JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2306-2312

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Mel Frequency Cepstral Coefficients (MFCC) Method and Multiple Adaline Neural Network Model for Speaker Identification

Sudi Mariyanto Al Sasongko [a,*], Shofian Tsaury [a], Suthami Ariessaputra [a], Syafaruddin Ch [a]

[a] *Department of Electrical Engineering, University of Mataram, Selaparang, Mataram, 83125, Indonesia*

Corresponding author: [*]mariyantosas@unram.ac.id

***Abstract*— Speech recognition technology makes human contact with the computer more accessible. There are two phases in the speaker recognition process: capturing or extracting voice features and identifying the speaker's voice pattern based on the voice characteristics of each speaker. Speakers consist of men and women. Their voices are recorded and stored in a computer database. Mel Frequency Cepstrum Coefficients (MFCC) are used at the voice extraction stage with a characteristic coefficient of 13. MFCC is based on variations in the response of the human ear's critical range to frequencies (linear and logarithmic). The sound frame is converted to Mel frequency and processed with several triangular filters to get the cepstrum coefficient. Meanwhile, at the speech pattern recognition stage, the speaker uses an artificial neural network (ANN) Madaline model (many Adaline/ which is the plural form of Adaline) to compare the test sound characteristics. The training voice's features have been inputted as training data. The Madaline Neural Network training is BFGS Quasi-Newton Backpropagation with a goal parameter of 0,0001. The results obtained from the study prove that the Madaline model of artificial neural networks is not recommended for identification research. The results showed that the database's speech recognition rate reached 61% for ten tests. The test outside the database was rejected by only 14%, and 84% refused testing outside the database with different words from the training data. The results of this model can be used as a reference for creating an Android-based real-time system.**

***Keywords*— MFCC; ANN; madaline; identification; speaker.**

Manuscript received 11 Nov. 2022; revised 17 Jan. 2023; accepted 8 Sep. 2023. Date of publication 31 Dec. 2023.

I. INTRODUCTION

Voice recognition recognizes a voice owner's identity by comparing the voice's features as input with each speaker's features inside and outside the existing database [1]. In some conditions, voice recognition becomes essential in human-computer interaction [2]. One of the mathematical computer technologies used to recognize the different characteristics of the human voice is the Fast Fourier Transform (FFT) [3]–[5]. FFT is a method for transforming a time zone signal into a frequency region signal and then storing it in digital form as a frequency-based signal spectrum [6].

Much research has been done with themes related to voice identification using artificial neural network methods, Self-Organizing Maps (SOM), Backpropagation, and other rules [7]. There are several speech features commonly used to extract speaker characteristics, including Linear predictive coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and Lateral Prefrontal Cortex (LPFC) [8]–[11]. MFCC has good results for feature extraction in sound and images[12]. MFFC is also combined with other methods to produce a high level of recognition, for example, using Self Organizing Maps (SOM).[13]. This study aims to determine speech recognition accuracy using the Self Organizing Maps (SOM) artificial neural network method using the MFCC model for voice feature extraction[14]. Researchers have also widely studied the speaker recognition system using several ways [15]–[17]. In addition, the speaker verification system has also been widely developed and researched [18]–[20].

An application of artificial neural network models Adaline and Madaline then compare the effectiveness in classifying goiter. Several studies using Adaline networks have also been carried out [21][22]. Madaline network performance is slightly superior to Adaline network [23]. Recognition level in identifying the speaker's voice by trying various SNR (Signal to Noise Ratio) values has been done. Sound with SNR from 20 dB to 80 dB has a success rate according to its SNR value. The greater the SNR value, the higher the identification success rate [24].

Based on the research that has been done, identification of the human voice spoken by the owner. According to the spoken words, the simulation could identify or detect the sound pattern's owner. Mel Frequency Cepstrum Coefficient (MFCC) feature extraction is the first step in the identification procedure. The approach may determine the Cepstrum coefficient based on how a person hears. It is linear for low frequencies and logarithmic for high frequencies. After getting the results of the cepstrum, training, and testing were carried out using the Madaline (Many Adaline) Artificial Neural Network (ANN). This method is a development of the Perceptron method and is the plural form of the Adaline (Adaptive Linear) ANN model. The difference with the Perceptron method is that in the Adaptive Linear method, the weight modification is carried out by a method known as the delta rule method or, commonly called the Least Mean Square (LMS) method.

II. Material and Method

Voice identification is made to determine the level of accuracy of speech pattern recognition produced by the owner [25].

