Artificial Intelligent for Human Emotion Detection with the Mel-Frequency Cepstral Coefficient (MFCC)

Anita Ahmad Kasim[1], Muhammad Bakri[2], Irwan Mahmudi[3], Rahmawati[4], Zulnabil[5]

[1]*Department of Information Technology, Faculty of Engineering, Universitas Tadulako, Indonesia*

[2]*Department of Architecture, Faculty of Engineering, Universitas Tadulako , Indonesia*

[3]*Department of Electrical engineering, Faculty of Engineering, Universitas Tadulako, Indonesia*

[4]*Department of Department of Social Sciences, Faculty of Engineering, Universitas Tadulako, Indonesia*

[5]*Study Program of Informatics Engineering, Faculty of Engineering, Universitas Tadulako, Indonesia*

[1]nita.kasim@gmail.com

Abstract - Emotions are an important aspect of human communication. Expression of human emotions can be identified through sound. The development of voice detection or speech recognition is a technology that has developed rapidly to help improve human-machine interaction. This study aims to classify emotions through the detection of human voices. One of the most frequently used methods for sound detection is the Mel-Frequency Cepstrum Coefficient (MFCC) where sound waves are converted into several types of representation. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The primary data used in this research is the data recorded by the author. The secondary data used is data from the "Berlin Database of Emotional Speech" in the amount of 500 voice recording data. The use of MFCC can extract implied information from the human voice, especially to recognize the feelings experienced by humans when pronouncing the sound. In this study, the highest accuracy was obtained when training with epochs of 10000 times, which was 85% accuracy.

Keywords: Human emotions, Voice Feature, MFCC

I. INTRODUCTION

AI in supporting human needs can be improved by allowing it to recognize emotions; with this ability, AI can provide various responses depending on the user's emotions. Emotions are inner feelings resulting from a person's reaction to experiences and events, such as emotions of fear, anger, frustration, and happiness. Emotions are quite important in our daily communications and recent years have witnessed a lot of research works to develop reliable emotion recognition systems based on various types data sources such as audio and video [1]. Speech Emotion Recognition is a challengeable task to improve human–computer interaction [2].

Various ways can be done; some studies do it through facial expressions. But in this study uses the human voice in identifying emotions. Facial expressions are not better than human voices. By studying how humans say something, important information can be extracted from speech, especially emotions. In recognizing human voices, several studies use the Linear Predictive Coding (LPC) feature extraction method, such as the research Detect Human Voice in Emotional State Using Linear Predictive Coding (LPC) With Coarse to Fine Search (CFS) Classification Based on Data Processing and feature extraction of Discrete Wavelet Transform (DWT) as in the research "Treatment Extraction and Recognition of Indonesian Vowel Voices by Gender in Real-Time" [3]-[4]. The MFCC feature extraction technique is widely used in speech recognition because it is robust, effective and simple to implement [5] . MFCC is a feature extraction that produces features or characteristics in the frame and cepstral coefficient parameters. Features are different from one another in the form of parameters. MFCC feature extraction can recognize more voice characters with non-linear voice signals while using Linear Predictive Code (LPC) only for linear ones [6]. The use of ANN in this study is because artificial neural networks can represent the learning of the human brain in classifying data. Artificial neural networks consist of a number of neurons that form several layers that are capable of processing data sent in each layer. The output of each layer is sent to the next layer. There is a nonlinear activation function that is different for each layer. This helps in the learning process

and output at every layer. Weights are related to neurons and are responsible for the overall classification process. Each Neural Network is equipped with functions that are minimized as learning progresses. The best weight is then used where the function gives the best result. The problem in this research is how to use the Artificial Neural Network algorithm in classifying emotions based on the MFCC feature extraction results' attributes and how accurate the MFCC method and the ANN algorithm are in recognizing human emotions? Research in the field of pattern recognition has been carried out in various areas, for example, in facial disease detection, image identification, and voice identification [7]-[8]. The research entitled Detecting Human Voices in Emotional Conditions Using Linear Predictive Coding (LPC) With Coarse to Fine Search (CFS) Classification Based on Data Processing. According to this study, the challenge in voice recognition is detecting the speaker's emotions. If you only observe what is being said without paying attention to how the words are pronounced, essential aspects of the speech may be lost, and misunderstandings may occur. This study acquired training data in sound frequencies from a computer or laptop voice recorder with the "WAV" format. Then tested to detect human emotions in 4 types, namely happy, angry, sad, and surprised [9]. The similarity of this research with the author's research is to recognize the emotions of the speaker's voice as an object.

However, the difference is that this study uses the LPC method to extract voice features from speakers who have a weakness, namely its time variant nature so that the determination of the beginning and end of the recording data must be precise, which requires the voice recording data set to have the same duration. The research entitled Implementation and Analysis of Emotion Detection Simulations Through Voice Recognition Using Mel-frequency Cepstrum Coefficient (MFCC) and Hidden Markov Model (HMM) Based on Internet of Things (IoT), have purpose to develop an existing emotion detection research [10].The accuracy is quite good, but the output produced by the machine is still in the form of emotional class data and is difficult for ordinary users to understand. So, this research makes an IOT-based indicator tool that can receive input from the machine. It is presented in the form of a light indicator representing each emotion based on the emotional reference that has been taugdevicehe machine. The similarity of this research with the author's research is the method used for the feature extraction stage, namely, using MFCC. The difference is in the classification algorithm; this study uses the HMM algorithm, while the author's research uses ANN. The study entitled Recognition of Human Emotion Patterns Based on Speech Using Mel-Frequency Cepstral Coefficients (MFCC) Feature Extraction. Emotion recognition is complex because of the differences in customs and dialects in different ethnicities, regions, and communities [10]-[11].

This problem also becomes difficult because of an objective assessment; emotion is an event that occurs in the human subconscious. This study aimed to determine the pattern of emotion recognition based on feature extraction from speech. Just like the feature extraction method used by the author, the feature extraction method used in this study is MFCC, which is a feature extraction method that approximates the human hearing system. The difference between this study and the author's research is in the stage after MFCC feature extraction. The author uses the Artificial Neural Network algorithm to classify the training data as emotional labels. However, this study only describes sound waves before and after feature extraction. Then from the sound wave diagram image, the patterns that characterize an emotion are studied, such as happy, bored, neutral, sad, and angry.

II. METHOD

There are two data collections in this study, namely primary data collection and secondary data collection. Primary data is obtained by directly recording the voice of a human using a voice recorder by expressing certain emotions and different genders. The data to be collected is 339 voice recording data consisting of 127 data labeled "angry", 71 data labeled "happy", 79 data labeled "neutral", and 62 data labeled "sad". The recorded data will be divided into two blocks, namely 80% training data (271 data) and 20% test data (68 data). The secondary data used is data from the "Berlin Database of Emotional Speech" in the amount of 500 voice recording data. Spoken by ten actors of different ages and genders with file name code information [12] :

1. 03 - male, 31 years old
2. 08 - female, 34 years
3. 09 - female, 21 years
4. 10 - male, 32 years
5. 11 - male, 26 years
6. 12 - male, 30 years
7. 13 - female, 32 years
8. 14 - female, 35 years
9. 15 - male, 25 years
10. 16 - female, 31 years

Also, with different text. With the filename code in Table I.

TABLE I
SOUND TEXT

code	text (german)	try of an english translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the frigde.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

It consists of happy, angry, anxious, afraid, bored, disgusted, and neutral emotions, with the filename code in Table II. The data collection technique used by the author was to ask students of the Department of Information Technology at Tadulako University to record their voices when saying a sentence that contains certain emotions.

The data that will be taken is 339 voice recording data consisting of 127 data with the label "angry," 71 data with the title "happy," 79 data with the brand "neutral," and 62 data with the label "sad." The recorded data will be divided into two blocks, namely 80% training data (271 data) and 20% test data (68 data). The method used in this study can be seen in Fig. 1.

