Comparison Method in Indonesian Emotion Speech Classification

[1] Jutono Gondohanindijo, [2]Edy Noersasongko, [3]Pujiono, [4]Muljono
[5]Ahmad Zainul Fanani, [6]Affandy, [7]Ruri Suko Basuki
[1,2,3,4,5,6,7] Informatics Engineering Department
Dian Nuswantoro University
Semarang, Indonesia
[1]jutono@mhs.dinus.ac.id, [2]edynur@dsn.dinus.ac.id, [3]pujiono@dsn.dinus.ac.id, [4]muljono@dsn.dinus.ac.id
[5]fanani@dsn.dinus.ac.id, [6]affandy@dsn.dinus.ac.id, [7]ruri.basuki@dsn.dinus.ac.id

***Abstract*— Emotion speech recognition aims to study the formation and change in emotional status based on human speech signals. This study aims to solve the problem of recognizing emotional speech, classifying it through sound processing to recognize sound patterns based on the classification of emotions and comparing the level of accuracy of some of the classifiers used. Stages the method starts from collecting or obtaining voice data, pre-processing, feature extraction, classification, results and evaluation. The pre-processing stage is carried out to clean the data first before processing. Extraction features used are MFCC and Classifier used are Support Vector Machine, Random Forest, Naïve Bayes and Neural Network. The system that is built can recognize 4 types of emotions namely anger, pleasure, sadness and disgust. The database used is the sound from audio recordings. From the results of observation, the highest accuracy rates are as follows: SVM of 100%, NN 99.7%, RF 97% and NB 94.1%. Evaluation stages using Confusion Matrix with 10 k-fold.**

Keywords – speech; emotion; classification; comparison; accuracy; recognition

I. INTRODUCTION

The recognition of voice-based emotions aims to automatically identify the human emotional state of his voice. This is based on an in-depth analysis of the mechanism of voice signal generation, extracting some features that contain emotional information from the speaker's voice, and taking appropriate pattern recognition methods to identify emotional states. Like a typical pattern recognition system, the emotional recognition system consists of four main modules: speech input, feature extraction, classification and emotion output [1]. General system architecture The introduction of sound-based emotions has the three steps shown in Figure 1:

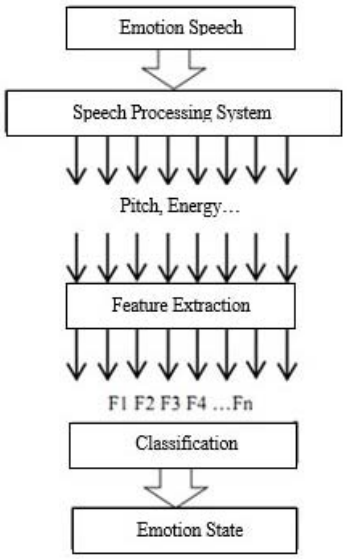

Fig. 1. The Speech Emotion Recognition System

- A speech processing system, extracting a number of appropriate signals, such as pitch or energy,
- This number is summarized into several features that are appropriate or just needed,
- A classifier of learning by taking sample data and how to connect features to emotions.

In Figure 1 it can be seen that after pre-processing, the speech is modeled based on its characteristics. Feature extraction is based on speech partitions in small intervals known as frames. To choose the appropriate features that carry information about emotions from sound signals is an important step in a sound-based emotion recognition system. Energy is the basic and most important feature in sound signals. To get the statistical value of the energy feature, we use the short-term function to extract the energy value in each frame of speech. Then we can obtain the statistical value of energy in the whole sample and calculate energy,

978-1-7281-3832-9/19/$31.00 ©2019 IEEE

such as mean values, max values, variants, range variations, energy contours [2].

Speech recognition, in general, aims to replace sensors or physical actions such as hands that function as a command to the machine. In addition, sound pattern recognition can be used to classify data from the speaker to recognize gender (gender), dialect and emotion. This can be used for further purposes, for example for security, namely recognizing passwords for certain sounds, recognizing someone's emotional state through his emotions, recognizing one's health through speech.

This research conducts to solve the problem of recognizing speech emotions, classification, and compare their accuracy levels based on the classifier used. The used method are getting speech data, pre-processing, extraction features using MFCC, classification (using Support Vector Machine, Random Forest, Naïve Bayes and Neural Network), results and evaluation. The results of this research system can recognize 4 types of emotions: angry, happy, sad and disgust. The Database used is the sound from the audio recordings. From the observation results, the highest degree of accuracy in succession as follows: SVM of 100%, NN 99.7%, RF 97% and NB 94.1%. Evaluation stages using Confusion Matrix with 10 K-fold.

II. RELATED WORK

Technical scientists and health scientists have been researching human feelings for many years [3]. Research on Speech Emotion Recognition describes emotional voice vowel utterances that are distinct from each person [3,4]. It is particular to each person for "Emotion" and "Emotional Reaction Learning," which has a set fundamental function for each individual[5].

