

Speaker and Speech Recognition Using Hierarchy Support Vector Machine and Backpropagation

Asti F. Fadlilah, Esmeralda C. Djamal*

Department of Informatics

Universitas Jenderal Achmad Yani

Cimahi, Indonesia

*Corresponding author: esmeralda.contessa@lecture.unjani.ac.id

Abstract— Voice signal processing has been proposed to improve effectiveness and facilitate the public, such as Smart Home. This study aims a smart home simulation model to move doors, TVs, and lights from voice instructions. Sound signals are processed using Mel-frequency Cepstrum Coefficients (MFCC) to perform feature extraction. Then, the voice is recognized by the speaker using a hierarchy Support Vector Machine (SVM). So that unregistered speakers are not processed or are declared not having access rights. For the process of recognizing spoken words such as "Open the Door", "Close the Door", "Turn on the TV", "Turn off the TV", "Turn on the Lights" and "Turn Off the Lights" are done using Backpropagation. The results showed that hierarchy SVM provided an accuracy of 71% compared to the single SVM of 45%.

Keywords: spoken word recognition, speaker recognition, MFCC, backpropagation, support vector machine.

I. INTRODUCTION

Voice signal processing has developed in the past, which allows computers to receive input in the form of voice commands. The application is suitable for people that have physical limitations. Also, it intends to people that want to do everything practically and efficiently daily work such as opening and closing doors. Voice command can change the operation of many devices. Sound signals consist of variables such as amplitude, wavelength, tempo, rhythmic and frequency that contain information about the words spoken, the identity of the owner of the voice, the emotions spoken or the combination.

Voice processing consists of three parts, namely identification of spoken words, identification of speaker identity, and emotional identification. In the process of identifying spoken words can be displayed in written form or can be read by technological devices. Recognition of words through voice signals has been carried out by many previous studies, one of them is Makhraj recognition for Al-Quran recitation [1], voice-based door access control system [2], pronunciation of the alphabet in English [3], language recognition Bengali automatically [4], play songs automatically [5]. While other study identified speaker [6], recognized of a person through footstep sound [7], identified Qori reciter in Arabic [8], identified speakers with whispered speech audio flow [9]. Emotional recognition has been carried out through voice signals for emotional recognition in speeches [10], analysis and classification of speech modes whispering through shouts [11] and identification of speakers for whispering speeches using training transformations [12].

Characteristics of a person's voice are determined physiology factor so that it can be verification tools of the speaker. The some one's physiology produce that sound depending on the dimensions of the vocal tract, mouth, nose

cavities, and other speech processing mechanisms in the human body [13]. The recognition of the speaker identity is divided into two types, namely saying a particular word or a random word. For the first type requires less computation, although it requires cooperative from volunteers as subjects [14].

The sound signal obtained needs to be extracted, which aims to distinguish specific patterns. The best parametric representation of sound signals determines recognition performance. The efficiency of parts affects the behavior of the identification stage. The most common method in extracting sound signals is using the Mel-Frequency Cepstrum Coefficients (MFCC) method considering the scale used is close to the level of human perceptions [15].

The Mel-Frequency Coefficients Cepstrum (MFCC) method as the extraction of sound signals and identification using the Backpropagation method produces an average accuracy of 97.5% in recognizing someone through footstep sound [7]. Meanwhile, another study evaluated the performance of LPC and Mel-Frequency Coefficients Cepstrum (MFCC). It got MFCC, and Artificial Neural Network (ANN) were able to identify speech signals better than using Linear Predictive Coding (LPC). The highest level of accuracy that achieved is almost 100% [16]. Research on Arabic numeral speech recognition systems using extraction of Delta-Delta Mel's frequency cepstral feature coefficient (MFCC) that results in the effectiveness of digit recognition in the dataset in 99.31% of cases [17].

Whereas in voice identification usually used Hidden Markov Model (HMM) [18], Gaussian Mixture Model (GMM) [8], Learning Vector Quantization (LVQ) [5], Backpropagation [7][19] or Support Vector Machine (SVM) [10] [20].

There is speech recognition to control the movement of the Robot Arm using MFCC extraction and Support Vector Machine (SVM). It has an average accuracy with 80% of data training, and 70% of new data [20].