MFCC is a feature extraction method that is widely used in the field of speech recognition. Introduced by Davis and Marmelstein in the 1980s, and has been the best ever since Davis & Mermelstein in 1980. Prior to the introduction of MFCC, its predecessor Linear Predictive Coding (LPC) was the main feature extraction method in Automatic Speech Recognition (ASR). Use of effective features for emotion recognition is a step towards better accuracy[13][14]. For the implementation of software in development is divided into two sides,

TABLE II
FILE NAME CODE

letter	emotion (english)	letter	emotion (german)
A	*Anger*	W	*Ärger (Wut)*
B	*Boredom*	L	*Langeweile*
D	*Disgust*	E	*Ekel*
F	*anxiety/fear*	A	*Angst*
H	*Happiness*	F	*Freude*
S	*Sadness*	T	*Trauer*

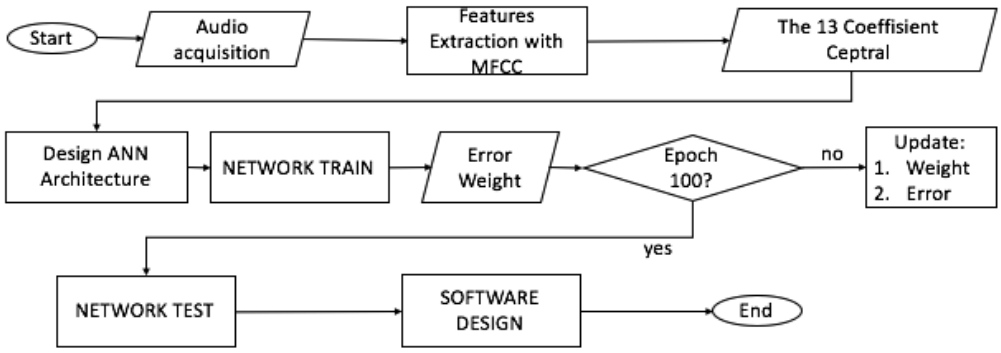

Fig. 1 Research method

namely the backend and frontend. The backend uses the programming language *Python* to run the ANN algorithm, the Integrated Development Environment (IDE) used is *PyCharm v2020.1.1*. Then the frontend uses the programming language *Javascript* which is supported by the *ReactJS* framework, the IDE used is *Visual Studio Code v1.49.2*. The following are the stages of the MFCC feature extraction process in Fig. 2.

A. DC Removal

DC Removal aims to normalize voice samples by removing unnecessary data in the next process. This is done by calculating the average of the voice sample data and then subtracting the value of each voice sample with that average value. The DC Removal process is shown in (1).

$$y[n] = x[n] - \bar{x}, \ 0 \leq n \leq N-1 \qquad (1)$$

where,
y[n] = DC Removal sample sound signal
x[n] = Sample original sound signal
x = Average value of the original sound signal sample
N = Length of sound signal

B. Pre-Emphasize

Pre-Emphasize is done to reduce noise in the input sound, so that the accuracy of the feature extraction process can be increased. This filter maintains high frequencies in a spectrum that are generally eliminated during the sound production process. The pre-emphasize process is shown in (2).

$$y[n] = s[n] - a.s[n-1], \ 0.9 \leq a \leq 1.0 \qquad (2)$$

where,
y[n] = Signal result of pre-emphasize filter
s[n] = Signal before pre-emhasize filter
a = alpha value

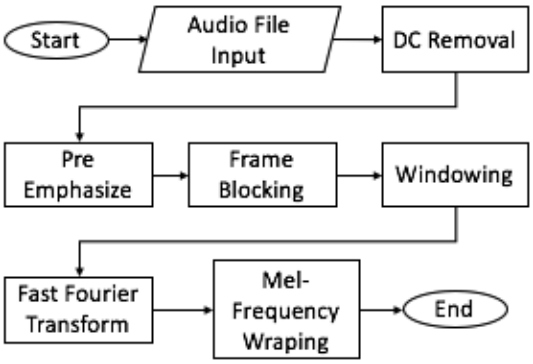

Fig. 2 MFCC feature extraction process

C. Frame blocking

Frame blocking is analyzing speech signals into frames. Each frame is represented by a single feature vector depicted in the averaged spectrum [15]-[16]. The sound signal is constantly changing, so it is necessary to divide it into several frames on a time scale between 20 and 40ms (the default is 25ms). If it is shorter than that, the spectral sample obtained is not good enough, and if it is longer the signal changes too much throughout the frame. For example, an audio file with a sample rate of 16kHz and assuming the framing timescale is 25ms (by default), this means the frame length for a 16kHz signal is 0.025 * 16000 = 400 samples. The frame step is usually around 10ms (160 samples), which may overlap the frame. The first 400 sample frames start at sample 0, then the next 400 sample frames start at sample 160. And so on until the end of the sound file is reached. The following is an illustration of frame blocking in Fig. 3. The frame blocking process is shown in (3).

$$number\ of\ frames = Ts/M \qquad (3)$$

where,
Ts = Voting duration (ms)
M = Frame length (ms)

D. Windowing

After the frame blocking process, the voice signal is divided into several overlapping frames and causes discontinuity. The windowing process is carried out to reduce the discontinuity or gap. It is calculated by (4) and (5).