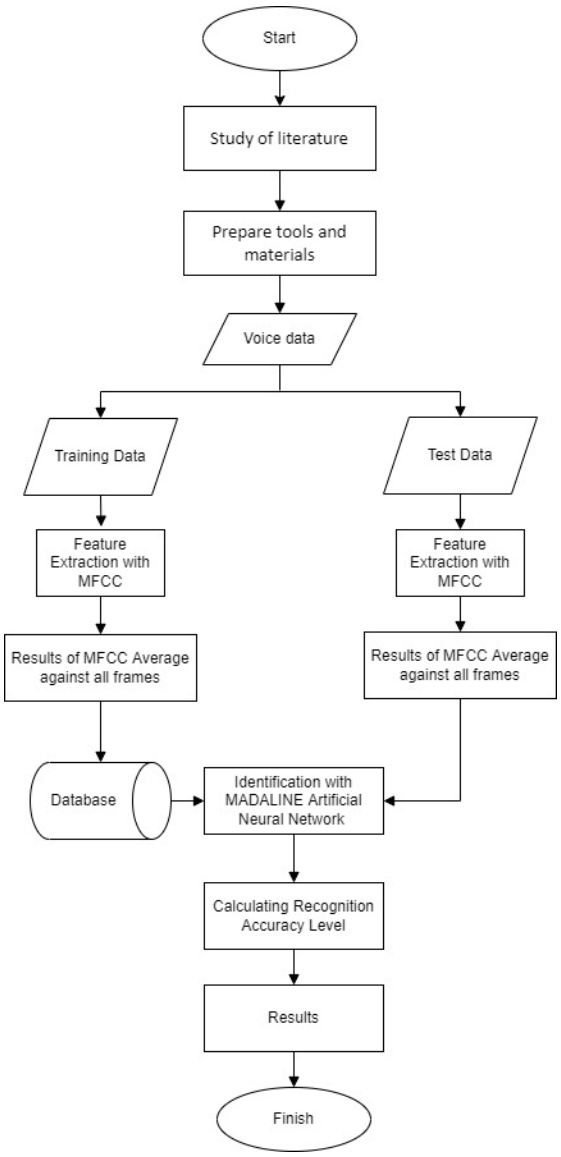

Fig. 1 Research stages

An aspect of the spoken voice was acquired using the Mel Frequency Cepstrum Coefficient (MFCC) approach and subsequently recognized by the Madaline model of Artificial Neural Networks (ANN). The system design process begins with conducting research and analyzing the system to be built. Here are some strategies that could be carried out in system design.

A. Voice Recording

Voice recording is used as a command input using the Goldwave software program [26]. Fig. 2 shows the sound recording process. Data from a speaker's voice signal is recorded using a microphone connected to a laptop. The recording is done on speakers with the GoldWave application with a duration of 5 seconds per sound at a sampling rate (Fs) of 16000 Hz and mono channels.

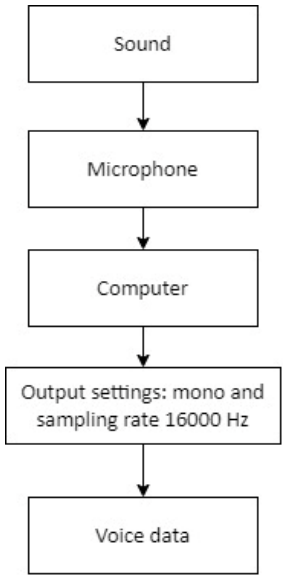

Fig. 2 Sound recording process

Twenty-five speakers could be divided into ten speakers included in the database and 15 other speakers outside the database. The speakers in the database consist of 8 men and two women who say the word "Telkom Laboratory." Speakers outside the database also comprised eight men and two women saying the same word, namely "Telkom Laboratory," and five other people saying different words from the database, namely "electrical engineering." Each speech data is saved as an audio file in ".wav" format, which is named after the speaker's name and followed by a pronunciation order index.

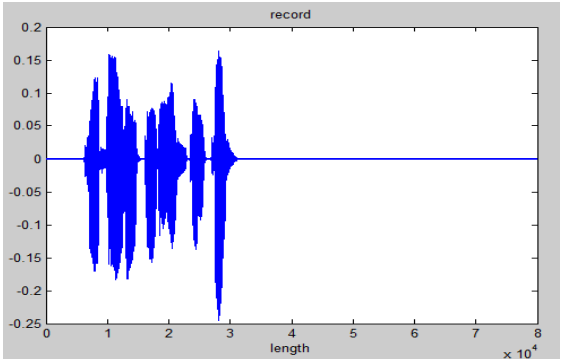

Fig. 3 Voice recording results

The recording was done 25 times for the data in the database. Almost 15 of the 25 recorded data are used as training and testing data. At the same time, the other ten are only used as test data. Ten speeches were recorded and used as test data for data outside the database. The results of the recording are then saved in .wav format. The recording results can be seen in Fig. 3.

B. Feature Extraction (MFCC)

Voice signal feature extraction in this study using MFCC. The parameters of the MFCC are:

- Input, namely voice input, comes from each speaker and is saved in a wav file. Each speaker had ten file records.
- Each file could be processed as a sampling step.
- The sampling rate is the number of values taken in one second. This study used a sampling rate of 16000 Hz[27].
- The time frame is the desired time for one frame (in milliseconds). The time frame used is 50 ms.
- Lap, which is overlapping, consists of N/2 data.
- The cepstrum coefficient is the desired number of cepstrum as the output of the frame. The cepstrum coefficient used is 13. The coefficient value of 13 is obtained from the spectrum value of the frequency value of the dominant voice data.

The stages of the MFCC process are as follows:

1) Frame Blocking: The result of voice recording is an analog signal in the time domain, a time-variant [28]. Therefore, it must be cut into specific time slots to be considered invariant.

