Relation between Emotional State and Speech Signal

Fergyanto E. Gunawan[1], Kanyadian Idananta[2], Sevenpri Candra[3], Benfano Soewito[4]
[1,2,4]Binus Graduate Programs, Bina Nusantara University, Jakarta, Indonesia 11480
[3]School of Business & Management, Bina Nusantara University, Jakarta, Indonesia 11480
Emails: [1]fgunawan@binus.edu, [2]kidananta@binus.edu, [3]scandra@binus.edu, [4]bsoewito@binus.edu

***Abstract*—Automatic emotion recognition from human speech signal has many important practical applications. For the reason, a number of studies has been performed on the basis of English, German, Mandarin, Persian, and Danish languages. This work intends to develop automatic emotion recognition system on the basis of speech signal in Indonesia language. The study is limited to four emotional states, namely, happy, sad, angry, and fear. The speech data are collected from amateur actors and actresses, and are further quantified using Mel-Frequency Cepstral Coefficient to provide 48 emotion-related features. Finally, these features are used for emotion classification using Support Vector Machine method. The results suggest that the recognition can achieve about 86% of the level of accuracy.**

***Keywords*—Automatic emotion recognition, Indonesia language, Mel-frequency cepstral coefficient (MFCC), Support vector machine (SVM), Speech features**

I. Introduction

Emotion is an intense feeling or reaction directed to a person or a thing. A human being demonstrates their emotion when they are happy, angry, or afraid of a person or a thing [1]. The emotional state is reflected in the forms of speeches, body gestures, and face expressions [2].

Many evidences in the daily life have indicated that misunderstanding of the human emotion could result in fatality as in the case of the accident of Korean Air Flight 801 on August 6, 1997 [3].

On the basis described above, a number of research works have been focused on the understanding of human emotion including developing a system to measure accurately the emotion state on the basis of speech signals of various languages: English [4], German [5], Mandarin [6], Assemese tribe [7], Persian [8], and Danish [9].

Various classification methods have been deployed. Reference [7] studied detection of emotion state of Assam tribe in India using the methods of Gaussian Mixture Model (GMM) classifier and Mel-Frequency Cepstral Coefficient (MFCC). The study concluded that the state feeling of surprise was the most difficult state to be identified correctly. In addition, Ref. [2] had also performed a similar study using the method of Discrete Wavelet Transform (DWT) and Linear Predictive Coefficient (LPC). Reference [2] found that the speeches could provide accurate information up to 95% of the human emotion. Furthermore, Ref. [10] had evaluated the method of Dynamic Time Warpimg (DTW) and MFCC and concluded that the methods were effective for the purpose. Finally, Ref. [11] compared the relative strengths of the speech signal analysis methods, see Table I.

The current study intends study a possibility of automatically classifying emotion state on the basis of speech signal of Indonesia language.

II. Research Methods

The research procedure is as the following. We collected the speech data in Indonesia language. The speakers were two actors and two actresses. All of them were amateur and were active members of Theater Student Activity of Bina Nusantara University in Jakarta, Indonesia. They were asked to simulate four emotional states, namely, happy, sad, angry, and fear. The speech scripts were prepared by the researchers containing 15 sentences in Indonesia language, see Table II. The speakers were asked to speech the sentences with the four emotional states. Several speech data were obtained per person, per sentence, and per emotional state. As the results, 291 speech signals in Indonesia language were collected. As much as 66% of the dataset were utilized for training, and the remainder were for testing.

A. Mel-Frequency Cepstral Coefficients

The speech data were processed to provide Mel-Frequency Cepstral Coefficients (MFCC) and Teager Energy (TE). MFCC is based on human hearing perceptions, which are difficult to perceive frequencies over 1 kHz. MFCC is based on known variation of the human ears critical bandwidth frequency [10]. In summary, the MFCC procedures are: apply the Fourier transform to every segment of the speech signal, map the frequency into the Mel scale and apply triangular overlapping windows on the powers of the spectra, apply logarithmic transformation on the power spectra on each Mel frequency, and apply the discrete cosine transform. The results are 12 MFCC coefficients on each segment.