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (4)$$

$$x(n) = x_i(n)w(n) \qquad (5)$$

where,
x(n) = The sample value of the windowing signal
$x_i(n)$ = Signal sample value from signal frame to i
w(n) = window function

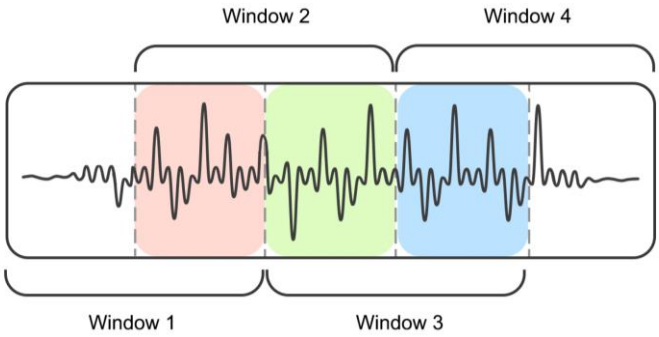

Fig. 3 Frame blocking

D. Fast Fourier Transform (FFT)

Inspired by the human cochlea (an organ in the ear) which can vibrate in different positions depending on the frequency of the incoming sound. Nerves tell the brain that there are certain frequencies. This step identifies which frequencies are in the frame. FFT is the stage for converting each frame consisting of N samples from the time domain into the frequency domain. The first step to interpreting the FFT is to calculate the frequency value of each sample. The FFT process is shown in (6).

$$f(n) = \sum_{K=0}^{N-1} y_k e^{-\frac{2\pi kn}{N}}, n = 0, 1, 2, \dots, N-1 \quad (6)$$

where,
f(n) = Frequency
N = Number of samples in each imaginary frame
k = 0, 1, 2, …, (N-1)

E. Mel-Frequency Warping

The identified frequencies still contain a lot of unnecessary information. Mel filterbank filters the voice signal that has been converted into the form of a frequency domain. A filterbank is a system that divides the input signal into signal analysis pools corresponding to different regions of the spectrum. Typically, the region of the spectrum given by collective signal analysis spans the entire range of sound audible to human hearing. The filter bank used in the MFCC method is Mel-Filterbank. The mel-filterbank consists of a series of overlapping triangular windows that will filter out N samples of the signal, as can be seen in Fig. 4. .

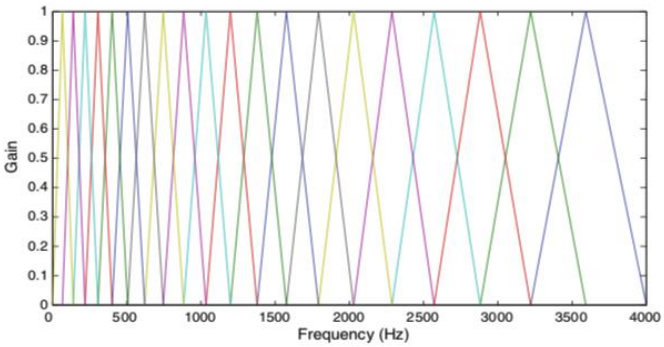

Fig. 4 Mel filter bank

When the frequency is high, the filterbank becomes wider. Mel scale is useful for knowing how wide the filter bank space is needed. The equation for converting from frequency to Mel Scale is shown in (7).

$$Mel = 2595 * log10\left(1 + \frac{f}{700}\right) \quad (7)$$

where,
Mel = Mel value (converted from frequency value)
f = Frequency

III. RESULT AND DISCUSSION

Voice acquisition is made by recording the human voice using a smartphone device. The recorded voices are four emotions: happy, sad, angry, and neutral. The following is an overview of the spectrum of the voice signal in Fig. 5.

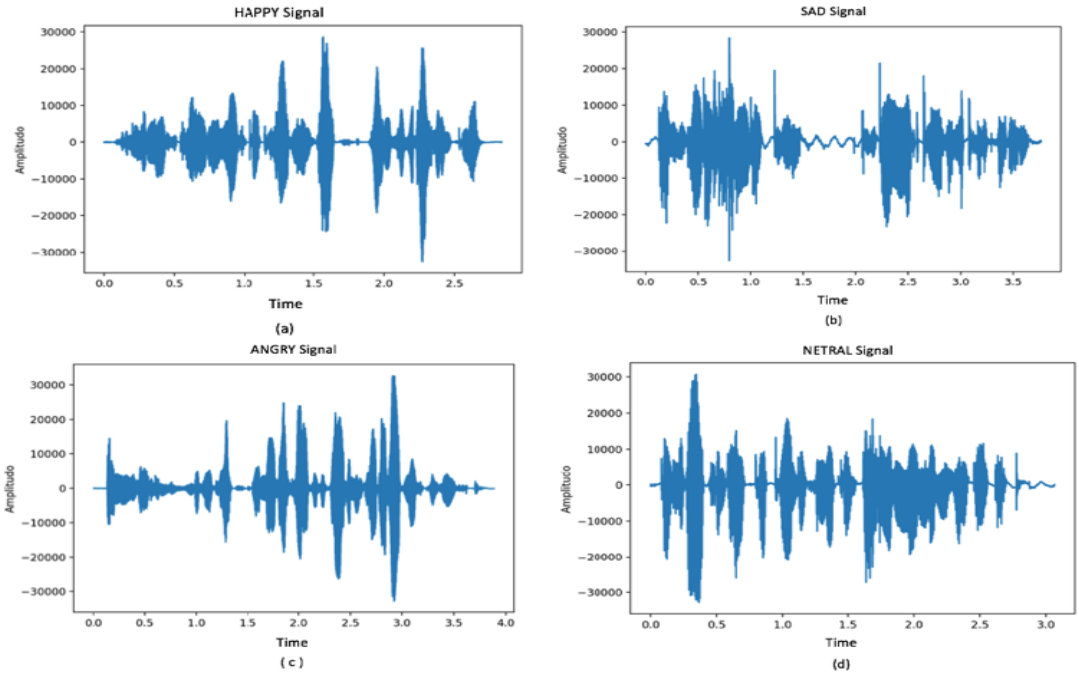

Fig. 5 Signal spectrum (a) happy, (b) sad, (c) angry, (d) neutral

The next stage is feature extraction, part of the Pattern Recognition technique, which aims to retrieve or extract 13 unique values from an object that distinguishes it from other things. This study uses MFCC feature extraction, where MFCC is a method for calculating cepstral coefficients based on critical frequency variations in the human hearing system. This value will be used in classification using an artificial neural network. An overview of MFCC features can be seen in Fig. 6.

The parameters used for testing the Artificial Neural Network algorithm are 13 coefficients. This coefficient is obtained from the results of feature extraction using MFCC. The author chose 13 coefficients because 13 is the most popular coefficient in the range of 10-20 coefficients in the field of speech recognition. From the results of testing the Artificial Neural Network algorithm which was carried out using 271 training data and 68 human speech voice test data, it was found that the classification results had a good level of accuracy in the voice file upload method and had a low level of accuracy in the direct voice recording method of test data.

The low accuracy of the direct recording method occurs due to several factors, including; the difference in the equipment used in the training data and the test data, the absence of pre-processing voice data when recording directly causes noise that is too high so that the MFCC is not good enough in processing the data.

From the results of the confusion matrix test, facts are obtained as shown in Table III.