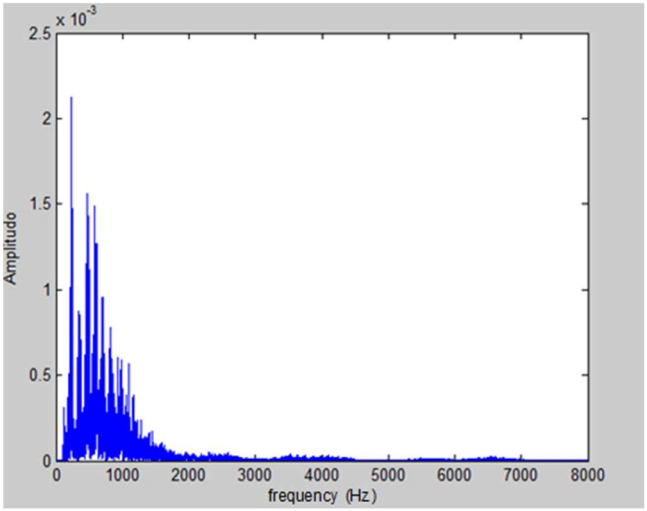

Fig. 4 The result of the FFT process of sound recording

One frame contains 800 samples, and another overlaps along 400 models or 50% of the total sample between shelves. Fig. 4 shows the results of the FFT process from voice recordings, and Fig. 5 shows the results of the frame-blocking process from sound recordings. Using a sampling frequency of 16000, the voice signal is cut by 50 milliseconds. Where the calculation is as follows:

Sampling rate (Fs) = 16000 Hz
Time frame (Ts) = 50 ms or 0.05 s
Frame size (N) = 16000 * 0.05 = 800 samples
Overlapping (M) = 800/2 = 400 samples

Then, the voice signal is cut along 800 at each overlapping 400. Each piece is called a frame. So, in one frame, there are 800 samples from 80000 existing samples.

2) Hamming Window: The sound signal cut into several frames could cause data errors in the Fourier transform process. A Hamming Window is needed to reduce the discontinuity effect of the frame-blocking process, especially at the beginning and end of each frame [30].

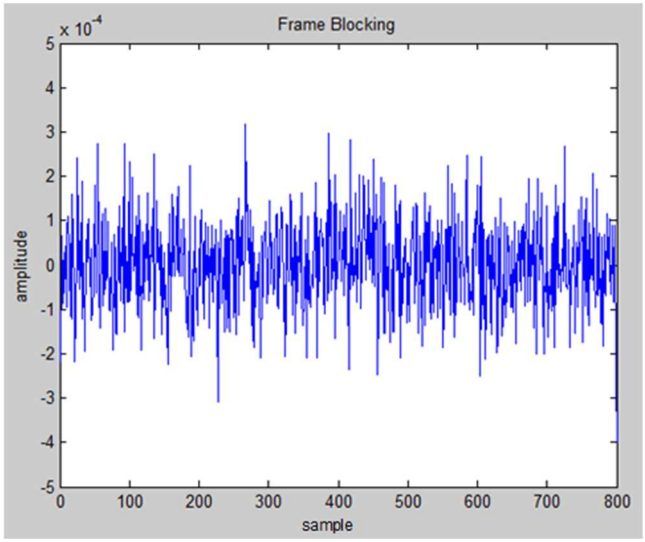

Fig. 5 Frame blocking result

The framing process causes a signal discontinuity (cut off/not connected). The windowing process reduces signal discontinuity from the beginning to the end of the frame. Fig. 6 shows the voice data after the Hamming Window process for voice with an SNR of 80 dB.

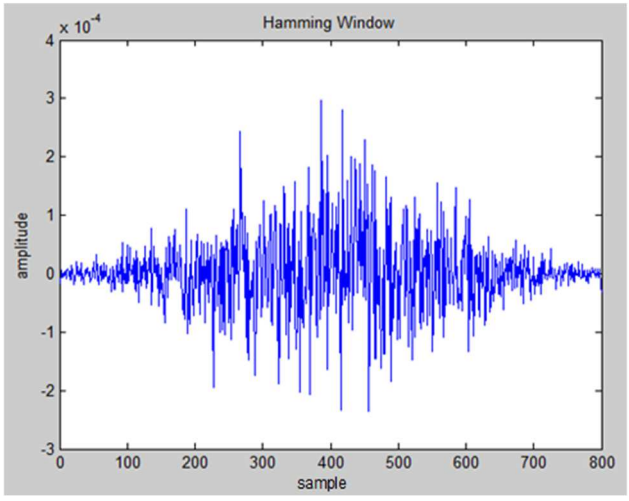

Fig. 6 Hamming windowing result

3) Fast Fourier Transform: In the Fourier transform process, there is a change in the shape of the input voice signal from the time domain into the frequency domain [29]. The following process is the Fast Fourier Transform (FFT) process. FFT is a process used to convert voice signals from the time domain into the frequency domain. The signal to be converted is a signal processed by frame blocking. Then each frame could be processed by FFT. Fig.7 shows the sound data after the Fourier transform process.