TABLE I. Advantages and disadvantages of Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Coefficient (LPC), and Dynamic Time Warping (DTW) in modeling speech signal [11].

Methods	Advantages	Disadvantages
MFCC	As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system.	MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise.
LPC	Provides a good approximation within vocal spectral envelope.	On certain condition, the noise may became dominant.
DWT	Reduced storing space for the reference template. Increased recognition rate.	Difficult to find the best reference template for certain words.

978-1-5090-5130-4/16/$31.00 ©2016 IEEE

The computation of the MFCC in detail is of the following [12], [13]. The pre-processing proses is called pre-emphasis process where the speech signal is filtered with the equation,

$$y(n) = x(n) - 0.95 \cdot x(n-1), \tag{1}$$

where n is an integer index, $x(n)$ is the value of the speech signal at the discrete time-step n, and $y(n)$ is the filtered signal. Then, the filtered signal is divided into several small frames where the length of each frame is within the range of 20–40 ms. The voice signal is divided into frames of N samples. The adjacent frames are separated by M where $M < N$. Typically, $M = 100$, and $N = 256$. The signal on each frame is scaled with a window function. The most widely window function is the Hamming window. The windowing process is mathematically written as:

$$z(n) = y(n) \cdot w(n), \qquad \text{and} \tag{2}$$

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \tag{3}$$

where $y(n)$ is a segment of signal prior windowing process, $z(n)$ is after windowing process, and $n \in [0, N-1]$. Then, Fourier transform is applied to each frame signal to transform the time-domain signal to the frequency-domain signal:

$$F(f) = \int_{-\infty}^{+\infty} f(t)\, e^{-2\pi \jmath t f}\, \mathrm{d}t, \tag{4}$$

where the signal in the time domain is denotes by $f(t)$ and in the frequency domain by $F(f)$. The computation of Eq. (4) is performed numerically by the fast Fourier transform algorithm. The obtained Fourier spectra are expressed in form of the power spectra.

The next step is transformation of the Fourier frequency f to Mel-scale frequency by

$$M(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right). \tag{5}$$

and applying triangular overlapping windows. The resulted spectra are subjected to logarithmic transformation. Finally, the discrete cosine transform is applied to the log power spectra.

TABLE II. THE FOUR SPEAKERS, TWO MEN AND TWO WOMEN, WERE ASKED TO SPEAK THE FOLLOWING SENTENCES IN FOUR EMOTIONAL STATES: HAPPY, SAD, ANGRY, AND FEAR.

No	Sentences
1	*Bukunya tadi aku taruh di meja.*
2	*Menurutmu gimana?*
3	*Tadi mama telpon kamu.*
4	*5 jam lagi acara dimulai.*
5	*Itu tas kenapa ditaruh disitu?*
6	*Sabtu ini aku mau pulang dan bertemu dengan Agnes.*
7	*Baru aja buku itu dibawa ke atas dan sekarang dibawa ke bawah lagi.*
8	*Aku disuruh tangkap capung.*
9	*Lisa mau ngumpulin berkasnya hari Rabu.*
10	*Kamu suka sama Budi? (cinta).*
11	*Aku bakal kasih tahu nanti malam.*
12	*Aku itu belum punya pacar.*
13	*Aku sudah melakukannya.*
14	*Makanannya ada di kulkas.*
15	*Ekky tadi kasih aku boneka.*

B. Teager Energy

The Teager energy of a speech signal $x(t)$ is defined by [14]:

$$\Psi\left[x(t)\right] = \dot{x}^2(t) - x(t)\ddot{x}(t) \tag{6}$$

where $\dot{x}(t)$ is the first derivative and $\ddot{x}(t)$ is the second derivative of $x(t)$.

C. Support Vector Machine

In the present study, we only use the Support Vector Machine (SVM) for linearly separable data. The SVM is a numerical method to compute an hyperplane for separating a two-class dataset. It can easily be extended to multiple-class problem. The SVM establishes the hyperplane, governed by $(\mathbf{w}, b)$, by using the support vectors, which are the data points that are closest to the hyperplane. The following SVM formulation is derived from Refs. [15], [16]; readers are advised to the two sources for detail exposition.