Data from Introduction to Emotional Speech is collected and trained by recording human emotional speech [7]. Human-spoken independently understanding of feelings is not always the same as the context of the spoken sentence [8]. Since 1980, Emotion Speech's introduction has been explored using acoustic parameter statistical methods [8].

This implementation of introducing emotional expression is commonly used for human and machine communication, but is also implemented in various fields, such as diagnosing the feelings of a person [9].

This attribute includes mental data for the extraction of emotional expression, spectral characteristics and prosodic features. Mel-frequency cepstral coefficients (MFCC) and linear cepstral predictive coefficients (LPCC) are spectral characteristics, while prosodic characteristics containing basic frequency characteristics are used for distinct emotional recognition [10].

Artificial Neural Network (ANN), k-nearest neighbors (KNN), Gaussian Mixtures Model (GMM), Hidden Markov Models (HMM) and Support Vector Machine (SVM) are some techniques of recognizing emotional speech recognition. As Schuller et al. put it. With pitch and energy characteristics, HMM can classify emotional speech by 86 percent [11]. Shen et al. says. Using SVM, emotional expression can be classified by 82.5% using the Berlin emotional database [12].

Ververidis and Kotropoulos [13] used a bayes classifier method to classify voice samples into five emotions (anger, happiness, neutral, sadness, and surprise).

Lee and Narayanan [14] used a linear discriminating classifier and neighbourhood classifier k-nearest to combine acoustic, lexical, and discourse methods to classify speech samples into two (adverse, non-negative) feelings. This technique had precision levels of over 80 percent of unspecified participants ' free phrases in the two emotions [15][16].

A fundamental set of 35 characteristics was chosen using statistical method and the use of artificial neural network and random forest classifiers [17]. There were seven categories, namely happiness, anger, fear / anxiety, sorrow, boredom, disgust and neutrality. Artificial neural network classification has achieved an precision of 83.17% in the speaker-dependent structure and 77.19% in the Random Forest. In the speaker-independent structure, a mean precision of 55% was achieved for artificial neural network classification, while Random Forest achieved a mean precision of 48%.

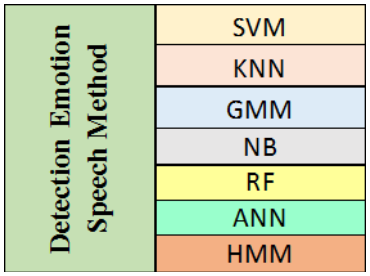

Fig. 2. Detection Emotion Speech Methods

III. METHOD

The method that will be used in carrying out the introduction and classification of emotional speech in this paper is SVM, NN, RF, NB. The process of introduction and classification can be seen in Figure 3 as follows:

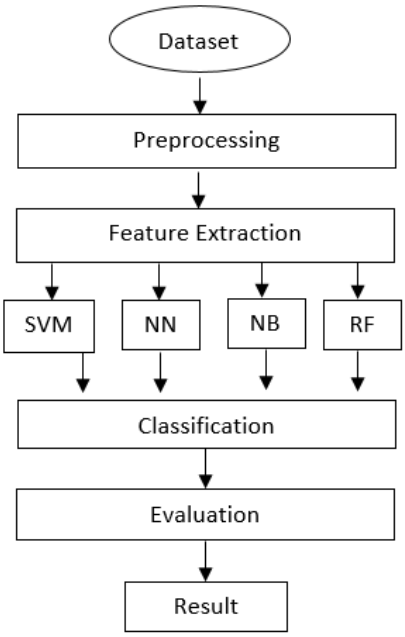

Fig. 3. Emotion Speech Classification

A. Support Vector Machine

Support Vector Machine is an algorithm of machine learning based on the concept of statistical learning. The primary concept or core of SVM is to convert the initial input into greater feature sizes using kernel functions and to attain the ideal classification level in the fresh feature space where there is a clear distinction between characteristics acquired from the ideal positioning of the separator hyperplane [5].

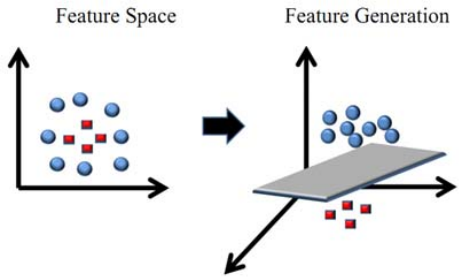

Fig. 4. Transformation from FS to FG

Figure 4 shows a method for classifying information that can not be divided linearly by transforming data into feature space sizes so that it can later be separated linearly by the mapping or conversion process. By using the transformation function, the function of the learning outcomes produced is

$$f(x) = wx + b \text{ atau } f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b \quad (9)$$

$$Dengan \quad w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (9.1)$$

$$Ld = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \alpha_i \alpha_j y_i y_j K(x_i x_j) \quad (9.2)$$

$$syarat : 0 \leq \alpha_i \leq C \text{ dan } \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$Ld = