Security System Using A Robot Based On Speech Recognition

Wayan Dadang
Dept. of Electrical Engineering
Universitas Sriwijaya
Palembang, Indonesia
wayandadangunsri@gmail.com

Hera Hikmarika
Dept. of Electrical Engineering
Universitas Sriwijaya
Palembang, Indonesia
herahikmarika@ft.unsri.ac.id

Hermawati
Dept. of Electrical Engineering
Universitas Sriwijaya
Palembang, Indonesia
herma08@gmail.com

Bhakti Yudho Suprapto
Dept. of Electrical Engineering
Universitas Sriwijaya
Palembang, Indonesia
bhakti@ft.unsri.ac.id

Suci Dwijayanti*
Dept. of Electrical Engineering
Universitas Sriwijaya
Palembang, Indonesia
sucidwijayanti@ft.unsri.ac.id

***Abstract*—This study describes a security system using a humanoid robot by utilizing speech recognition. The robot has two main parts, namely, Raspberry Pi 3 and two Arduino UNO R3 as a slave. This robot is designed as a combination of speech recognition and voice biometric. The instruction given by a speaker must be obeyed by the robot using servo motor. Meanwhile, for voice biometric, robot may give access to an authorized person using speech recognition. Mel Frequency Cepstral Coefficients (MFCCs), their delta, and delta-delta are used as feature extraction which is fed to a classifier, Gaussian Mixture Model (GMM). Results of this study show that the robot may recognize the speaker with an accuracy of 99.4% and 99% for 50% of testing data and 20% of testing data, respectively. Thus, this suggests that the combination of MFCC and GMM can be implemented in speech recognition for security system performed by the robot.**

Keywords*—*speech recognition, voice biometric, robot, MFCC, GMM

I. Introduction

Humanoid robot can be defined as a machine which can be programmed and has a capability to mimic human as well as resembling human’s body [1]. There are various application of humanoid robot, for example as an assistant in building construction, education, and security system. However, not many studies focus on robot for security system which utilizes the speech. Thus, this study aims to utilize a humanoid robot which is designed to give an access to an authorized person into a room. The access is given based on speech recognition. Some studies have implemented voice biometric for security system. Kong Aik Lee et al. implemented speech technologies for smart home [2]. Meanwhile, [3] utilized speech for fire safety.

In this study, speech will be used as an access control for entrance and robot will give permission based on speaker recognition. Hence, the designed humanoid robot is not only able to communicate to human but also to control the security system through speech recognition. In speech recognition, utterances spoken by speaker are processed to generate features which is then sent to the speech engine. Thus, features extraction is important in speech recognition. The combination of Mel Frequency Cepstal Coefficients (MFCCs) and Dynamic Time Wraping (DTW) was used in [4]. MFCC and probabilistic neural network have been implemented for speech recognition in [5]. MFCCs have been utilized as feature extraction because they may capture the characteristic of voice and have low complexity in the speech recognition [6]. Thus, this study also uses MFCC because it is based on human hearing perception. Later, speech will be classified using Gaussian Mixture Model (GMM).

The paper will be organized as follows. In Section 2, methods implemented in the study are described. The experimental results and a discussion of the results are provided in Section 3. The conclusion is presented in Section 4.

II. Methods

The proposed security system was designed using a humanoid robot which can recognize an input speech to give a signal so the access will be opened. There are two parts that must be considered in the robot design, namely electronic system and body design.

Electronic system relates to wiring among components so they can work well. The wiring must suit with the communication sets between Rapberry Pi as a master and other components such as Arduino Uno, servo motor, Radio-frequency identification (RFID) MFRC522, and LED matrix. Communication between these components uses serial, serial peripheral interface (SPI), and Inter-integrated circuit (I2C) since they need less cable so the design becomes simpler. The design of electronic system of the proposed robot can be seen in Fig. 1. After the electronic system, another importan part in designing the proposed robot is the body which is made by a mannequin. Fig. 2 shows the visual of the proposed robot.

