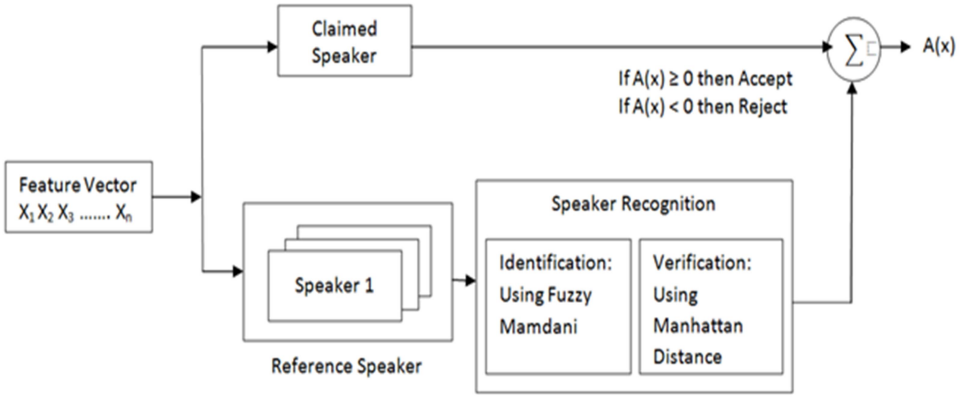

Figure 1. The block diagram of speaker recognition system

Figure 2, the schematic flowchart show three main stages. The stages were the pre-processing stage, the processing stage, and the post-processing stage. In the pre-processing stage is the extraction process, the extraction process is an image extraction algorithm, the stage of extracting features from objects in the image can be used to identify speaker with other speakers. In the processing stage generate the feature vector. The feature vectors can be used to represent the features from an object for easily analyzed. And then, the post-processing stage identify the speaker and to verify the accuracy of the speech recognition.

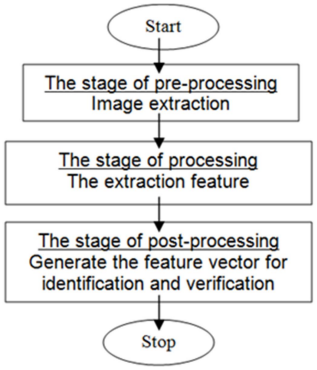

Figure 2. The flowchart for speaker recognition system

3. RESULTS AND ANALYSIS

3.1. Identification Using Fuzzy Mamdani

The fuzzy Mamdani method is known as the Max-Min method. This method was introduced by Ebrahim Mamdani in 1975. The schematic flowchart of the fuzzy Mamdani method for speaker recognition is shown in Figure 3, in this method, the input is a feature vector. The feature vector is the average value of the result of the kekre-transformation. The input of kekre-transformation is the sub-image of the spectrogram image. We used evaluation in error rate to identify the speaker. If the error rate was too high, the spectrogram image is not passed to the verification process, but the digital image is returned again in the process of feature vectors. The Mamdani method consists of four stages, as shown in Figure 4.

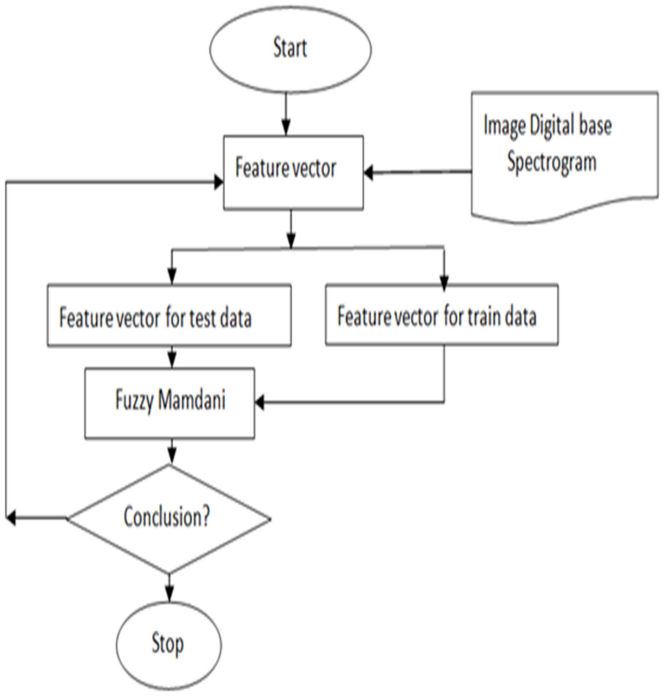

Figure 3. The schematic flowchart of identify speaker using the Mamdani method

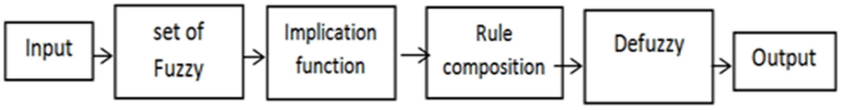

Figure 4. The stage of the fuzzy Mamdani method

In Figure 2, spectrogram image from the speaker was divided with the size of 256x256, 128x128, 64x64, 32x32 and 16x16 as sub-images, the vector features process each sub-image as input and output variables on the Mamdani method. The features mapped into crisp value (numeric) into the set of fuzzy and determine membership degree in the set of fuzzy. All input and output of the features data were processed on set fuzzy logic. The table of the data for the sub-image1 features was the input variable, as shown in Table 1.

Table 1. Input Variable for the Sub-image1 Features

Code	Set of input fuzzy		Domain
	Name	Notation	
1	Low	r	[0, 35]
2	Medium	s	[25, 35, 45]
3	High	t	[35,65]

In Table 1, the degree of membership value for each fuzzy variable is decided based on experimentation data. The domain value decreases can be represented as the fuzzy set low, shape of membership function: linear. The domain value of an increase and decrease can be represented as the fuzzy set medium, shape of membership function: triangle. The domain value increases can be represented as the fuzzy set high, the shape of membership function: linear. The fuzzy system was based on the Mamdani method using MATLAB 7.10.0. The programming codes are creating a fuzzy inference system (FIS) variable and adding input variable in membership function. The programming codes are shown in Figure 5.

```
%-----Create FIS variable;
a=newfis('speakerrecognition');
%---Add input feature_vector_sub_image_1;
a=addvar(a,'input','feature_vector_sub_image_1',[0 65]);
% Add membership function feature_vector_sub_image_1: Low, Medium, High;
a=addmf(a,'input',1,'Low','trimf',[0 35]);
a=addmf(a,'input',1,'Medium','trimf',[25 35 65]);
a=addmf(a,'input',1,'High','trimf',[35 65]);
%plotinput feature_vector_sub_image_1 to see the result;
plotmf(a,'input',1);
```

Figure 5. The programming codes of create FIS variable

After fuzzy membership sets of the input and output variables, the next process is product implication function in the form of IF-THEN rule, the parts of implication function are antecedent and consequences, the IF part of a rule is antecedent and the THEN part is consequent. One of implication function in the form of IF-THEN rule is shown in Figure 6.

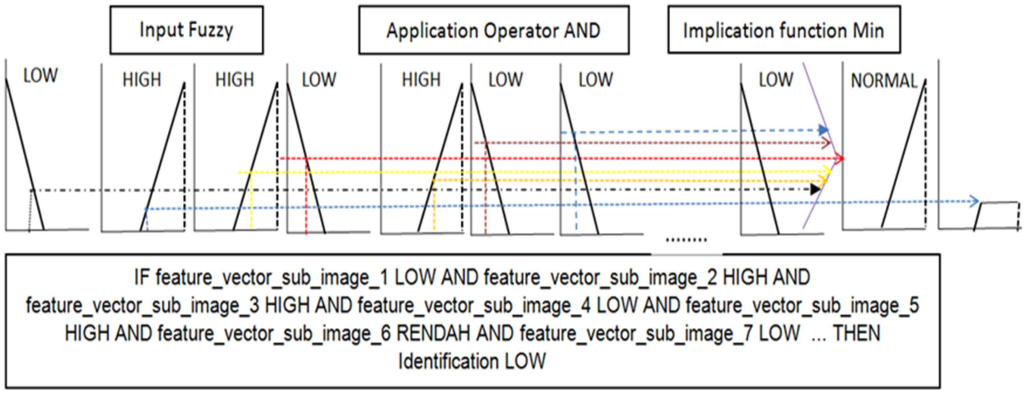

Figure 6. MIN implication function

The rules of the fuzzy inference system consist of fifty-four rules. The inference was obtained from set and correlation of fifty-four rules. The max method used to a fuzzy inference system. The maximum value would be at the center of the fuzzy set. All rules in proposition are evaluated in parallel. The output of the fuzzy inference system reflects contribution from every proposition. Generally, the output can be to be written in mathematics Equation 1.