This research builds identification of words and identification of speakers in speech recognition for smart home simulation instructions. The process starts with the extraction of voice signals using MFCC. Then, the speaker recognition was accomplished by compared to the list of speakers who are authorized — this way used Hierarchy Support Vector Machine. The speaker registered as many as five people, outside it was recognized as unknown. Known voice signals (in class 5 people enrolled), are processed to identify spoken words in order to carry out actions in the form of words in the Indonesian language "Buka Pintu", "Tutup Pintu", "Nyalakan TV", "Matikan TV", "Nyalakan Lampu" dan "Matikan Lampu".

II. PROPOSED METHODS

A. Data Acquisition

This research used voice signal of recording using the Audacity application with a frequency sampling of 8000Hz Channel mono 16bit .wav file format. Sounds are recorded through the microphone on the laptop offline in 2 seconds using 46 subjects and will be divided into 40 trainers and six test data. Six instructions for words spoken in Indonesian are "Open the Door", "Close the Door", "Turn on the TV", "Turn Off the TV", "Turn on the Lights" and "Turn Off the Lights". Pronunciation must be taken with precise articulation and the duration about two seconds. The data consist of training data and test data. Each recording produces sampling data of 16000 [1][5].

B. Speaker and Speech Recognition

The design of the speaker and spoken word identification starts with the pre-processing stage and feature extraction using MFCC. Then, the Cepstrum coefficient result as input in recognition of the speaker using hierarchy SVM and spoken word recognition using the Backpropagation algorithm. As in Fig. 1.

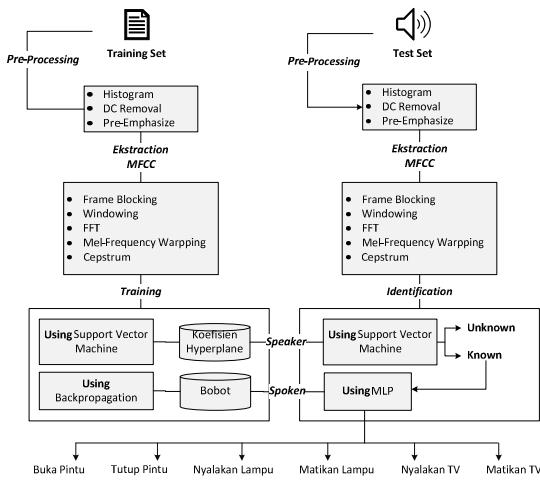


Fig. 1. Design of identification

C. Pre-Processing

Preprocessing is accomplished through three stages, namely Histogram Equalization, DC Removal, and Pre-Emphasize. Histogram Equalization is used to get the length of sample data that matches the specified data length. Several other studies flattened the sample data length to 16000.[1][5].

1) Histogram Equalization

Histogram Equalization aims to obtain cumulative distributive values. Histogram Equalization will be calculated using (1).

$$D'n = D[n] + D[n - 1] \quad (1)$$

After getting the cumulative distributive value, the level of sample data is performed using (2).

$$h[v] = \left(\frac{(D'[v] - \min(D'))}{\max(D') - 1} \right) N + 1 \quad (2)$$

2) DC-Removal

DC Removal is achieved by calculating the average of the sound sample data and subtracting the value of the sound sample data by the average value to obtain normalization using (3).

$$D[i] = s[i] - \frac{\sum_{i=1}^n s[i]}{n} \quad (3)$$

3) Pre-Emphasize

Pre-Emphasize is used to maintain high frequencies on a spectrum that is eliminated during sound processing. Pre-Emphasize is calculated using (4).

$$y[n] = s[n] - \alpha[n - 1] \quad (4)$$

D. Mel-Frequency Cepstrum Coefficients (MFCC)

Davis and Mermelstein introduced Mel-Frequency Cepstrum Coefficients (MFCC). MFCC is the most popular and common feature for sound recognition systems [21] because the methods approach human hearing perception. The frequencies component can represent better sounds [22]. To perform extraction, MFCC has five process steps that are sequentially processed, namely Frame Blocking, Windowing, Fast Fourier Transform, Mel-Frequency Warping, and Cepstrum.