We consider the point sets $\mathbf{x}_i \in \Re^d$, as the support vectors, with the categories $y_i \in [-1, +1]$. The hyperplane that separates $y_i = -1$ from those of $y_i = +1$ should satisfy

$$< \mathbf{w}, \mathbf{x} > + b = 0, \tag{7}$$

where $\mathbf{w} \in \Re^d$, $< \mathbf{w}, \mathbf{x} >$ denotes the inner dot product of $\mathbf{w}$ and $\mathbf{x}$, and b is a scalar constant. The hyperplane is obtained by solving:

$$\min_{\mathbf{w},b} L_p = \frac{1}{2} < \mathbf{w}, \mathbf{w} > - \sum_i \alpha_i \left[y_i \left(< \mathbf{w}, \mathbf{x}_i > + b \right) - 1 \right], \tag{8}$$

where $\alpha_i \geq 0$.

D. Accuracy Indicator

The classification accuracy is computed by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{9}$$

where TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

III. RESULTS AND DISCUSSION

Firstly, we discuss how the accuracy of the emotion-state classification changes as the number of features is increased. We should note that in the MFCC analysis, we utilized 12 Mel filter banks from the lowest frequency to the highest frequencies. The coefficients in each bank are preserved in a cepstral vector. For each vector, we determine the minimum and maximum coefficients, and compute its average and standard deviation. Those statistical characteristics are used in the feature vector.

Table III shows the computed accuracies with the increasing of the number feature vectors. The first row shows the accuracy level when only the Teager energy is used as the feature. The second row shows for the case where the features involve the Teager energy and the statistical characteristics of the MFCC in the first bank. The characteristics on the second bank are added in the third row case. The analysis is repeated until all banks are taken into account.

TABLE III. THE CHANGES IN THE LEVEL OF ACCURACY AS THE FUNCTION OF THE CONTENTS OF THE FEATURE VECTOR. THE TE DENOTES THE TEAGER ENERGY, C_i DENOTES THE FEATURE VECTOR OF THE i-BANK OF MFCC WHERE THE VECTOR CONTENTS ARE $C_i = [\min(\mathrm{MFCC}_i)\ \max(\mathrm{MFCC}_i)\ \mathrm{mean}(\mathrm{MFCC}_i)\ \mathrm{deviation}(\mathrm{MFCC}_i)]$, AND $i \in [1, 12]$ IS THE BANK NUMBER.

Feature Vector	Accuracy (%)	Change in Accuracy (%)
$[\mathrm{TE}]$	41.41	-
$[\mathrm{TE}, C_1]$	72.73	31.32
$[\mathrm{TE}, C_1, C_2]$	77.78	5.05
$[\mathrm{TE}, C_1, C_2, C_3]$	85.86	8.08
$[\mathrm{TE}, C_1, C_2, \ldots, C_4]$	85.86	0.0
$[\mathrm{TE}, C_1, C_2, \ldots, C_5]$	86.87	1.01
$[\mathrm{TE}, C_1, C_2, \ldots, C_6]$	85.86	-1.01
$[\mathrm{TE}, C_1, C_2, \ldots, C_7]$	84.85	-1.01
$[\mathrm{TE}, C_1, C_2, \ldots, C_8]$	86.87	2.02
$[\mathrm{TE}, C_1, C_2, \ldots, C_9]$	87.88	1.01
$[\mathrm{TE}, C_1, C_2, \ldots, C_{10}]$	86.87	-1.01
$[\mathrm{TE}, C_1, C_2, \ldots, C_{11}]$	86.87	0.0
$[\mathrm{TE}, C_1, C_2, \ldots, C_{12}]$	85.86	-1.01

TABLE IV. THE COMPUTED CONFUSION MATRIX FOR THE CASE WITH 49 SPEECH FEATURES.