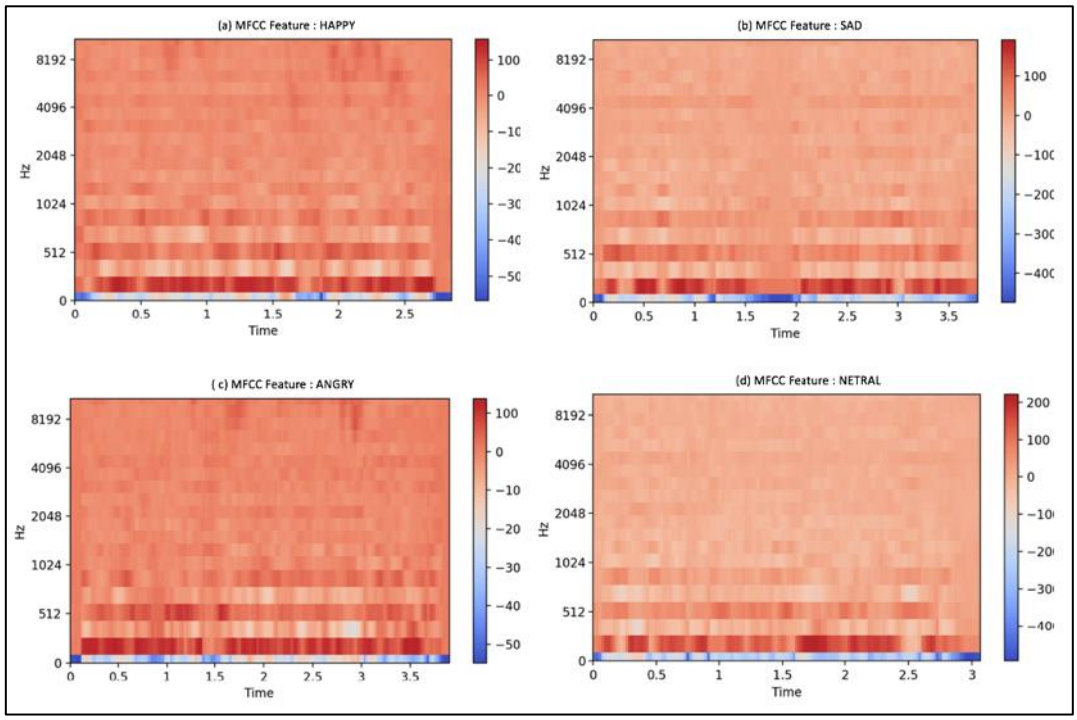

Fig. 6 Visualization of MFCC features (a) happy, (b) sad, (c) angry, (d) neutral

TABLE III
CONFUSION MATRIX FOR HUMAN EMOTION DETECTION

		Output Class			
	Class of emotion	**Angry**	**Happy**	**Neutral**	**Sad**
Target Class	**Angry**	True Angry	False Happy	False Neutral	False Sad
	Happy	False Angry	True Happy	False Neutral	False Sad
	Neutral	False Angry	False Happy	True Neutral	False Sad
	Sad	False Angry	False Happy	False Neutral	True Sad

The equation for calculating the accuracy and the error rate, shown in (8) and the error rate, shown in (9).

$$Accuracy = \frac{TrueAngry+TrueHappy+TrueNeutral+TrueSad}{\sum All\ The\ Data\ of\ Class\ of\ emotion} \times 100 \quad (8)$$

$$Error\ Rate = 100\% - Accuracy \quad (9)$$

In the first experiment, which was 100 epochs, the total correct data with classification results that were also correct or true positive was 35 and the total data that was not true positive was 33, so the calculation was carried out to obtain an accuracy of 51%. Then precision is calculated for each class and then averaged so that it gets a value of 35%, as well as recall which gets a value of 42%. Then the error rate is the difference between 100% and accuracy, which is 49%. In this experiment, the model that was built was still weak in recognizing emotions because the amount of training carried out was still very low, namely 100 times of training.

In the second experiment with 1000 epochs, the total data with true positive obtained increased significantly compared to the first experiment, which was 52 data with true positive classification results. Meanwhile, the total data that were not true positive also directly decreased to 16 data. The accuracy obtained from the calculation is 76%. The average precision in this experiment is slightly lower than the accuracy, which is 74%, then the recall average is 72% and the error rate is only 24%.

In the third experiment with 10000 epochs achieved the highest accuracy of 85%. This means that the learning built on the model succeeded in making the model acquire knowledge like humans as learning was carried out. The result of the experiment is show in Table IV, Table V and Table VI.