1) Frame Blocking

Frame Blocking in this process divides the sound signal into several frames; one frame consists of several samples depending on time and the amount of sound frequency. To see the operation of the Frame Blocking stage can be seen in Fig. 2. In (5) the number of frames is 99 frames/second:

$$\left(\frac{(I-N)}{M} + 1 \right) \quad (5)$$

$$I = \text{sample rate} : \frac{Fs}{Ts} = \frac{16000}{2} = 8000$$

$$N = \text{sample point} : I * t = 8000 * 0,02 = 160$$

$$M = \frac{N}{2} = \frac{160}{2} = 80$$

$$\text{Sum of frame} = \left(\frac{(8000-160)}{80} + 1 \right) = 99 \text{ frame/seconds.}$$

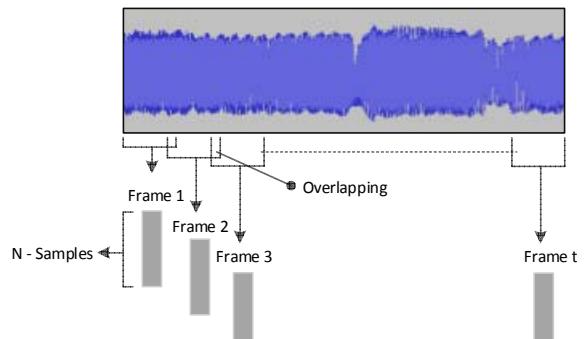


Fig. 2. Frame blocking process

2) Windowing

At this stage, the sound signal that has been divided into several frames is carried out by windowing to minimize

signal discontinuity, the windowing used is Hamming Window.

$$w(n) = 0,5 + 0,46 \cos \left[\frac{2\pi n}{N-1} \right], 0 \leq n \leq N-1] \quad (6)$$

3) Fast Fourier Transform

Fast Fourier Transform (FFT) is the process of converting each frame sample N from the time domain to the frequency domain, calculated using (7).

$$X(m) = \sum_{n=0}^{N-1} x(n) \cdot \left[\cos \left(\frac{2\pi mn}{N} \right) - j \sin \left(\frac{2\pi nm}{N} \right) \right] \quad (7)$$

4) Mel-Frequency Wrapping

Mel - Frequency Wrapping The Mel scale is used to map the signal frequency scale to a logarithmic scale for frequencies higher than 1 kHz. This scale makes the spectral frequency of the signal following human hearing [23]. This scale is defined by Stanley Smith, John Volkman, and Edwin Newman on (8).

$$mel(f) = 1127 \ln \left(1 + \frac{f}{700} \right) \quad (8)$$

5) Cepstrum

The cepstrum stage converts log Mel spectrum into the time domain. Calculated using (9). The number of MFCC coefficients used for the speech recognition process is 13 [24].

$$c_n = \sum_{k=1}^K (S[k]) \cos \left[\frac{\pi n(m+1)}{K} \right], n = 1, 2, \dots, K \quad (9)$$

E. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a technique for making predictions, both in the case of classification and regression. The concept of SVM can be explained simply as looking for the best hyperplane that functions as a separator of two classes. Fig. 3 shows several patterns that are members of two types +1 and -1. The model incorporated in the +1 class is symbolized as green (box), while the trend in class 1 is symbolized as blue (circle). In high-dimensional space, a hyperplane can maximize the distance (margin) between the data classes.

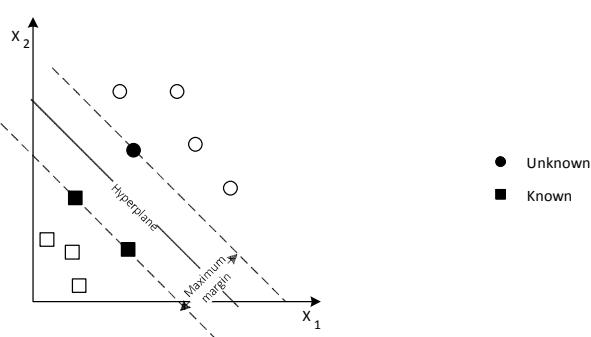


Fig. 3. Linear Support vector machine

The hyperplane margin was used to find the best hyperplane of two classes. Margin is the distance between the

hyperplane and the closest pattern to each class. The model most intimate to the hyperplane line is called a support vector. The line with the hyperplane on Fig. 3 shows the best hyperplane, which is located right in the middle of the two classes, while the right pattern on the dashed line is a support vector. In the search for the hyperplane function, the coefficients of w_i and b are required as bias (10).

$$y(x_i w_i + b) \geq \pm 1 \quad (10)$$

SVM can only separate from two classes. Therefore, to identify more than two classes is done in stages called hierarchy SVM, as in Fig 4. The image illustrates to recognize five registered speakers and the rest known as unknown.