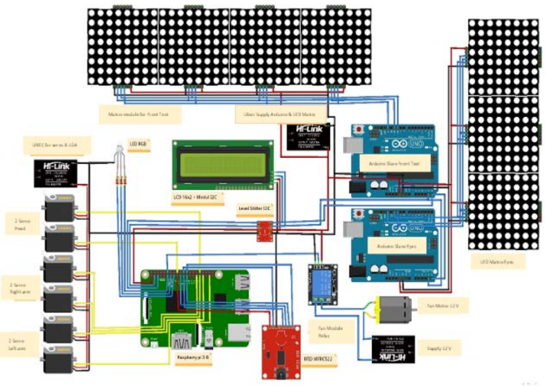

Fig. 1. Design of Electronic System

* Corresponding author

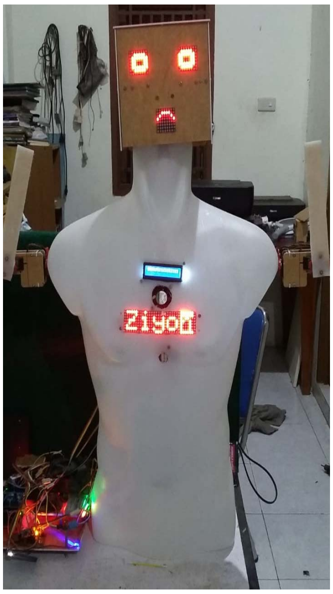

Fig. 2. Visual of Robot

In the robot, two main controls utilize Rapberry pi and two Arduino UNO as control slaves which control two LED matrix. These LED matrices functionize as the robot's eyes and robot's name as shown in Fig. 2.

The designed robot must be able to recognize the speech as the biometric to access the security system. The input speech which is captured by a microphone will be processed by Raspberry Pi to open the access. To prevent the failure of accessing the security system, radio-frequency identification (RFID) is used as the additional key. Working system of the proposed method can be seen in diagram as in Fig. 3.

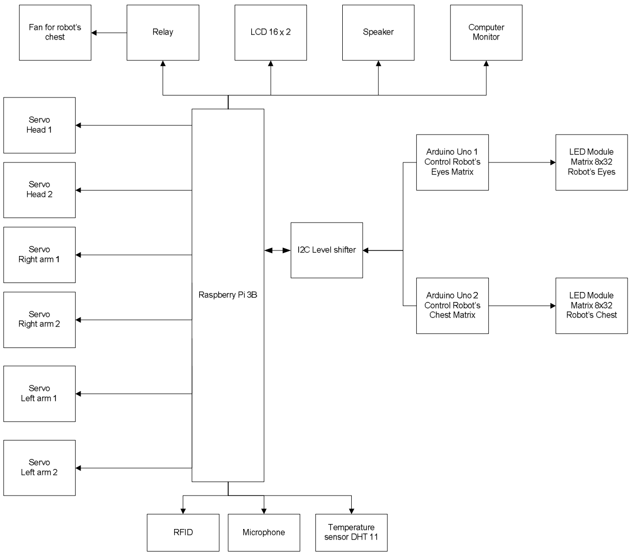

Fig. 3. Working System of Robot

Flowchart of speech recognition which is used as the biometric for a security system is shown in Fig. 4. At the beginning, the speech is recorded by a microphone and then the feature is extracted to capture characterictics of the speech. Here, Mel Frequency Cepstral Coefficients (MFCCs), their delta, and delta-delta are used as the features trained by Gaussian Mixture Model (GMM). After training, the new data is input to the trained GMM to identify the speaker. If the speaker is recognized, then this information is sent to the Arduino in robot to turn LED on or relay ON which indicates the speaker is the same as the trained dataset.

