Contemporary Mathematics and Applications
Vol. 7, No. 1, 2025, pp. 36-45.

Voice-Based Emotion Identification Based on Mel Frequency Cepstral Coefficient Feature Extraction Using Self-Organized Maps and Radial Basis Function

Asrivatun Nikmah[1], Auli Damayanti[2], Edi Winarko[3]

[1,2,3]Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga
Jl. Mulyorejo, Surabaya

[2]Corresponding author: aulid@fst.unair.ac.id

Abstract. Speech recognition is one of the most popular research fields, one of which is about emotion identification. Voice-based emotion identification is carried out to determine the pattern of emotions using the depth analysis mechanism of voice signal development and feature extraction that carries the emotional characteristic parameters of the speaker's voice. Furthermore, the emotional characteristics of the speaker's voice are classified using an artificial neural network method to recognize patterns. In this study, emotion identification from voice signal data is classified into angry, sad, happy, and neutral emotions. The stages of voice-based emotion identification, including the feature extraction stage using the mel frequency cepstral coefficient, produce coefficient values, which will be used in the identification stage using the Self Organized Maps method on the Radial Basis Function.

Keywords: *Voice Based Emotion Identification; Artificial Neural Network*; *Mel Frequency cepstral Coefficient*; *Self Organizied Maps; Radial Basis Function.*

1 Introduction

Emotion is a complex phenomenon influenced by biological, social, and cognitive factors [1]. Furthermore, emotions can be defined as reactions to external and internal influences on an individual by a state or feeling that can describe a person's behavior or actions. Emotions are one of the important aspects of human life because they influence human intentional behavior [2]. In the advancement of more modern technology today, where human-machine interaction is no longer a new phenomenon, machines will provide better responses if they can recognize user emotions [3].

Voice recognition in the field of research has become one of the most sought-after areas. The goal is to improve the quality of interaction between humans and voice-based machines (computing) [4]. Emotion recognition based on voice is carried out to identify emotional patterns using a depth analysis mechanism of voice signal development and feature extraction that carries the emotional characteristic parameters of the speaker's voice [5]. According to [6], voice recognition generally has three main parts, namely first,

Received: January 17, 2025
Accepted: February 27, 2025

feature extraction to represent the voice signal in a characteristic manner such as pitch, signal coefficients, and others. The second is the modeling scheme to determine the characteristics of the voice signal using a statistical approach. And third is the classification scheme to recognize patterns from the audio signal.

There are several feature extraction methods that can be performed for audio signal processing, such as linear predictive coding (LPC), Fast Fourier Transform (FFT), mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) cepstral coefficients, and so on. MFCC is a feature extraction based on perception that closely resembles human voice. This method is often used by previous researchers due to its ability to closely resemble the human voice [7].

The Self-Organizing Maps (SOM) network was first introduced by Prof. Teuvo Kohonen in 1982 [8]. In the SOM network, the layer containing neurons will organize itself based on certain input values into groups known as clusters. The cluster that has the weight vector most closely matching the input pattern (the shortest distance) becomes the winning vector [9]. Radial Basis Function (RBF) networks are a method of artificial neural network learning that simultaneously applies supervised and unsupervised learning algorithms [8].

Based on the journal by [3], several previous studies on emotion recognition based on voice have been conducted. Research using the Bayesian method was conducted in [10], classifying emotions into 6 categories: neutral, angry, happy, sad, empathetic, and panicked. With the use of 62 features, a result of 80.46% was achieved in recognizing happy and sad emotions, and 62% in recognizing neutral, empathetic, angry, and panicked emotions. The research conducted by [11] discusses emotion recognition for script-based acting guidance by classifying emotions into 8 categories, namely anger, despair, empathy, happiness, sadness, pride, and joy, achieving an accuracy rate of 57.1% using 18 features.

In journal [12], research [5] used MFCC feature extraction and the K-Nearest Neighbour (K-NN) algorithm, achieving an accuracy rate of 86.02%. Research [13] conducted a study using MFCC feature extraction and the Gaussian Mixture Model (GMM) algorithm, achieving the highest accuracy of 92% and 72% for angry emotions using the K-NN algorithm. A similar study was conducted by [14] using MFCC feature extraction and the Hidden Markov Model (HMM) algorithm, resulting in an average accuracy of 86.66%.