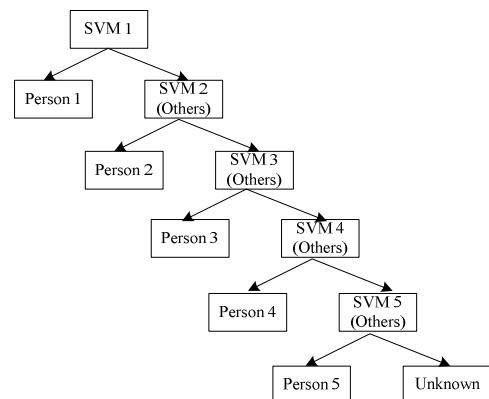


Fig. 4. Hierarchy Support vector machine

F. Backpropagation Algorithm

In Artificial Neural Networks, there is an activation function that is used to determine the output of a neuron. The activation function used must fulfill several conditions, namely continuous, easily differentiated, and non-descending functions [25]. The function that satisfies this requirement is the binary sigmoid shown in (11) which has a range of 0 and 1 with derivatives at (12).

$$f(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

$$f'(x) = f(x)(1 - f(x)) \quad (12)$$

The extracted data used backpropagation training. The number of neurons of the input layer (X) is obtained from the extraction process before. It is a 13 point for each frame. If there are 198 frames in two seconds, so that get 198 + 1 frame or 2587 as input neurons (n) in each learning. Because it has six keyword classes so that the number of output neurons (Y) is 6 neurons (m), so the number of hidden neurons (Z) is $\sqrt{(m * n)} = \sqrt{(6 * 2587)} = 124$. The algorithm used Multilayer Perceptron (MLP), as shown in Fig. 5.

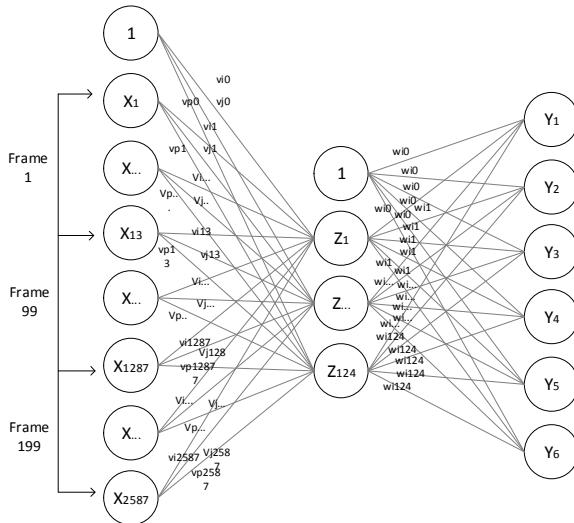


Fig. 5. Multi-Layer Perceptron model

III. RESULT AND DISCUSSION

Tests are carried out using trained data and new data. New data is obtained from the results of the same narcotic recording by directly using the microphone on the laptop. The tests carried out were the test of the effect of the number of MFCC extraction coefficients, the effect test of the optimization of training parameters, the examination of the impact of the amount of training data on SVM, offline identification testing and real-time testing. The discussion is explained in the following description.

A. MFCC Coefficients

The experiment of the number of MFCC coefficients aims to find the best amount of coefficients that have the highest accuracy. The results of the test of the effect of the number of MFCC coefficients can be seen in Table I.

TABLE I. CEPSTRUM COEFFICIENT TESTING

MFCC Coefficients	Time	MSE	Accuracy %	
			Training Data	New Data
7	663.51	0.126110824	88	56
10	699.43	0.103326918	89	65
13	616.14	0.056338934	95	71

The results in Table I, the magnitude of the number of coefficients in MFCC affect increasing the value of accuracy. In the results of this test, the number of coefficients used, the higher the accuracy value obtained. This result because the sound signal requires a large pattern distinguishing feature. Therefore, with the tests that have been carried out, the coefficient of 13 is the coefficient that has the highest accuracy with the value of correctness for training data that is 95% and for the new data 71%.

B. Optimization parameter

The effect of optimizing the training parameters on backpropagation aims to find the optimal settings that provide the best accuracy, the parameters tested using learning rate 0.01 and epoch 10000. The results of the test to optimize the training parameters can be seen in Table II.