$$\mu sf[xi] \longleftarrow \max(\mu sf[xi], \mu kf[xi]) \quad (1)$$

The μsf[xi] is the value of fuzzy membership for rule number-I, and μkf[xi] is the value of membership consequent for rule number-i, If there are three rules (proposition) then the programming code for creating rules in Matlab is shown in Figure 7. Max method can be used to select the best solution the speaker recognition, as shown in Figure 8.

```
% -Create rules;
% Rule1: IF feature_vector_sub_image_1 Low AND
feature_vector_sub_image_2 High AND
feature_vector_sub_image_3 High AND
feature_vector_sub_image_4 Low AND
feature_vector_sub_image_5 High AND
feature_vector_sub_image_6 Low AND
feature_vector_sub_image_7 Low ... THEN Identification Low;
% Rule2: ;
% Rule3: ;
rule1 = [1 1 1 1 ... 2];
rule2 = [2 0 2 1 ... 0];
rule3 = [3 2 3 1 ... 2];
%-------- Input rules;
listRules= [rule1;rule2;rule3];
a = addrule(a,listRule);
```

Figure 7. The programming codes of create rule

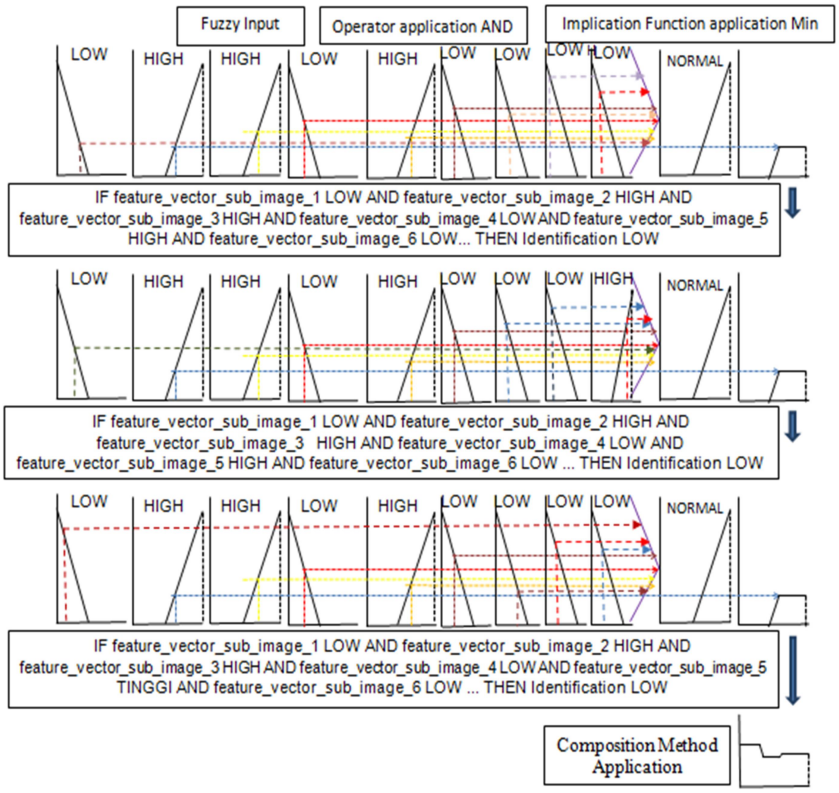

Figure 8. Fuzzy rule for MAX method

The input for the defuzzification process is a fuzzy set from rules of fuzzy inference and the output is a single value from a number in the set fuzzy domain. If set of fuzzy number in the certain range, the crisp value must be taken as the output. The centroid method is used for defuzzification. In this method, the crisp solution was obtained by taking the center point (z*) fuzzy area. This method can be to be written in mathematics formula (2). To get identification of feature vector can use programming code in Figure 8. Figure 9 shows the programming codes of evaluation for feature vector