In this article, emotion identification based on voice will be conducted to determine the characteristics of emotional patterns, namely angry, sad, happy, and neutral emotions. The feature extraction method used in this research is the Mel Frequency Cepstral Coefficient (MFCC), due to its capability to closely resemble human voice. The classification of the results from the feature extraction process uses the Radial Basis Function (RBF) network algorithm by applying Self-Organizing Maps to determine the cluster centers. This RBF network is used because it has a unique learning pattern by using both supervised and unsupervised learning methods simultaneously.

2 Mel Frequency Cepstral Coefficient Feature Extraction

Mel-Frequency Cepstrum Coefficient (MFCC) is a feature extraction method in speech recognition. MFCC has good resistance to noise. MFCC takes the average or mean logarithmic value of the spectrum after the Mel FilterBank and frequency wrapping. The goal is to sharpen the differences in patterns so that it can facilitate the process of emotion pattern recognition. The stages of MFCC feature extraction are as follows. [7]

2.1 Signal Cutting

Audio signals contain important attributes in the time domain that carry significant information such as amplitude and others. However, these attributes change over time, so it is necessary to estimate them periodically by segmenting the signal.

2.2 Windowing

Windowing aims to reduce the effects of discontinuities from signal truncation when the signal is transformed to the frequency domain. The Hamming window coefficients (w) are shown in Equation (1) with the window function (y) shown in Equation (2) where N is the frame length.

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & , 0 \leq n \leq N-1 \\ 0 & , otherwise \end{cases}, \tag{1}$$

$$y(n) = x(n).w(n) \tag{2}$$

2.3 Fast Fourier Transform

Fast Fourier Transform (FFT) is a method for converting signals from the time domain to the frequency domain. FFT applies the divide and conquer algorithm by breaking down the DFT algorithm into problems with even and odd indices, thereby reducing the time complexity to solve the algorithm more quickly. The FFT function (X) is shown in Equation (3).

$$|X(k)| = \left|\sum_{n=0}^{\frac{N}{2}-1} y(2n)e^{-\frac{2\pi kni}{N/2}} + e^{-\frac{2\pi ki}{N/2}}\sum_{n=0}^{\frac{N}{2}-1} y(2n+1)e^{-\frac{2\pi kni}{N/2}}\right| \tag{3}$$

2.4 Mel Filter Bank

The human voice does not follow a linear frequency scale, so it is necessary to convert it to a mel frequency scale whose representation approximates the human voice. The mel filterbank process is carried out by converting the linear frequency scale (f) into the mel frequency scale (B) to obtain the filterbank boundaries using Equation (4). Then the boundaries of the filter bank (l) are divided by the number of filters using Equation (5). The number of filters (m) used is usually between 24 to 40 filters.

$$B(f) = 1125 * ln\left(1 + \frac{f}{700}\right), \tag{4}$$

$$l(m) = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_{min}) + m\frac{B(f_{max}) - B(f_{min})}{M+1}\right). \tag{5}$$

The filter method that is often used is the triangular filter bank (H) as shown in Equation (6).

$$\mathrm{H(k)} = \begin{cases} 0, & \mathrm{k} < f(m-1), \\ \frac{\mathrm{k-f(m-1)}}{\mathrm{f(m)-f(m-1)}} & \mathrm{f(m-1) \le k \le f(m)}, \\ \frac{\mathrm{f(m+1)-k}}{\mathrm{f(m+1)-f(m)}} & \mathrm{f(m) \le k \le f(m+1)}, \\ 0, & \mathrm{k} > f(\mathrm{m}+1). \end{cases} \tag{6}$$

Next, to generate the mel filter bank values (D) by multiplying the FFT result with the triangular filter bank as shown in Equation (7).

$$\mathrm{D(m)} = \log\left(\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{X(k)^2 . H_m(k)}\right). \tag{7}$$

2.5 Discreate Cosine Transform

Discrete cosine transform (DCT) is used to convert the signal back into the time domain. The obtained value is called the Cepstral Coefficient or MFCC coefficient (C). MFCC feature extraction in speech recognition usually only uses the first 13 cepstral coefficients (M). The value of the first coefficient is not used because it contains little information.