TABLE II. OPTIMIZATION OF TRAINING PARAMETER

Hidden Neuron	Minimum Error	MSE	Accuracy %	
			Training Data	New Data
124	0.001	0.215	78	52
	0.010	0.229	77	53
	0.100	0.214	78	51
248	0.001	0.204	79	55
	0.010	0.200	80	57
	0.100	0.184	81	58
496	0.001	0.063	93	60
	0.010	0.048	95	71
	0.100	0.138	86	63

Based on the test results in Table II, it shows the higher the value of the Hidden Neuron and the target MSE value set, the higher the amount of accuracy. Accuracy results for training data are in the range of 70 - 95% while for new data is in the range of 52 - 71%. The training graph of the optimal parameter is learning rate 0.01, MSE target 0.01, and epoch 10000 can be seen in Fig. 6.

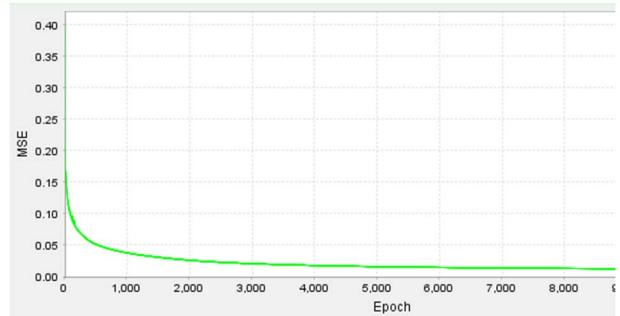


Fig. 6. Decrease of learning rate in training

C. The experiment of Support Vector Machine

Using SVM to recognize the speaker is done in two ways. The first method is to use a single SVM so that all registered speakers (five people) are placed in one class, and the other speakers become the second class. Whereas the second method is to use SVM hierarchy, as shown in Fig. 4. Accuracy comparisons between the two ways are shown in Table III.

It can be seen that if all the speakers are put together, the accuracy decreases while the use of SVM Hierarchy provides better recognition accuracy.

Meanwhile, the amount of data during training using Support Vector Machine influenced the accuracy of the system so that it needs to be tested with several different training data. Tests are carried out from the amount of training data, which is then added to the highest number of training data so that changes accuracy can be seen in Table III.

Table III, the identification system is the best development with a total of 1200 training data. During testing, the accuracy was below 95% for testing training data because it was confirmed that all were recognized. If the training data of more than 1200 data that improve system accuracy. Besides, a large number of classes is one of the triggers of the results of low precision.

TABLE III. SUPPORT VECTOR MACHINE EXPERIMENT

No	Amount of Train Data	Time (second)	Accuracy %	
			Training Data	New Data
Hierarchy SVM				
1.	200	0.024	91	66
2.	400	0.038	92	65
3.	600	0.047	93	63
4.	800	0.059	93	59
5.	1000	0.062	94	68
6.	1200	0.075	95	71
Single SVM				
7.	200	0.034	81	40
8.	400	0.044	82	38
9.	600	0.051	83	44
10.	800	0.064	83	45
11.	1000	0.077	84	42
12.	1200	0.089	85	45

The time of identification is influenced by a large amount of data so that the more the amount of training data, the longer the identification time will be. This result gave the number of coefficients that follow the amount of training data.

D. Speaker and Speech Recognition

If a set of 5 speakers are entitled to carry out instructions (person 1-5) and are also tested against unauthorized speakers (person other 1-3), give accurate results as in Table IV.

TABLE IV. ACCURACY RESULT

No	Test Data	Speech	Accuracy %	
			Training Data	New Data
1	Person 1	Buka Pintu	98	70
2		Tutup Pintu	96	72
3		Nyalakan TV	94	74
4		Matikan TV	95	67
5		Nyalakan Lampu	93	68
6		Matikan Lampu	94	70
7	Person 2	Buka Pintu	97	71
8		Tutup Pintu	98	73
9		Nyalakan TV	94	71
10		Matikan TV	94	70
11		Nyalakan Lampu	95	75
12		Matikan Lampu	96	70
13	Person 3	Buka Pintu	97	72
14		Tutup Pintu	96	70
15		Nyalakan TV	94	75
16		Matikan TV	95	69
17		Nyalakan Lampu	96	70
18		Matikan Lampu	94	69
19	Person 4	Buka Pintu	96	71
20		Tutup Pintu	97	73
21		Nyalakan TV	91	71
22		Matikan TV	94	70
23		Nyalakan Lampu	95	74
24		Matikan Lampu	92	75
25	Person 5	Buka Pintu	97	70
26		Tutup Pintu	98	72