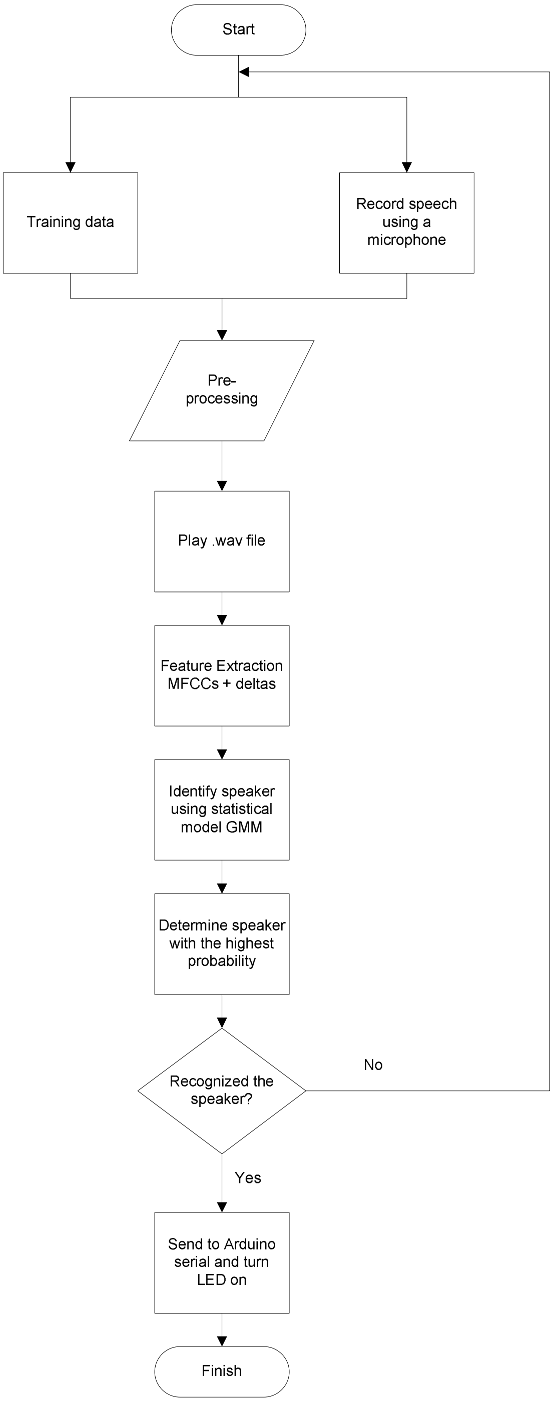
Start
Training data
Record speech using a microphone
Pre-processing
Play .wav file
Feature Extraction MFCCs + deltas
Identify speaker using statistical model GMM
Determine speaker with the highest probability
Recognized the speaker?
No
Yes
Send to Arduino serial and turn LED on
Finish

Fig. 4. Flowchart of Speaker Recognition for Security System

In the proposed speech recognition, the input speech will be pre-processed using short time energy to distinguish speech from non-speech before extracting the features. MFCCs are chosen as the feature because they represent the human's auditory characteristic. To get better extraction, delta and delta-delta are added because MFCC can only obtain spectral from a frame. In fact, an utterance also has dynamics [7][8]. Thus, delta and double delta will be useful to improve the accuracy of recognition. The process to obtain MFCC is shown in Fig. 5. The extracted features are then trained using a GMM which utilizes probability distribution approach as follows:

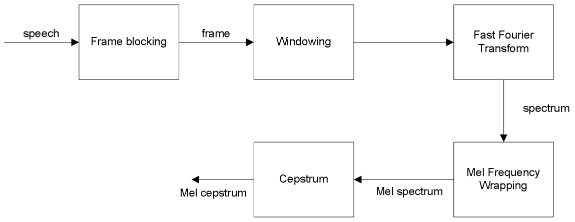

Fig. 5. Diagram Block of MFCC

$$p(x) = \sum_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k) \qquad ...(1)$$

where x is vector dimension d, π_k is weight of gaussian component k^{th}, μ_k is vector dimension d of gaussian component k^{th}, and Σ_k is d of covariant matrix d for gaussian component k^{th}. $\mathcal{N}$ is gaussian density function as

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \qquad ...(2)$$

where $|\Sigma|$ is determinant of Σ

To get optimal parameter of gaussian mixture, expectation maximization (EM) is used [9].

III. Results and Discussion

The speech recorded is the students' identity number from 100 students and the sampling frequency is 16 kHz. Each student repeats the speech ten times. Thus, there are 1000 dataset to be processed. After pre-processing the speech data, the MFCCs, their delta and double delta are extracted into 40 coefficents which are then trained into GMM.

To test the recognition system, the testing dataset is divided into two, 50% and 20 % of testing data, respectively. The purpose is to show the performance of the proposed system in recognizing testing data which are different from the training data. Those two testings show the results as the name of the speaker and LED turns on using Arduino which is communicated in serial.

A. Testing Using 50% of Data

Here, the 50% of data is used to train the speech and the rest 50% is for testing the proposed system.Thus, there are totally 500 data used in this testing. Then, testing is performed for five testing data for each speaker. The results are shown in Table 1.