TABLE IV
CONFUSION MATRIX FOR HUMAN EMOTION DETECTION (epoch=100)

Epoch = 100		**Output Class**			
	Class of emotion	**Angry**	**Happy**	**Neutral**	**Sad**
Target Class	**Angry**	24	0	0	0
	Happy	14	0	0	0
	Neutral	5	0	11	0
	Sad	14	0	0	0
Accuracy= (24+11)/(24+14+5+11+14) x 100 = 51,4%					
Error Rate= 100%-51,4%=48,6%					

TABLE V
CONFUSION MATRIX FOR HUMAN EMOTION DETECTION (epoch=1000)

Epoch = 1000		**Output Class**			
	Class of emotion	**Angry**	**Happy**	**Neutral**	**Sad**
Target Class	**Angry**	19	5	0	0
	Happy	2	11	1	0
	Neutral	0	0	14	2
	Sad	0	0	0	14
Accuracy= (19+11+14+14)/86 x 100 = 76,4%					
Error Rate = 100%-76,4%=23,6%					

TABLE VI
CONFUSION MATRIX FOR HUMAN EMOTION DETECTION (epoch=1000)

Epoch=10000		**Output Class**			
	Class of emotion	**Angry**	**Happy**	**Neutral**	**Sad**
Target Class	**Angry**	23	1	0	0
	Happy	12	1	1	0
	Neutral	1	0	14	1
	Sad	0	0	0	14
Accuracy= (23+1+14+14)/68x100%=85,2%					
Error Rate =100%-85,2%=14,8%					

In the fourth experiment the author implemented the stop condition for learning, namely with an accuracy threshold of 85%, and it turned out that the learning process reached 85% accuracy in the 3000th epoch, so that 3000 epoch was the optimal number of repetitions to achieve maximum accuracy from this system because when the author changed the termination conditions at 90% accuracy, the learning process that is carried out never stops because the accuracy is stuck at 85% and can even be reduced due to over training. The more trained, the higher the accuracy of the model in classifying. However, the amount of learning carried out does not absolutely make the model have good knowledge. When the author changes the stop condition to 90% accuracy, the learning process never stops. This is caused by many factors such as data quality, data quantity, and others. Just like humans, if they are trained with knowledge with low categories and small amounts, then no matter how many people are trained their knowledge will be stuck at one point and will not increase. So, what must be done for better knowledge is to improve the quality and quantity of that knowledge.

The training phase works through the process iteratively using a set of training data, comparing the expected value of the network with each training data. In each iteration process, the weights in the network are updated to minimize the error by reducing the expected or predicted value with the reality or output value. The weight modification is done backward, from the output layer to the first layer of the hidden layer, so this method is called backpropagation.

IV. CONCLUSION

This system can classify angry, happy, neutral, and sad emotions by utilizing the MFCC feature extraction and the ANN algorithm. The use of MFCC can extract implied information from the human voice, especially to recognize the feelings experienced by humans when pronouncing the sound. In this study, the highest accuracy was obtained when training with epochs of 10000 times, which was 85% accuracy. From the results of the confusion matrix test on the first try is 100 epochs, the accuracy is only 51% but the second experiment with 1000 epochs, the built model managed to get a good accuracy of 76%, the third experiment with 10000 epochs achieved the highest accuracy of 85%. This means that the learning built on the model has succeeded in making the model acquire knowledge like humans as learning is carried out. The more trained, the higher the accuracy of the model in classifying. The author has trained more than 10000 epochs, but the accuracy results obtained are only up to 85% and can't be more than that. This is caused by many factors, such as data quality, data quantity, and others. Just like humans, if trained with knowledge with low categories and small amounts, then no matter how much the human being is trained, his learning will not increase.

ACKNOWLEDGEMENT

This work was supported and funded by Universitas Tadulako in the scheme “Penelitian Penugasan” in 2021 with Grant Number 430. ab/UN28.2/PL/2021 on April 26th, 2021.

REFERENCES

[1] Md. Z. Uddin and E. G. Nilsson, “Emotion recognition using speech and neural structured learning to facilitate edge intelligence,” *Eng Appl Artif Intell*, vol. 94, p. 103775, 2020, doi: 10.1016/j.engappai.2020.103775, access time November 1st, 2022

[2] S. Lalitha, D. Geyasruti, R. Narayanan, and S. M, “Emotion Detection Using MFCC and Cepstrum Features,” *Procedia Comput Sci*, vol. 70, pp. 29–35, 2015, doi: 10.1016/j.procs.2015.10.020, access time November 1st, 2022