No	Test Data	Speech	Accuracy %	
			Training Data	New Data
27		Nyalakan TV	94	75
28		Matikan TV	92	67
29		Nyalakan Lampu	90	68
30		Matikan Lampu	98	68
Total Recognized			95	71

Speech and speaker identification accuracy obtained by 71%. Need to test the effect of speaker characteristics and spoken words. Hearing from the speaker identification experiments, found that the average accuracy per person ranges from 70-72.33% and 0.99 so that the less significant deviation affects the accuracy. Meanwhile, the subsequent analysis that the characteristics of the spoken words were average accuracy ranged from 68.6 to 73.2% with a variation of 1.55%. The experiment shows that the features of the spoken word affect accuracy compared to the speaker characteristics. It is seen in Table IV that the lowest accuracy when saying "Matikan TV" (Turn off the in Indonesian).

Another factor is a problem in recording with noisy environmental conditions. The issue will take place if the environment in training data recording is unprecedented from recording from test data.

The graph of the accuracy show in Fig. 7 with the orange line is the accuracy for the new data, and the blue line is the correctness for training data.

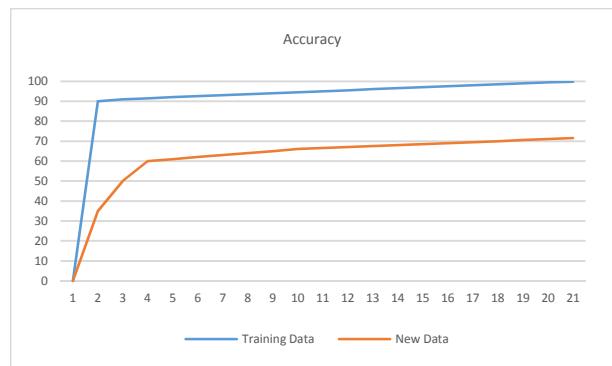


Fig. 7. Accuracy of identification

IV. CONCLUSION

This research developed identification using MFCC, SVM hierarchy, and Backpropagation, which provides 71% accuracy of new data. It can be a voice command for intelligent home simulations to move doors, TV, and lights.

In training using Learning Rate 0.001 and with Neuron Hidden 496, they are obtaining accuracy for training data by 95% and test data 71%. The small accuracy in new data is caused by several factors such as the environment when recording and pronunciation of speech of different people. The background of recording can affect the quality of recording so that decrease of accuracy. Judging from the characteristics of the sound, the effect of the spoken word significantly influences the accuracy compared with the speaker factor. Although both factors also affect the value of accuracy.

The results also showed that SVM hierarchy provided more accuracy compared to single SVM given the

characteristics of different registered speakers. Low correctness in a single SVM considers that all speakers with various voice characteristic are divided into two classes only. That is "accessible" and "inaccessible". Even though the two types have different speaker sound characteristics, meanwhile, the SVM method can only split into two classes. Therefore, the use of the SVM hierarchy can accommodate a large number of speaker classes and provide better accuracy.

ACKNOWLEDGMENT

The research was funded by "PTUPT –Penelitian Terapan Unggulan Perguruan Tinggi" from Ministry of Research Technology and Higher Education, Republik Indonesia 2019 with contract 2900/L4/PP/2019.