$$\mathrm{C(k)} = \mathrm{w(k)} \sum_{\mathrm{m=0}}^{\mathrm{M-1}} \mathrm{D(m)} \cos\left(\frac{\pi \mathrm{k}}{\mathrm{M}}\left(\mathrm{m} + \frac{1}{2}\right)\right), \tag{8}$$

$$w(k) = \begin{cases} \frac{1}{\sqrt{M}}, & k = 1, \\ \sqrt{\frac{2}{M}}, & 2 \le k \le M. \end{cases} \tag{9}$$

2.6 Liftering

Liftering aims to refine the extracted signal by emphasizing the low-order coefficients that are sensitive to spectral slope and the high-order coefficients that are sensitive to noise with a lifter coefficient. The liftering function (C') is shown in Equation (10) [15].

$$\mathrm{C'(k)} = \mathrm{C(k)}\left(1 + \frac{\mathrm{L}}{2}\sin\left(\frac{\pi\mathrm{k}}{\mathrm{L}}\right)\right) \tag{10}$$

with

C′(k) = the value of the k-th liftering coefficient; k = 1,2, … , M
L = lifter coefficient.

3 Self Organizing Maps

The self-organizing maps (SOM) neural network was first introduced by Prof. Teuvo Kohonen in 1982. In the SOM network, neurons in a layer will organize themselves based on specific input values within a cluster. The training algorithm for the SOM network is as follows [8].

a. Initialize the weight parameters (w_{ij}), neighborhood parameters (R), and learning rate parameters (α).
b. As long as the stopping condition is not met, repeat steps c - h.
c. For each input data x, do steps d – f.
d. For each index j, calculate the distance value (D) of the input data (x) and the weight with Equation (11).

$$D(j) = \sqrt{\sum_i \left(x_{ij} - w_j\right)^2} \quad (11)$$

e. Determine the index J as a class by determining the minimum distance.
f. For each unit at index j, update the weights with Equation (12)

$$w_{ij}(new) = w_{ij}(old) + \alpha * h * \ [x_i - w_{ij}(old)] \quad (12)$$

g. Update the learning rate (α) and neighbourhood values with Equation (13) and (14)

$$\alpha(new) = \mathrm{\alpha}(old) * \exp\,(-\mathrm{t}/\mathrm{T}_1) \quad (13)$$

$$h(new) = \mathrm{h}(old) * \exp\left(\frac{\mathrm{t}}{\mathrm{T}_2}\right) \quad (14)$$

where t is the current iteration value, T_2 is the maximum iteration value.
h. Calculate the network error value and the iteration value increases by one for each iteration.

At the end of the SOM algorithm, the cluster center (centroid) will be obtained, namely the optimal w_{ij} weight values.

4 Radial Basis Function

Radial Basis Function (RBF) Neural Network is one type of learning in the multilayer artificial neural network method that uses a hybrid learning method in its training. RBF network learning consists of supervised learning and unsupervised learning. The RBF network consists of an input layer, a hidden layer with its activation function being a radial basis function and an output layer with a purelin/identity activation function [16]. The RBF training algorithm is as follows [17].

a. Calculating the value of the radial basis function. The function that is often used is the Gaussian function (φ) shown in the Equation (15).

$$\varphi(x) = \exp\left(-\frac{\|x_{ik}-c_k\|}{2\sigma^2}\right) \tag{15}$$

b. Constructing a Gaussian matrix $(\boldsymbol{G})$. The Gaussian matrix is formed from the activation values of the radial basis functions. The constructed matrix is shown as follows in Equation (16).

$$\boldsymbol{G} = \begin{pmatrix} \varphi_{11} & \cdots & \varphi_{1k} & b \\ \varphi_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \varphi_{n1} & \cdots & \varphi_{nk} & b \end{pmatrix} \tag{16}$$

with n is the number of input data or the number of units in the input layer and k is the number of clusters. The last column is the bias value.

c. Calculating the weight vector($\mathbf{w}$) of the hidden layer neurons to the output layer by multiplying the pseudoinverse of the Gaussian matrix and the target vector $(\mathbf{t})$ as shown in the following Equation (17) and (18).

$$\mathbf{w} = (\mathbf{G^T G})^{-1}.\mathbf{G^T}.\mathbf{t} \tag{17}$$

$$\mathbf{w} = (\mathrm{w}_1\mathrm{w}_2 \dots \mathrm{w}_\mathrm{n}\ \mathrm{b})^{\mathbf{T}} \tag{18}$$

Where $\mathbf{G^T}$ is transpose of Gaussian matrix and $\boldsymbol{t}$ is a column matrix containing target values.