[3] I. Rahmawanthi, J. Raharjo, and A. Rusdinar, “Detection Human Voice in Emotion Condition Using Linear Predictive Coding (LPC) with Coarse to Fine Search (CFS) Classification Based on Data Processing,” in *e-proceeding of engineering*, 2019, pp. 656–663, access time October 31th, 2022

[4] R. Via Yuliantari, R. Hidayat, and O. Wahyunggoro, “Ekstrasi Ciri dan Pengenalan Suara Vokal Bahasa Indonesia Berdasarkan Jenis Kelamin Secara Real Time,” in *Prosiding SNATIF3*, 2016, pp. 1–6, access time October 31th, 2022

[5] P. Thu and Z. Tun, “Audio Feature Extraction Using Mel-Frequency Cepstral Coefficients,” vol. 2, p. 12, 2020, doi: 10.5281/zenodo.1342401, access time December 11th, 2022

[6] H. Heriyanto and D. A. Irawati, “Comparison of Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction, With and Without Framing Feature Selection, to Test the Shahada Recitation,” *RSF Conference Series: Engineering and Technology*, vol. 1, no. 1, pp. 335–354, Dec. 2021, doi: 10.31098/cset.v1i1.395, access time December 11th, 2022

[7] A. A. Kasim, R. Wardoyo, and A. Harjoko, “Batik classification with artificial neural network based on texture-shape feature of main ornament,” *International Journal of Intelligent Systems and Applications*, vol. 9, no. 6, pp. 55–65, Jun. 2017, doi: 10.5815/ijisa.2017.06.06, access time October 31th, 2022

[8] A. A. Kasim and Agus Harjoko, "Klasifikasi Citra Batik Menggunakan Jaringan Syaraf Tiruan Berdasarkan Gray Level Co-Occurrence Matrices (GLCM) Agus Harjoko," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2014, pp. C7-C–13, access time October 31th, 2022

[9] N. J. Nalini and S. Palanivel, "Music emotion recognition: The combined evidence of MFCC and residual phase," *Egyptian Informatics Journal*, vol. 17, no. 1, pp. 1–10, 2016, doi: 10.1016/j.eij.2015.05.004, access time November 1st, 2022

[10] A. A. Sundawa, A. G. Putrada, and N. A. Suwastika, "Implementasi dan Analisis Simulasi Deteksi Emosi Melalui Pengenalan Suara Menggunakan Mel-Frequency Cepstrum Coefficient dan Hidden Markov Model Berbasis IOT," in *e-Proceeding of Engineering*, 2019, pp. 1–8, access time October 31th, 2022

[11] S. Helmiyah, A.Fadlil, and A.Yudhana, "Pengenalan Pola Emosi Manusia Berdasarkan Ucapan Menggunakan Ekstraksi Fitur Mel-Frequency Cepstral Coefficients (MFCC)," *Cogito Smart Journal*, vol. 4, no. 2, pp. 372–381, 2018, doi: 10.31154/cogito.v4i2.129.372-381, access time October 31th, 2022

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech*, 2005. doi: 10.21437/Interspeech.2005-446, access time October 31th, 2022

[13] M. N. Mohanty and H. K. Palo, "Child emotion recognition using probabilistic neural network with effective features," *Measurement*, vol. 152, p. 107369, 2020, doi: 10.1016/j.measurement.2019.107369, access time November 1st, 2022

[14] Y. R. Prayogi, "Modifikasi Metode MFCC untuk Identifikasi Pembicara di Lingkungan Ber-Noise," *JOINTECS) Journal of Information Technology and Computer Science*, vol. 4, no. 1, pp. 2541–3619, 2019, doi: 10.31328/jointecs.v4i1.999, access time November 1st, 2022

[15] Ranny, I.S. Suwardi, T.L.E. Rajab, and D.P. Lestari, "Study of Sound Processing and Application on Information Technology," *JUITA*, vol. VII, no. 1, pp. 1–10, 2019, doi: 10.30595/juita.v7i1.3491, access time November 1st, 2022

[16] Heriyanto, S. Hartati, and A.E. Putra, "Ekstraksi Ciri Mel Frequency Cepstral Coefficient (MFCC) dan Rerata Coefficient Untuk Pengecekan Bacaan Al-Qur'an," *Telematika*, vol. 15, no. 02, pp. 99–108, 2019, doi: 10.31315/telematika.v15i2.3123, access time November 1st, 2022