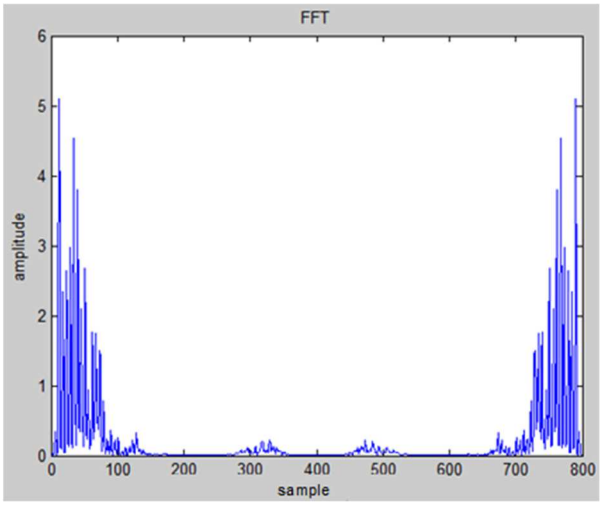

Fig. 7 Voice data after going through the FFT process

4) Mel Frequency Wrapping: Mel frequency wrapping aims to filter the spectrum of each frame. Signals that have passed the FFT process could then be filtered using a filter bank. The frequency scale of the filter bank is the same as the concept of human hearing, so the frequency scale is often used as an extraction parameter in sound signal processing. Fig. 8 shows the triangular filter bank.

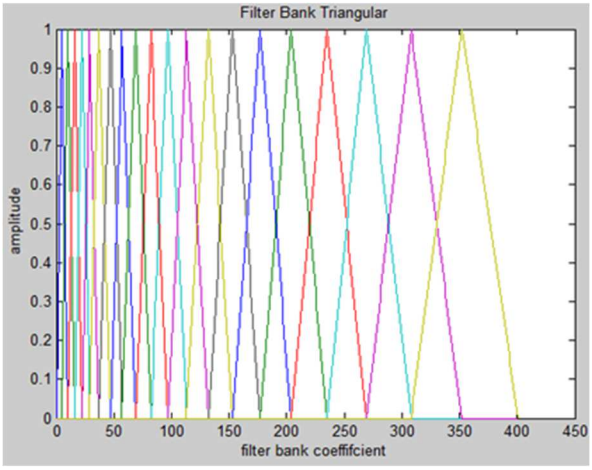

Fig. 8 Triangular Bank Filter process results

The mapping between hertz and Mel scale frequencies is linear for frequencies below 1000 Hz and logarithmic for frequencies above 1000 Hz [31]. Equations 1 and 2 show the formula for forming Mel frequency wrapping. Formula 1 is used for conversion from frequency scale to Mel scale.

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

Formula 2 is used to calculate the Mel scale to the frequency scale.

$$Mel\text{-}1(f) = 700 * \left[\exp\left(\frac{f}{1127}\right) - 1\right] \quad (2)$$

Furthermore, a filter array is formed, which contains some M triangular filters with M triangular filters used 20. Fig. 9 shows the sound characteristics after the filter bank process for sound without noise. The filtering results could produce 20 cepstrum parameters according to the number of triangular filters [32].

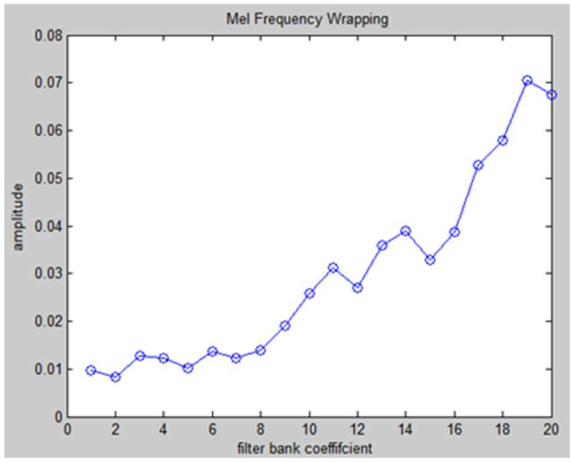

Fig. 9 Mel Frequency Wrapping Results

5) Cepstrum: Cepstrum results from the log Mel spectrum from the frequency domain converted into the time domain using DCT, which produces a matrix measuring the number of frames * coefficient [33].

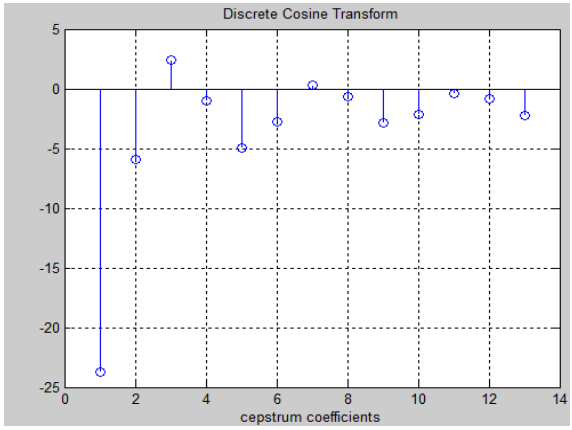

Fig. 10 DCT (Discrete Cosine Transform) Results

Cepstrum is the last process and is carried out after the filterbank process. Cepstrum is used to convert log Mel spectrum into cepstrum using DCT (Discrete Cosine Transform). Fig. 10 shows the sound characteristics after the DCT process for noiseless sound. In this case, 13 dominant cepstrum parameters are used.