In RBF network training, the values obtained are the weight matrix and bias values. In the neuron layer towards the output

5 Discussion and Result

The data used in this study were taken from The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database of 100 voice recording files in .wav format, in the form of happy, sad, angry, and neutral emotions, each with 25 files. The data was divided into 80 training data and 20 validation test data with the same number of emotions. Voice-based emotion identification was designed using the Java programming language on the NetBeans IDE 8.2 software.

The emotion identification process begins with reading the voice recording data and then taking the feature extraction value using MFCC in the form of the MFCC coefficient value. On the feature extraction data, the training process was carried out using the SOM and RBF algorithms. The unsupervised learning stage on the RBF network uses the SOM algorithm, which is to determine the cluster center used in the hidden layer. In this study, the parameter variations used in the training process are the number of filters, the number of MFCC coefficients, the learning rate value, the maximum iteration value, the maximum epoch value, and the error limit value. The results of the training process and testing of training data can be seen in Table 1. The results of the tests carried out on the validation test data with several variations of the best parameters from the SOM-RBF network training results are presented in Table 2 and Table 3. Table 2 is the result of the validation test taken from the training parameters with the smallest MSE error value.

Table 1. The Result of Training and Testing on Training Data

Number of MFCC Coefficients	**Num. of Filter**	**α**	**Max. Iterations**	**Result of Testing on Training Data Latih**	
				MSE	**Accuration**
10	26	0,04	100	0,04254	95
			300	0,04236	92
			500	0,0435	95
		0,05	100	0,04192	96
			300	0,04351	96
			500	0,04374	95
		0,06	100	0,0437	96
			300	0,04391	95
			500	0,04402	95
	40	0,04	100	0,04477	91
			300	0,04439	93
			500	0,04481	93
		0,05	100	0,04524	92
			300	0,04683	92
			500	0,04236	95
		0,06	100	0,04439	92
			300	0,0413	96
			500	0,04423	93
13	26	0,04	100	0,04115	96
			300	0,04127	96
			500	0,04085	97
		0,05	100	0,04119	97
			300	0,04082	96
			500	0,04155	97
		0,06	100	0,04088	95
			300	0,04126	96
			450	0,04086	96
	40	0,04	100	0,041	96
			300	0,04168	93
			500	0,04206	92
		0,05	100	0,04095	95
			300	0,04124	97
			450	0,04116	95
		0,06	100	0,04167	95
			300	0,04086	96
			500	0,04093	93

Table 2. Result of Validation Test From Smallest MSE

No.	Number of MFCC Coefficients	Num. of Filter	α	Max. Iterations	MSE of Training	Result of Validation Test	
						MSE	Accuration (%)
1.	13	26	0,04	500	0.04085	0,168517	65
2.	13	26	0,06	500	0.040858	0,124498	85
3.	**13**	**40**	**0,06**	**300**	**0.040864**	**0,100717**	**90**
4.	13	40	0,06	500	0.040933	0,167685	80
5.	10	40	0,06	300	0.041295	0,227338	65
6.	10	26	0,05	100	0.041919	0,197227	60
7.	10	40	0,05	500	0.042363	0,183067	60
8.	10	26	0,04	300	0.042357	0,218251	70

Table 3 shows the results of the validation test using the training result parameters with the highest accuracy value.

Table 3. Result of Validation Test From Highest Accuration

No.	Number of MFCC Coefficients	Num. of Filter	α	Max. Iterations	MSE of Training	Result of Validation Test	
						MSE	Accuration (%)
1.	13	26	0,05	100	97	0,142965	70
2.	13	26	0,05	500	97	0,160183	70
3.	13	40	0,05	300	97	0,143600	75
4.	13	40	0,04	100	96	0,118444	80
5.	10	26	0,05	300	96	0,243135	60
6.	10	26	0,06	100	96	0,258932	70
7.	10	40	0,04	300	93	0,266040	55
8.	10	40	0,06	500	93	0,212114	65

Based on Table 2 and Table 3, the highest percentage of validation test result accuracy is 90% with an MSE error value of 0.100717 obtained from the parameter variation of the number of MFCC coefficients 13; number of filters 40; learning rate 0.06; with maximum iterations and epochs of 300 iterations.