$$z^* = \frac{\int_{z1} z\mu(z)dz}{\int_z \mu(z)dz} \quad (2)$$

```
% Perform evaluation for feature_vector_sub_image_1 = 7
AND feature_vector_sub_image_2 = 8
AND feature_vector_sub_image_3 = 9
AND feature_vector_sub_image_4 = 10
AND feature_vector_sub_image_5 = 11
AND feature_vector_sub_image_6 =12
AND feature_vector_sub_image_7 =13...;
evalfis ([7 8 9 10 11 12 13 ...], a)
```

Figure 9. The programming codes of evaluation for feature vector

The evaluation process of identification based twelve of test data. The evaluation process as the error rate was made can show in Table 2, an error rate is the difference between spectrogram image new and spectrogram image train. The identification can be accepted if error below or equal of 13%, the identification rejected if error above of 13%. In this research, the lowest error rate is 9.34%, in Table 2, the data with a dark color indicated as the most suitable speakers. The set of parameters (code, speaker) in the evaluation process is accepted if the set of parameters (code, speaker) has an error rate of error below or equal to 13%, the evaluation of the set of parameters (code=1, speaker=3) is accepted because the set of parameters (code=1, speaker=3) has an error rate below or equal to 13%, the next process for the set of parameters (code=1, speaker=3) need to verification process.

Table 2. The Error Rate of Identification System

Code	Speaker	identification system Size					Conclusion
		16x16	32x32	64x64	128x128	256 x 256	
1	Speaker 1	19.33	09.34	17.66	15.27	15.32	Accepted
2	Speaker 1	18.77	09.01	15.17	13.92	19.72	Accepted
3	Speaker 1	15.43	12.90	14.21	17.33	19.00	Accepted
4	Speaker 1	14.76	08.75	15.33	16.87	18.95	Accepted
5	Speaker 2	13.41	13.21	13.36	15.33	17.21	Rejected
6	Speaker 2	17.97	13.99	16.22	17.99	18.81	Rejected
7	Speaker 2	14.55	16.03	13.11	15.76	17.23	Rejected
8	Speaker 2	16.32	14.00	16.42	14.44	18.88	Rejected
9	Speaker 3	17.47	13.71	16.22	17.99	18.81	Rejected
10	Speaker3	16.52	12.67	16.17	16.93	18.76	Accepted
11	Speaker 3	15.54	13.44	18.99	20.01	21.44	Rejected
12	Speaker3	14.65	14.11	16.66	20.11	20.11	Rejected
	Mean	16.22	12.43	15.79	16.82	18.68	

In the Table 2, the lowest mean of error rate was size 32x32. The sub-image is the best frame size. In this research mean of error rate was 12.43. In Table 2, the five tests data from twelve were to be assessed as accepted, the five test data use to verification process in table 4. The five test data are set of parameters (code=1, speaker=1), set of parameters (code=2, speaker=1), set of parameters (code=3, speaker=1), set of parameters (code=4, speaker=1), and set of parameters (code=10, speaker=3).

3.2. Verification Using Manhattan Distance

Manhattan distance method is also called "city block distance". This method was the sum of the distances from entire attributes [11]. Generally, this method can be written in mathematics formula as Equation 3.

$$d_{ik} = \sum|x_{ik} - c_{ik}| \quad (3)$$

where; d_{ik}=Distance between x_{ik} and c_{ik}
x_{ik}=Target
c_{ik}=Comparative
k=Number of attributes in each case
i=Individual attributes from 1 to n

This method used a similarity measurement to speaker recognition based quantitative data [12]. This method can use to verification process for speaker recognition. The schematic flowchart for speaker verification system based spectrogram image can show in Figure 10. The verify process use the distance between spectrogram images new and spectrogram image train. The closest distance is accepted. For spectrogram images new in decline will be returned to the vector feature encoding process.