The feature extraction result using MFCC has a feature matrix of *nxk*, n is the number of frames, and k is the coefficient. It produces a matrix of the same size in each vote, namely a matrix of size *lxk*. The coefficients are averaged for each row. The results of this cepstrum could be used as input to the Madaline process.

C. Madaline Artificial Neural Network Process

Before training the Madaline model of artificial neural networks, consider the network architecture [34]. The network architecture was chosen with a constructive approach: a small Adaline network with one or more hidden layers. The Adaline Neural Network model's activation and threshold function equations also obtain the hidden layer. Then, it develops the number of hidden units and additional weights until the desired solution is obtained.

Each neuron in the input layer consists of feature extraction results with the MFCC method and a predetermined weight.

The number of neurons in the input layer corresponds to the number of variables selected as network input plus one biased neuron. Fig. 11 shows that the number of input layer neurons is between 1 and 10 according to the number of speakers used as input data in the system, plus one bias neuron.

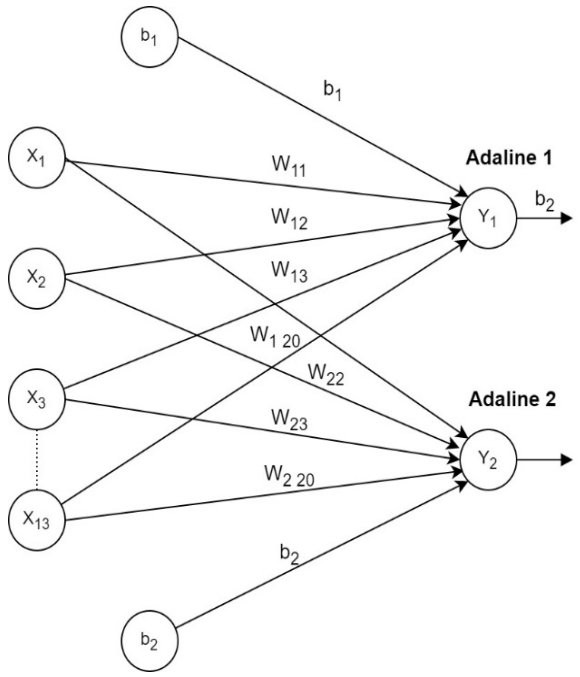

Fig. 11 Madaline architecture

The initial weights and biases are initialized with a small random number between 0 to 1. The initial weight will affect whether the network will reach a local minimum or global minimum and the duration of its convergence. The initial weight that is too large makes the derivative value of the activation function minimal. It causes the weight change to be tiny as well. The size of the input layer weight matrix is 10 x 13.

Another thing to note also is that the parameters that must be set in the network include:

1) *Learning rate:* The learning rate selected was 0.01 to 0.99 during the training. Generally, the automatic learning rate is 0.01

2) *Goal parameters:* The performance objective is the target value of the performance function. The iteration will be stopped if the value of the performance function is less than or equal to the performance objective.

3) *Maximum number of iterations:* Maximum iteration is the maximum number of epochs performed during the training process. The iteration will be stopped if the number that has been trained exceeds the maximum number of iterations.

III. Results and Discussion

A. Network Training with the Madaline Neural Network Model.

Network training is carried out to see the system's performance and find the slightest error value during the training process by changing several parameters.

1) Looking for a type of training with a target: Table 1 shows that the best type of training used in the Madaline Neural Network training is BFGS Quasi-Newton Backpropagation. In training, the results followed the target with few iterations and a small number of errors. In the BFGS Quasi-Newton Backpropagation, there has also been a change in the learning rate (lr) from 0.1 – 0.9. Meanwhile, the results do not affect the training process, both the output results and the errors caused.

2) Looking for the best goal parameters: Table 2 shows the output of each training course with changes in goal parameters from 0.01 to 0.00001. By paying attention to the production of the two Adaline networks in training, the best goal parameter value is 0.0001.

TABLE I
NETWORK TRAINING WITH CHANGING TYPES OF EXERCISE.

No	Training	Output	Error
1	Gradient Descent with Momentum and Adaptive Learning Rule Backpropagation	[2.6 1.5 3 3.8 6.4 4.8 6.09 7.4 8.8 10.4]	[-1.6 0.48 -0.06 0.1 -1.4 1.14 0.9 0.52 0.18 -0.41]
2	BFGS Quasi-Newton Backpropagation	[1.0001 1.9999 2.9999 4.0 5.0 6.0 7.0 8.0 8.9999 9.9999]	[1.0e-03 (-0.0548 0.089 0.059 -0.0624 -0.159 -0.067 -0.005 -0.07 0.085 0.1)]
3	Batch Training with Weight and Bias Learning Rules	[NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN]	[NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN]
4	Gradient Descent with Momentum Backpropagation	[2.7 1.47 3.0 3.7 6.5 4.77 6.0 7.4 8.8 10.4]	[-1.7 0.52 -0.6 0.2 -1.4 1.22 0.97 0.57 0.2 -0.44]
5	Bayesian Regulation Backpropagation	[2.95 3.59 2.74 2.78 6.6 5 5.26 6.33 8.16 8.77]	[-1.95 -1.59 0.25 1.2 -1.61 0.99 1.73 1.66 0.83 1.22]
6	Levenberg-Marquardt Backpropagation	[0.99 1.997 2.998 4.003 5.004 6.002 7.003 7.99 8.998 9.99]	[1.0e-03(0.5 2.4 2 -0.3 -4.2 -2.1 -3.6 1.6 1.5 2.4)]
7	Resilient Backpropagation	[2.95 3.59 2.74 2.78 6.6 5 5.26 6.3 8.16 8.77]	[-1.95 -1.59 0.25 1.22 -1.61 0.99 1.7 1.66 0.83 1.22]