6 Conclusion

The Self Organizing Maps method on the Radial Basis Function network can be applied to voice-based emotion identification using MFCC feature extraction as a feature identifier. The best percentage result obtained from emotion identification in this study was 90% with an MSE error of 0.100717 with parameter variations of the number of MFCC coefficients 13; number of filters 40; learning rate 0.06; maximum iteration and maximum epoch of 300 iterations.

7 References

[1]. Deak, A., 2011, *Brain and Emotion: Cognitive Neuroscience of Emotion*, Review of Psychology, **18**(2). 71-80.

[2]. Prawitasari J. E., 1995, *Mengenal Emosi Melalui Komunikasi Nonverbal*, Bulletin Psikologi, **3**(1), 27-43.

[3]. Maria E., Matthias L., Sten H., 2019, *Emotion Recognition from Physiological Signal Analysis; A Review*, Elsevier: Electronic Notes in Theoretical Computer Science, **343**(C), 35-55.

[4]. Albornoz E. M., Milone D. H., dan Rufiner H. L., 2011, *Spoken Emotion Recognition Using Hierarchical Classifier*, Elsevier: Computer speech and Language, **25**(3), 556-570.

[5]. Bombatkar, A., Bhoyar, G., Morjani, K., Gautam, S., & Gupta, V., 2014, *Emotion Recognition Using Speech Processing Using K-Nearest Neighbour Algorithm*, International Journal of Engineering Research and Application (IJERA), 68-71.

[6]. Sahidullah, M.D., & Saha, G., 2012, Design, *Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computational for speaker recognition*, ELSEVIER: Speech Communication, **54**(4), 543-565.

[7]. Tiwari, V., 2010, *MFCC and Its Application in Speaker Recognition,* International Journal on Emerging Technologies, **1**(1), 19-22.

[8]. Fausett, L., 1994. *Fundamental of Neura Networks Architectures, Algorithms, and Application*, London: Prentice-Hall, Inc.

[9]. Kusumadewi, S. & Hartati, S., 2006, *Neuro-Fuzzy Integrasi Sistem Fuzzy dan Jaringan Saraf*, Yogyakarta:Penerbit Graha Ilmu.

[10]. Dai, K., Fell, H.J., & MacAuslan, J., 2008, *Recognizing Emotion in Speech Using Neural network, International conference on Telehealth and Assistive Technologies* (IASTED). Vol. **4**, pp. 31-36. http://www.scopus.com/inward/record.url?eid-2-s2.0-62649110768&par tnerID-tZ0tx3y1.

[11]. Kessous, L., Castellano, G., & Caridakis G., 2010, *Multimodal Emotion Recognition in Speech based interation using facial expression, body gesture and acoustic analysis*, Journal on Multimodal User Interface, Vol. **3**, pp. 33-48.

[12]. Helmiyah, S., Fadlil, A., & Yudhana, A., 2018, *Pengenalan Pola Emosi Berdasarkan Ucapan Menggunakan Ekstraksi Fitur Mel-Frequency Cepstral Coefficient*, Cogito Smart Journal, Yogyakarta.

[13]. Lanjewar, R.B., Mathurkar, S., & Patel, N., 2015, *Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model* (GMM) *and K-Nearest Neighbour* (K-NN) *Techniques*, Procedia Computer Science, Vol. **49**, hal. 50-57.

[14]. Prasetio, B.H., 2017, Pengenalan Emosi Berdasarkan Suara Menggunakan Algoritma HMM, **4(**3), 168-172.

[15]. Eyben, F., 2015, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Institute for Human-Machine Communiacatio, Germany, Springer: Switzerland. ISBN: 978-3-319-27299-1

[16]. Bishop, C.M., 1995, *Natural Network for Pattern Recognition.* UK: Oxford University Press, Cambridge.

[17]. Howlett, R.J. & Jain, L.C., 2001, *Radial Basis Function Networks 2: New Advances in Design,* Springer: Verlag Berlin Heidelberg. ISBN: 978-3- 7908-1826-0.