REFERENCES

- [1] A. Wahidah *et al.*, "Makhraj Recognition for Al-Quran Recitation using MFCC," *International Journal of Intelligent Information Processing*, vol. 4, no. 2, pp. 45–53, 2013.
- [2] K. F. Akingbade, O. M. Umanna, and I. A. Alimi, "Voice-Based Door Access Control System Using the Mel Frequency Cepstrum Coefficients and Gaussian Mixture Model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 4, no. 5, pp. 643–647, 2014.
- [3] T. B. Adam and M. Salam, "Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks," *International Journal of Computer Applications*, vol. 42, no. 12, pp. 21–27, 2012.
- [4] M. Hossain, M. N. Bhuiyan, and S. Engineer, "Automatic Speech Recognition Technique for Bangla Words," *International Journal of Advanced Science and Technology*, vol. 50, pp. 51–60, 2013.
- [5] E. C. Djamal, N. Nurhamidah, and R. Ilyas, "Spoken Word Recognition Using MFCC and Learning Vector Quantization," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, no. September, pp. 19–21.
- [6] C. Kumar, F. Rehman, S. Kumar, A. Mehmood, and G. Shabir, "Analysis of MFCC and BFCC in a Speaker Identification System," *International Conference on Computing, Mathematics and Engineering Technologies – iCoMET*, 2018.
- [7] J. E. Riwurohi, J. E. Istiyanto, K. Mustofa, and A. E. Putra, "People Recognition through Footstep Sound Using MFCC Extraction Method of Artificial Neural Network Backpropagation," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 18, no. 4, pp. 28–35, 2018.
- [8] T. S. Gunawan, N. Atikah, M. Saleh, and M. Kartiwi, "Development of Quranic Reciter Identification System using MFCC and GMM Classifier," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 1, pp. 372–378, 2018.
- [9] V. M. Sardar and S. D. Shirbahadurkar, "Speaker Identification of Whispering Sound using Selected Audio Descriptors," *International Journal of Applied Engineering Research ISSN*, vol. 13, no. 9, pp. 6660–6666, 2018.
- [10] H. Aouani and Y. Ben Ayed, "Emotion Recognition in Speech Using MFCC with SVM , DSVM and Auto-encoder," *International Conference on Advanced Technologies For Signal and Image Processing – ATSIP*, 2018.
- [11] C. Zhang and J. H. L. Hansen, "Analysis and Classification of Speech Mode : Whispered through Shouted," *INTERSPEECH, Eighth Annual Conference of the International Speech Communication Association*, pp. 1–4, 2007.
- [12] X. Fan and J. H. L. Hansen, "Speaker Identification for Whispered Speech Using A Training Feature Transformation From Neutral To Whisper," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1408 – 1421, 2011.
- [13] D. Bhattacharyya, R. Ranjan, F. Alisherov, and C. Minkyu, "Biometric Authentication : A Review Biometric Authentication: A Review," *International Journal of u- and e- Service, Science and Technology*, vol. 2, no. 3, pp. 13–27, 2009.
- [14] A. H. Mansour, G. Zen, A. Salh, H. Hayder, and Z. Alabdeen, "Voice recognition Using back propagation algorithm in neural networks," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 23, no. 3, pp. 133–139, 2015.
- [15] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal of Computing*, vol. 2, no. 3, pp. 138–143, 2010.
- [16] E. Mansour, M. S. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification," *international Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 3, pp. 55–66, 2013.
- [17] N. Hammami, M. Bedda, N. Farah, and R. O. Lakehal-ayat, "Spoken Arabic Digits recognition based on (GMM) for e-Quran voice browsing : Application for blind category," in *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, 2013, pp. 123–127.
- [18] Fitriolina, R. Kurnia, and S. Aulia, "Pengenalan Ucapan Metoda MFCC-HMM untuk Perintah Gerak Robot Mobil Penjelajah Identifikasi Warna," *Jurnal Nasional Teknik Elektro*, vol. 2, no. 1, pp. 31–40, 2013.
- [19] S. J. Subavathi and T. Kathirvalavakumar, "Adaptive modified backpropagation algorithm based on differential errors," *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, vol. 1, no. 5, pp. 21–34, 2011.
- [20] I. O. P. C. Series and M. Science, "The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm," 2018.
- [21] D. O. Shaughnessy, "Automatic speech recognition : History , methods and challenges," *Pattern Recognition*, vol. 41, pp. 2965–2979, 2008.
- [22] N. J. Ibrahim *et al.*, "Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, no. 8, pp. 207–215, 2008.
- [23] G. Giroti, T. Nakhat, M. Laddha, and P. M. Sarve, "Person Identification through Voice using MFCC and Multi-class SVM," *International Research Journal of Engineering and Technology (IRJET)*, no. May, pp. 690–693, 2018.
- [24] Abriyono and A. Harjoko, "Pengenalan Ucapan Suku Kata Bahasa Lisan Menggunakan Ciri LPC, MFCC, dan JST," *Indonesian Journal of Computing and Cybernetics Systems*, vol. 6, no. 2, pp. 23–34, 2012.
- [25] S. Amalia, "Pengenalan Digit 0 Sampai Digit 9 Menggunakan Ekstraksi Ciri MFCC dan Jaringan Syaraf Tiruan Backpropagation," *Teknik Elektro ITP*, vol. 6, no. 1, pp. 1–14, 2011.