TABLE I. Results Using 50% Data

No	Sample's Name	Five Testing Data				
		Test 1	*Test 2*	*Test 3*	*Test 4*	*Test 5*
1	Student 1	✓	✓	✓	✓	✓
2	Student 2	✓	✓	✓	✓	✓
3	Student 3	✓	✓	✓	✓	✓
4	Student 4	✓	✓	✓	✓	✓
5	Student 5	✓	✓	✓	✓	✓
6	Student 6	✓	✓	✓	✓	✓
7	Student 7	✓	✓	✓	✓	✓
8	Student 8	✓	✓	✓	✓	✓
9	Student 9	✓	✓	✓	✓	✓
10	Student 10	✓	✓	✓	✓	✓
11	Student 11	✓	✓	✓	✓	✓
12	Student 12	✓	✓	✓	✓	✓
13	Student 13	✓	✓	✓	✓	✓
14	Student 14	✓	✓	✓	✓	X
15	Student 15	✓	✓	✓	✓	✓
16	Student 16	✓	✓	✓	✓	✓
17	Student 17	✓	✓	✓	✓	✓
18	Student 18	✓	✓	✓	✓	✓
19	Student 19	✓	✓	✓	✓	✓
20	Student 20	✓	✓	✓	✓	✓
21	Student 21	✓	✓	✓	✓	✓
22	Student 22	✓	✓	✓	✓	✓
23	Student 23	✓	✓	✓	✓	✓
24	Student 24	✓	✓	✓	✓	✓
25	Student 25	✓	✓	✓	✓	✓
26	Student 26	✓	✓	✓	✓	✓
27	Student 27	✓	✓	✓	✓	✓
28	Student 28	✓	✓	✓	✓	✓
29	Student 29	✓	✓	✓	✓	✓
30	Student 30	✓	✓	✓	✓	✓
31	Student 31	✓	✓	✓	✓	✓
32	Student 32	✓	✓	✓	✓	X
33	Student 33	✓	✓	✓	✓	✓
34	Student 34	✓	✓	✓	✓	✓
35	Student 35	✓	✓	✓	✓	✓
36	Student 36	✓	✓	✓	✓	✓
37	Student 37	✓	✓	✓	✓	✓
38	Student 38	✓	✓	✓	✓	✓
39	Student 39	✓	✓	✓	✓	✓
40	Student 40	✓	✓	✓	✓	✓
41	Student 41	X	✓	✓	✓	✓
42	Student 42	✓	✓	✓	✓	✓
43	Student 43	✓	✓	✓	✓	✓
44	Student 44	✓	✓	✓	✓	✓
45	Student 45	✓	✓	✓	✓	✓
46	Student 46	✓	✓	✓	✓	✓
47	Student 47	✓	✓	✓	✓	✓
48	Student 48	✓	✓	✓	✓	✓
49	Student 49	✓	✓	✓	✓	✓
50	Student 50	✓	✓	✓	✓	✓
51	Student 51	✓	✓	✓	✓	✓
52	Student 52	✓	✓	✓	✓	✓
53	Student 53	✓	✓	✓	✓	✓
54	Student 54	✓	✓	✓	✓	✓
55	Student 55	✓	✓	✓	✓	✓
56	Student 56	✓	✓	✓	✓	✓
57	Student 57	✓	✓	✓	✓	✓
58	Student 58	✓	✓	✓	✓	✓
59	Student 59	✓	✓	✓	✓	✓
60	Student 60	✓	✓	✓	✓	✓
61	Student 61	✓	✓	✓	✓	✓
62	Student 62	✓	✓	✓	✓	✓
63	Student 63	✓	✓	✓	✓	✓
64	Student 64	✓	✓	✓	✓	✓
65	Student 65	✓	✓	✓	✓	✓
66	Student 66	✓	✓	✓	✓	✓
67	Student 67	✓	✓	✓	✓	✓
68	Student 68	✓	✓	✓	✓	✓
69	Student 69	✓	✓	✓	✓	✓
70	Student 70	✓	✓	✓	✓	✓
71	Student 71	✓	✓	✓	✓	✓
72	Student 72	✓	✓	✓	✓	✓
73	Student 73	✓	✓	✓	✓	✓
74	Student 74	✓	✓	✓	✓	✓
75	Student 75	✓	✓	✓	✓	✓