		Prediction			
		Happy	Sad	Angry	Fear
Actual	Happy	34	0	6	1
	Sad	0	22	0	1
	Angry	2	0	15	1
	Fear	2	1	0	14

The results suggest that the most important feature is the Teager energy and then followed by the statistical characteristics in the MFCC first bank. The Teager energy only is capable to achieve 41.4% of the level of accuracy. The features from the first MFCC bank is able to increase the accuracy by 31%. The features from the third MFCC bank have a slightly better quality to increase the accuracy in comparison to those features in the second MFCC bank. The features from the fourth to the last banks can be ignored without sacrificing the accuracy.

The confusion matrix for the case where the feature vector consisting $[\mathrm{TE}, C_1, C_2, \ldots, C_{12}]$ is shown in Table IV. The table shows that a few cases of happy-emotion state were detected as angry emotion. Meanwhile, the motion states of sad, angry, and fear were classified with high accuracy.

IV. CONCLUSION

Automatic emotion recognition on the basis of human speech signal is important as it has many practical benefits. For the reason, many studies have been performed particularly using the speech signals in English, German, Mandarin, Persian, and Danish languages. This work has similar nature but different in the aspects of the language of interest and of the feature vector. The study focuses on Indonesia language and the Teager energy is taken into account as an important feature alongside the features of the statistical descriptive of MFCC. The performed numerical trials suggest that the Teager energy indeed is an important feature as it contributes to the classification accuracy by about 41%. In addition, some statistical descriptive of MFCC associated with low frequencies power spectra are also crucial for accurate classification. The final finding is that the happy-emotion state seems slightly difficult to be differentiated from the angry-emotion state and the emotion states of angry, sad, and fear are detectable from the speech signal at a rather high accuracy.

REFERENCES

[1] N. Frieda, *Moods, Emotion Episodes, and Emotions*. New York: Guilford Press, 1993.

[2] C. Chibelushi and F. Bourel. (2003, January) Facial expression recognition: A brief tutorial overview. Retrieved on January 2015. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/CHIBELUSHI1/CCC_FB_FacExprRecCVonline.pdf

[3] M. Gladwell, *The Ethnic Theory of Plane Crashes "Captain the Weather Radar Has Helped Us a Lot". In Success, Outliers: The Story of Succes*. New York, United States of America: Little, Brown and Company, 2008.

[4] A. Sapra, N. Panwar, and S. Panwar, "Emotion recognition from speech," *International journal of emerging technology and advanced engineering*, vol. 3, no. 2, pp. 341–345, 2013.

[5] B. V. Sathe-Pathak and A. R. Panat, "Extraction of pitch and formants and its analysis to identify three different emotional states of a person," *Internasional journal of computer science*, vol. 9, no. 4, pp. 296–299, 2012.

[6] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and J.-J. Lu, "Detecting emotions in mandarin speech." in *ROCLING*, 2004, pp. 365–373.

[7] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from assamese speeches using mfcc features and gmm classifier," in *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE, 2008, pp. 1–5.

[8] M. Hamidi and M. Mansoorizade, "Emotion recognition from persian speech with neural network," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 5, p. 107, 2012.

[9] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, vol. 8. IEEE, 2005, pp. 4898–4901.

[10] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010. [Online]. Available: http://arxiv.org/abs/1003.4083

[11] S. B. Magre and R. R. Deshmukh, "A review on feature extraction and noise reduction technique," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 2, pp. 352–356, 2014.

[12] S. K. Kopparapu and M. Laxminarayana, "Choice of mel filter bank in computing mfcc of a resampled speech," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 121–124.

[13] J. Kaur and A. Sharma, "Emotion detection independent of user using mfcc feature extraction," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 6, pp. 230–234, 2014.

[14] E. Kvedalen, "Signal processing using the teager energy operator and other nonlinear operators," Master Thesis, Department of Informatics, University of Oslo, 2003.

[15] N. Christianni and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2008.

[17] Hindarto and Sumarno, "Feature Extraction of Electroencephalography Signals using Fast Fourier Transform," *CommIT (Communication & Information Technology) Journal*, vol. 10, no. 1, pp. 49–52, 2016. http://journal.binus.ac.id/index.php/commit/article/view/1548/1421