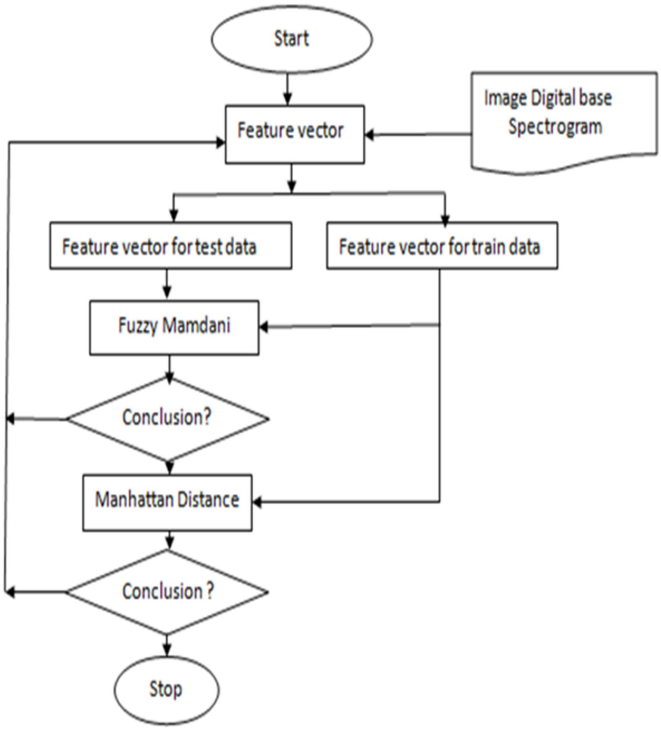

Figure 10. The schematic flowchart for verification using Manhattan distance

This research gives the steps to be done to develop verification, in Table 3, using this method; we construct a verification system for calculating the distance between the query and all image samples. For example, the spectrogram image for train data can show in Table 3. The new for test data can show Table 4.

Table 3. The Calculate the Distance for Train Data

Code	FV_1	FV_2	FV_3	FV_4	FV_5	FV_6	FV_7	Distance	C
1	10	9	8	9	4	5	9	22	1
2	4	5	7	8	9	3	4	14	2

Table 4. The Calculate the Distance for Test Data

Code	FV_1	FV_2	FV_3	FV_4	FV_5	FV_6	FV_7	Distance	C
1	5	6	9	6	12	5	7	22	1

Where, FV_i=Feature Vector of sub-image-i, C=Classification, to calculated distance between the new image spectrogram and the old image spectrogram from Table 1 and Table 2 can show in Equations 4 and 5.

$$|5-10|+|6-9|+|9-8|+|6-9 |+|12-4|+|5-5|+|7-9|=22 \quad (4)$$

$$|5-4|+|6-5|+|9-7|+|6-8|+|12-9|+|5-3|+|7-4|=14 \quad (5)$$

The evaluation of verification system using six test data. The verification system compares the distance rate according to the size sub-image from the speaker spectrogram image. In Table 5, the verification system use distance rate, the verification system of speaker recognition accepted if distance rate is below or equal to 1,000, the verification rejected if distance rate is above 1,000. The verification was accepted if the distance rate is below or equal to 1,000, and the distance rate is above 1,000. In Table 5, the four tests data from five is to be assessed as accepted, the four data test shown with dark color in row table, the four data test are set of parameters (code=1, speaker=1), set of parameters (code=2, speaker=1), set of parameters (code=3, speaker=1), and set of parameters (code=4, speaker=1). The data declared most suitable has the least distance as the best. The distance rate for set of parameters (code=1, speaker=1) is 965 in size 32x32. The mean for low distance is size 32x32. This size is the best in the process of speaker recognition. The lowest distance is 998.75. The data for set of parameters (code=10, speaker=3) rejected because the distance is above 1,000.