No	Sample's Name	Five Testng Data				
		Test 1	*Test 2*	*Test 3*	*Test 4*	*Test 5*
76	Student 76	✓	✓	✓	✓	✓
77	Student 77	✓	✓	✓	✓	✓
78	Student 78	✓	✓	✓	✓	✓
79	Student 79	✓	✓	✓	✓	✓
80	Student 80	✓	✓	✓	✓	✓
81	Student 81	✓	✓	✓	✓	✓
82	Student 82	✓	✓	✓	✓	✓
83	Student 83	✓	✓	✓	✓	✓
84	Student 84	✓	✓	✓	✓	✓
85	Student 85	✓	✓	✓	✓	✓
86	Student 86	✓	✓	✓	✓	✓
87	Student 87	✓	✓	✓	✓	✓
88	Student 88	✓	✓	✓	✓	✓
89	Student 89	✓	✓	✓	✓	✓
90	Student 90	✓	✓	✓	✓	✓
91	Student 91	✓	✓	✓	✓	✓
92	Student 92	✓	✓	✓	✓	✓
93	Student 93	✓	✓	✓	✓	✓
94	Student 94	✓	✓	✓	✓	✓
95	Student 95	✓	✓	✓	✓	✓
96	Student 96	✓	✓	✓	✓	✓
97	Student 97	✓	✓	✓	✓	✓
98	Student 98	✓	✓	✓	✓	✓
99	Student 99	✓	✓	✓	✓	✓
100	Student 100	✓	✓	✓	✓	✓

✓ = Recognized
X = Unrecognized

As shown in Table 1, most of the data is able to be recognized as the correct person by the proposed speech recognizer. When the speech is recognized, the LED in robot turns on which means that it gives an access to the authorized person and vice verse. However, error still occurs as indicated by "failed" in the table. The error may be caused by the noise while recording the speech.

The accuracy of the proposed method can be calculated as:

$$\text{accuracy} = \frac{\text{total testing data - total error}}{\text{total testing data}} \times 100\% \qquad \text{... (3)}$$

From total 500 samples and total error of 3, the accuracy is 99.4%.

This accuracy indicates that the proposed speech recognition which utilizes combination of MFCCs, delta, and delta-delta as features and GMM as classifier works well in recognizing the speaker.

B. *Testing Using 20% of Data*

Testing using 20% data means that there are 200 data used. The rest 80% of data is used to train the GMM. In this experiment, testing is performed for 2 data for each speaker. The results are shown in Table 2.

TABLE II. RESULTS USING 20% OF DATA

No	Sample's Name	Two Testing Data	
		Test 1	*Test 2*
1	Student 1	✓	✓
2	Student 2	✓	✓
3	Student 3	✓	✓
4	Student 4	✓	✓
5	Student 5	✓	✓
6	Student 6	✓	✓
7	Student 7	✓	✓
8	Student 8	✓	✓
9	Student 9	✓	✓
10	Student 10	✓	✓
11	Student 11	✓	✓
12	Student 12	✓	✓
13	Student 13	✓	✓
14	Student 14	✓	X
15	Student 15	✓	✓
16	Student 16	✓	✓
17	Student 17	✓	✓
18	Student 18	✓	✓
19	Student 19	✓	✓
20	Student 20	✓	✓
21	Student 21	✓	✓
22	Student 22	✓	✓
23	Student 23	✓	✓
24	Student 24	✓	✓
25	Student 25	✓	✓
26	Student 26	✓	✓
27	Student 27	✓	✓
28	Student 28	✓	✓
29	Student 29	✓	✓
30	Student 30	✓	✓
31	Student 31	✓	✓
32	Student 32	✓	X
33	Student 33	✓	✓
34	Student 34	✓	✓
35	Student 35	✓	✓
36	Student 36	✓	✓
37	Student 37	✓	✓
38	Student 38	✓	✓
39	Student 39	✓	✓
40	Student 40	✓	✓
41	Student 41	✓	✓
42	Student 42	✓	✓
43	Student 43	✓	✓
44	Student 44	✓	✓
45	Student 45	✓	✓
46	Student 46	✓	✓
47	Student 47	✓	✓
48	Student 48	✓	✓
49	Student 49	✓	✓
50	Student 50	✓	✓
51	Student 51	✓	✓
52	Student 52	✓	✓
53	Student 53	✓	✓
54	Student 54	✓	✓
55	Student 55	✓	✓
56	Student 56	✓	✓
57	Student 57	✓	✓
58	Student 58	✓	✓
59	Student 59	✓	✓
60	Student 60	✓	✓
61	Student 61	✓	✓
62	Student 62	✓	✓
63	Student 63	✓	✓
64	Student 64	✓	✓
65	Student 65	✓	✓
66	Student 66	✓	✓
67	Student 67	✓	✓
68	Student 68	✓	✓
69	Student 69	✓	✓
70	Student 70	✓	✓
71	Student 71	✓	✓
72	Student 72	✓	✓
73	Student 73	✓	✓
74	Student 74	✓	✓
75	Student 75	✓	✓
76	Student 76	✓	✓
77	Student 77	✓	✓