TABLE II
NETWORK TRAINING WITH CHANGING TYPES OF EXERCISE

No	Goal	Y	1	2	3	4	5	6	7	8	9	10
1	0.01	Y1	0.975	2.06	2.982	3.979	5.027	5.928	7.047	8.000	9.025	10.002
		Y2	1.004	1.960	2.986	3.987	5.027	5.999	6.994	8.004	9.029	9.991
2	0.001	Y1	1.000	2.000	2.999	4.001	4.999	5.999	6.999	8.000	9.000	9.999
		Y2	0.996	1.962	2.984	3.993	5.039	5.982	7.042	7.966	9.028	9.994
3	0.0001	Y1	1.000	2.000	3.000	4.000	5.000	6.000	7.000	8.000	9.000	10.000
		Y2	0.999	2.001	2.999	4.001	4.998	6.000	7.000	8.000	9.000	10.000
4	0.00001	Y1	1.000	1.999	2.999	4.000	5.000	5.999	7.000	7.999	9.000	10.000
		Y2	1.000	2.000	2.999	4.000	5.000	6.000	6.999	7.999	9.000	9.996

3) The results of the change in weight at the 15th training: Furthermore, the weights listed in Table 3 could be used in the testing process.

TABLE III
NETWORK TRAINING WITH CHANGING TYPES OF EXERCISE.

No	Network	
	1st Network Weight(W_1)	2nd Network Weight (W_2)
1	-791.302	-790.348
2	1710.265	1708.189
3	-766.349	-765.38
4	-1213.16	-1211.75
5	1563.804	1561.905
6	286.3674	286.1335
7	-1912.1	-1909.92
8	1148.174	1146.752
9	1192.689	1191.519
10	-2705.69	-2702.84
11	2352.769	2350.283
12	-1110.43	-1109.27
13	244.465	244.2159

B. Test Results in the Database

Table 4 shows speakers' speech recognition accuracy using the Madaline artificial neural network model (ANN).

TABLE IV
TESTING IN DATABASE

No	Voice	Number of Pronunciations	Recognized speakers	
			Amount	Level of accuracy (%)
1	Person 1	10	9	90%
2	Person 2	10	7	70%
3	Person 3	10	6	60%
4	Person 4	10	4	40%
5	Person 5	10	5	50%
6	Person 6	10	6	60%
7	Person 7	10	8	80%
8	Person 8	10	6	60%
9	Person 9	10	3	30%
10	Person 10	10	7	70%
	Total	100	61	61%

Table 4 shows that ANN Madaline cannot adequately recognize some voices, so the recognition percentage is only 61%.

C. Test Results Outside the Database

Table 5 shows the speaker's voice rejection accuracy using the Madaline artificial neural network model (ANN). The tables found that some votes were still well recognized by ANN Madaline, so the rejection percentage was only 14%.

TABLE V
TESTING OUTSIDE THE DATABASE

No	Voice	Number of Pronunciations	Recognized speakers	
			Amount	Level of accuracy (%)
1	Person A	10	10	100%
2	Person B	10	10	100%
3	Person C	10	8	80%
4	Person D	10	10	100%
5	Person E	10	0	0%
6	Person F	10	10	100%
7	Person G	10	10	100%
8	Person H	10	9	90%
9	Person I	10	9	90%
10	Person J	10	10	100%
	Total	100	86	86%

D. Testing outside the Database with Different Words

Table 6 shows that the test with different pronunciations showed imperfect rejection, which was 84%.

TABLE VI
TESTING OUTSIDE THE DATABASE WITH DIFFERENT PRONUNCIATION

No	Voice	Number of Pronunciations	unrecognized speakers	
			Amount	Level of accuracy (%)
1	Person 21	5	1	80%
2	Person 22	5	0	100%
3	Person 23	5	0	100%
4	Person 24	5	0	100%
5	Person 25	5	2	60%
	Total	25	3	84%

Testing the exact words from outside the database did not produce good results. As many as ten people said ten times, only one person was wholly rejected. As many as ten people said ten times, and only one was deserted. Likewise, testing with a different word, namely "electrical engineering," results in an imperfect rejection. Even though it produces a pretty good percentage, it still has a recognizable sound. This is due to the large amount of training data used in the training process, and the sound that contains silence is still legible during the sound recording process.