Table 5. Distance Rate for Verification Process

Code	Speaker	Distance rate					Conclusion
		Size					
		16x16	32x32	64x64	128x128	256 x 256	
1	Speaker 1	1,079	965	1,189	1,186	1,295	Accepted
2	Speaker 1	1,285	974	1,099	1,288	2,100	Accepted
3	Speaker 1	1,088	973	1,298	1,176	2,106	Accepted
4	Speaker 1	1,108	978	1,152	1,226	1,907	Accepted
10	Speaker 3	1,195	1,083	1,101	1,090	1,109	Rejected
	Mean	1,161.75	998.75	1,171.75	1,185.75	1,652.5	

The standard error is used in this research. To calculate the standard error can show in Equation 6.

$$SE = \frac{\sqrt{\frac{\sum(x-\mu)^2}{N}}}{\sqrt{n}} \quad (6)$$

Where, N=Population size, n=sample size, μ=mean and x=sample data. The level of standard error for less than 40% is said to be good and dependable [12]. The speaker recognition system has a standard error 5% and the system accuracy is 95%.

4. CONCLUSION

Speaker recognition has been successfully applied using fuzzy Mamdani and Manhattan distance. The fuzzy Mamdani used for identification process and the Manhattan distance used to the verification process. The best of distance means is size 32x32. The best of feature vector for the distance rate is 965 and the speaker recognition system has a standard error of 5%, the system accuracy is 95%, and the system can be said to be good.

ACKNOWLEDGEMENTS

The authors acknowledge UIN Maulana Malik Ibrahim Malang for the research grant of scheme BOPTN 2018.

REFERENCES

[1] Magdalena Igras-Cybulska, Bartosz Ziółko, Piotr Żelasko, and Marcin Witkowski, Structure of pauses in speech in the context of speaker verification and classification of speech type , EURASIP Journal on Audio, Speech, and Music Processing, 2016(8)

[2] Wu, D., Cao, J., and Wang, J., Speaker Recognition Based on i-vector and Improved Local Preserving Projection, TELKOMNIKA (Telecommunication, Computing, Electronics and Control) Indonesian Journal of Electrical Engineering. Vol.12, No.6, June 2014, pp. 4299 - 4305.

[3] Meng, Lei, Yin, S., and Xinyuan Hu. An Improved Mamdani Fuzzy Neural Networks Based on PSO Algorithm and New Parameter Optimization. *Indonesian Journal of Electrical Engineering and Computer Science*, 2016, 1(1): 201.

[4] Dutu L. C., Mauris G., and Bolon P. A Fast and Accurate Rule-Base Generation Method for Mamdani Fuzzy Systems, *IEEE Transactions on Fuzzy Systems,* 2017; 99:1–1.

[5] Sevugapandi, N. and Chandran, C.P. Classification algorithm for Gene Expression Graph and Manhattan Distance, *Indonesian Journal of Electrical Engineering and Computer Science*. Vol. 5, No. 2, February 2017, pp. 472-478.

[6] Silva W. L. S., and Serra G. L. d. O. *Proposal of an Intelligent Speech Recognition System*. In 2012 Third Global Congress on Intelligent Systems, 2012; 1:356–359.

[7] Lydia Abdul Hamid and Dzati Athiar Ramli, Quality based speaker verification systems using fuzzy inference fusion scheme, Journal of Computer Science, 2014, 10 (3): 530-543.

[8] Kulkarni, V., Kekre, H.B. Gaikar, P., and Gupta, N., 2012, Speaker Identification using Spectrogram of Varying Frame Sizes, *International Journal of Computer Applications.*2012; 50(20):27-33.

[9] Kekre H. B., and Shah K. *Performance Comparison of Kekre's Transform with PCA and Other Conventional Orthogonal Transforms for Face Recognition*. In 2009 Second International Conference on Emerging Trends in Engineering & Technology,2009;1:873–79.

[10] Kekre, H. B., Vaishali Kulkarni, Sunil Venkatraman, Anshu Priya, and Sujatha Narashiman. Speaker Identification Using Row Mean of DCT and Walsh Hadamard Transform, *International Journal on Computer Science and Engineering*, 2011;3(1):47-56

[11] Muhammad K. Farooq, Malik Jahan Khan, Shafay Shamail, and Mian M. Awais. *Intelligent project approval cycle for local government: case-based reasoning approach.* In *Proceedings of the 3rd international conference on Theory and practice of electronic governance* (ICEGOV '09), Tomasz Janowski and Jim Davies (Eds.). ACM, New York, NY, USA. 2009; 1; 68-73.

[12] Brooks, Peter. Metrics for Service Management. Van Haren, 2012:301