No	Sample's Name	Two Testing Data	
		Test 1	*Test 2*
78	Student 78	✓	✓
79	Student 79	✓	✓
80	Student 80	✓	✓
81	Student 81	✓	✓
82	Student 82	✓	✓
83	Student 83	✓	✓
84	Student 84	✓	✓
85	Student 85	✓	✓
86	Student 86	✓	✓
87	Student 87	✓	✓
88	Student 88	✓	✓
89	Student 89	✓	✓
90	Student 90	✓	✓
91	Student 91	✓	✓
92	Student 92	✓	✓
93	Student 93	✓	✓
94	Student 94	✓	✓
95	Student 95	✓	✓
96	Student 96	✓	✓
97	Student 97	✓	✓
98	Student 98	✓	✓
99	Student 99	✓	✓
100	Student 100	✓	✓

✓ = Recognized
X = Unrecognized

As shown in Table 2, recognition error may still occur. When the speaker is not recognized, the security system does not allow the user to accesss the room which is indicated by LED off in the robot. It might be caused by the noise in recoding. Using (3), the accuracy obtained is 99%. This result is slightly lower than the testing using 50% data. It may imply that the more data used in training, the less error may be obtained. Thus, the more data used in the training may improve the accuracy.

IV. Conclusion

In this work, the security system is performed using speech recognition which is used in the robot to give an access to the authorized person based on the recognized speech. The LED in the robot turns on when the speaker is recognized so only the authorized person gets an access to open the security system. From the testing performed to the system, MFCCs and their delta and double delta and GMM as the classifier are good combination to the speech recognition. The accuracy achieved are 99.4 % and 99% for testing 50% and 20% of data, respectively. The error occured in the proposed speech recognizion may be caused by the noise in the recording process.

Thus, in the future, we plan to use different algorithm for recognition such as deep learning which may be robust for noisy environment.

References

[1] V. Graefe and R. Bischoff, "Past , Present and Future of Intelligent Robots," in IEEE International Symposium on Computational Intelligence in Robotics and Automation for New Millenium, vol. 2, pp. 801-810, August 2003.

[2] K. A. Lee, A. Larcher, H. Thain, B. Ma, and H. Li, "Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home," in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[3] S. Park, Y. Kim, E. T. Matson, C. Lee, and W. Park, "An Intuitive Interaction System for Fire Safety Using A Speech Recognition Technology," in 6th International Conference on Automation, Robotics, and Application ICARA, pp. 388–392, 2015.

[4] B. J. Mohan and R. B. N, "Speech Recognition using MFCC and DTW," in International Conference on Advances in Electrical Engineering (ICAEE), pp. 1-4, January, 2014.

[5] K. S. Ahmad, A. S. Thosar, J. H. Nirmal, and V. S. Pande, "A Unique Approach in Text Independent Speaker Recognition using MFCC Feature Sets and Probabilistic Neural Network," in 8th International Conference on Advances in Pattern Recognition (ICAPR), pp. 1-6, 2015.

[6] A. M. Warohma, P. Kurniasari, S. Dwijayanti, Irmawan, and B.Y. Suprapto, "Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network," in 2018 Int. Semin. Appl. Technol. Inf. Commun., pp. 522–527, 2018.

[7] K. Kumat, C. Kim, and R.M. Stern, "Delta-spectral Cepstral Coefficiemts for Robust Speech Recognition," in IEEE International Conference on Acoistic, Speech, and Signal Processing (ICASSP), pp. 4784-4787, 2011

[8] S. Dwijayanti and M. Miyoshi, "Evaluation of Features for Voice Activity Detection Using Deep Neural Network," Journal of Theoretical and Applied Information Technology, vol. 96, no. 4, pp. 1114–1127, 2018.

[9] D. Yu and L. Deng, Automatic Speech Recognition, Springer, London, 2016