IV. CONCLUSION

Testing with the speaker's voice in the database obtained an introduction percentage of 61%. Meanwhile, testing with test data with speakers outside the database only rejected the introduction of 14%. The test with test data outside the database with different words resulted in a denial of 86%. In addition, the Madaline artificial neural network is not suitable for identification and is more suitable for classification and prediction research. This paper describes the process of identifying sounds based on the words spoken by the speaker. MFCC is one of the features of human voice feature extraction based on ear response filters, linear at low frequencies and logarithmic at medium frequencies. The Madaline-based speaker speech recognition algorithm gives good results for one-dimensional system identification, although it is not superior. Furthermore, the identification process should be attempted in real-time with MFCC and the Madaline detection algorithm or other algorithms.

The results of identifying the speaker's speech are not good. It is estimated that the Madaline algorithm used is a type I. In Madaline type I, the input layer is directly connected to the output layer, so the updated weight function depends on an error variable in the output. For further research, it can be tried to identify the speaker's utterance with the Madaline type II algorithm, where a hidden neuron layer is added between the input and output layers. With the addition of the hidden layer, it is hoped that each neuron in the Adaline network could be better at updating the weight of the disturbance. The Madaline algorithm should be implemented in one-dimensional cases that the Adaline algorithm can handle. The two algorithms can identify or predict weather-related cases, grayscale-based image patterns, or predict facial features with the characteristic variables being worked out as one-dimensional vectors.

REFERENCES

[1] B. Alkhatib and M. M. W. Kamal Eddin, "Voice Identification Using MFCC and Vector Quantization," Baghdad Science Journal, vol. 17, no. 3(Suppl.), p. 1019, Sep. 2020, doi:10.21123/bsj.2020.17.3(suppl.).1019.

[2] R. Jahangir et al., "Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network," IEEE Access, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/access.2020.2973541.

[3] A. Riyani, "A Identifying Human Voice Signals Using the Fast Fourier Transform (Fft) Method Based on Matlab," Journal of Informatics, Information System, Software Engineering and Applications (INISTA), vol. 1, no. 2, pp. 42–50, May 2019, doi:10.20895/inista.v1i2.52.

[4] M. S. A. Apsari and I. M. Widiartha, "Classification of Women's Voices Using Fast Fourier Transform (FFT) Method," JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), vol. 10, no. 1, p. 59, Aug. 2021, doi: 10.24843/jlk.2021.v10.i01.p08.

[5] M. Muhathir, S. Susilawati, and R. Muliono, "Analisis Fast Fourier Tansform untuk Pengenalan Voice Register Wanita dalam Teknik Bernyanyi," Journal of Informatics And Telecommunication Engineering, vol. 2, no. 2, p. 92, Jan. 2019, doi:10.31289/jite.v2i2.2166.

[6] J. F. Mahdi, "Frequency analyses of human voice using fast Fourier transform," Iraqi Journal of Physics, vol. 13, no. 27, pp. 174–181, Feb. 2019, doi: 10.30723/ijp.v13i27.276.

[7] M. H. Widianto, H. I. Pohan, and D. R. Hermanus, "Introduction to Indonesian Syllables Using the LPC Method and the Neural Network of Backpropagation," International Journal of Engineering Trends and Technology, vol. 69, no. 5, pp. 137–146, May 2021, doi:10.14445/22315381/ijett-v69i5p220.

[8] L. Wang, "A Machine Learning Assessment System for Spoken English Based on Linear Predictive Coding," Mobile Information Systems, vol. 2022, pp. 1–12, Sep. 2022, doi: 10.1155/2022/6131572.

[9] P. Choubey, "Warped Linear Predictive Coding of Speech Signal of Processing," International Journal for Research in Applied Science and Engineering Technology, vol. 9, no. 5, pp. 1819–1827, May 2021, doi:10.22214/ijraset.2021.34680.

[10] R. Mohd Hanifa, K. Isa, and S. Mohamad, "Speaker ethnic identification for continuous speech in malay language using pitch and MFCC," Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, no. 1, p. 207, Jul. 2020, doi:10.11591/ijeecs.v19.i1.pp207-214.

[11] R. R. Huizen and F. T. Kurniati, "Feature extraction with mel scale separation method on noise audio recordings," Indonesian Journal of Electrical Engineering and Computer Science, vol. 24, no. 2, p. 815, Nov. 2021, doi: 10.11591/ijeecs.v24.i2.pp815-824.

[12] S. M. Al Sasongko, E. D. Jayanti, and S. Ariessaputra, "Application of Gray Scale Matrix Technique for Identification of Lombok Songket Patterns Based on Backpropagation Learning," JOIV : International Journal on Informatics Visualization, vol. 6, no. 4, p. 835, Dec. 2022, doi: 10.30630/joiv.6.4.1532.

[13] H. Kim and S. H. Jung, "SOGN: novel generative model using SOM," Electronics Letters, vol. 55, no. 10, pp. 597–599, May 2019, doi:10.1049/el.2019.0202.

[14] A. de Oliveira, M. Dajer, and J. Teixeira, "Clustering Pathologic Voice with Kohonen SOM and Hierarchical Clustering," Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, 2021, doi: 10.5220/0010210901580163.

[15] N. N. An, N. Q. Thanh, and Y. Liu, "Deep CNNs With Self-Attention for Speaker Identification," IEEE Access, vol. 7, pp. 85327–85337, 2019, doi: 10.1109/access.2019.2917470.

[16] F. Ye and J. Yang, "A Deep Neural Network Model for Speaker Identification," Applied Sciences, vol. 11, no. 8, p. 3603, Apr. 2021, doi: 10.3390/app11083603.

[17] O. Mamyrbayev, A. Toleu, G. Tolegen, and N. Mekebayev, "Neural architectures for gender detection and speaker identification," Cogent Engineering, vol. 7, no. 1, p. 1727168, Jan. 2020, doi:10.1080/23311916.2020.1727168.

[18] C. Xu, W. Rao, J. Wu, and H. Li, "Target Speaker Verification With Selective Auditory Attention for Single and Multi-Talker Speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2696–2709, 2021, doi:10.1109/taslp.2021.3100682.

[19] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," Computer Speech & Language, vol. 60, p. 101027, Mar. 2020, doi:10.1016/j.csl.2019.101027.

[20] H. Shim, J. Jung, J. Kim, and H. Yu, "Integrated Replay Spoofing-Aware Text-Independent Speaker Verification," Applied Sciences, vol. 10, no. 18, p. 6292, Sep. 2020, doi: 10.3390/app10186292.

[21] N. A. Norani, M. S. Mohd Kasihmuddin, Mohd. A. Mansor, and N. S. N. Khurizan, "Logic Learning in Adaline Neural Network," Pertanika Journal of Science and Technology, vol. 29, no. 1, 2021, doi:10.47836/pjst.29.1.16.

[22] F. Salehi, M. Jaloli, R. Coben, and A. M. Nasrabadi, "Estimating brain effective connectivity from EEG signals of patients with autism disorder and healthy individuals by reducing volume conduction effect," Cognitive Neurodynamics, vol. 16, no. 3, pp. 519–529, Nov. 2021, doi: 10.1007/s11571-021-09730-w.

[23] "A Research on Different Filtering Techniques and Neural Networks Methods for Denoising Speech," Special Issue, vol. 8, no. 9S2, pp. 503–511, Aug. 2019, doi: 10.35940/ijitee.i1107.0789s219.

[24] S. Kumar, S. S. Gornale, R. Siddalingappa, and A. Mane, "Gender Classification Based on Online Signature Features using Machine Learning Techniques", Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 260–268, May 2022. doi: 10.31838/jcr.07.09.222.

[25] T. S. Gunawan, M. F. Alghifari, M. A. Morshidi, and M. Kartiwi, "A Review on Emotion Recognition Algorithms using Speech Analysis," Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 6, no. 1, Mar. 2018, doi: 10.52549/ijeei.v6i1.409.

[26] V. Kumar and O. P. Roy, "Formant Measure of Indian English Vowels for Speaker Identity," Journal of Physics: Conference Series, vol. 2236, no. 1, p. 012011, Mar. 2022, doi: 10.1088/1742-6596/2236/1/012011.

[27] S. Tirronen, S. R. Kadiri, and P. Alku, "The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection," Journal of Voice, Apr. 2022, doi: 10.1016/j.jvoice.2022.03.021.

[28] H. Nurdiyanto, H. Kurniawan, and S. Karnila, "Human Voice Recognition Using Artificial Neural Networks," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 9, pp. 1070–1077, 2021, [Online]. Available: https://www.proquest.com/scholarly-journals/human-voice-recognition-using-artificial-neural/docview/2623465013/se-2.

[29] A. Revathi, C. Ravichandran, P. Saisiddarth, and G. S. R. Prasad, "Isolated Command Recognition Using MFCC and Clustering Algorithm," SN Computer Science, vol. 1, no. 2, Mar. 2020, doi:10.1007/s42979-020-0093-x.

[30] J.-S. Sheu and C.-W. Chen, "Voice Recognition and Marking Using Mel-frequency Cepstral Coefficients," Sensors and Materials, vol. 32, no. 10, p. 3209, Oct. 2020, doi: 10.18494/sam.2020.2860.

[31] S. M. Qaisar, S. Bahanshal, and H. Alwazani, "A Cloud Assisted Hybrid Model Based Speaker Recognition and Resource Sharing," Procedia Computer Science, vol. 163, pp. 410–416, 2019, doi:10.1016/j.procs.2019.12.123.

[32] S. Nagarajan, S. S. S. Nettimi, L. S. Kumar, M. K. Nath, and A. Kanhe, "Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales," Digital Signal Processing, vol. 104, p. 102763, Sep. 2020, doi:10.1016/j.dsp.2020.102763.

[33] I. Pires et al., "Recognition of Activities of Daily Living Based on Environmental Analyses Using Audio Fingerprinting Techniques: A Systematic Review," Sensors, vol. 18, no. 2, p. 160, Jan. 2018, doi:10.3390/s18010160.

[34] D. M. Midyanti, S. Bahri, and H. I. Midyanti, "ADALINE Neural Network For Early Detection Of Cervical Cancer Based On Behavior Determinant," Scientific Journal of Informatics, vol. 8, no. 2, pp. 283–288, Nov. 2021, doi: 10.15294/sji.v8i2.31064.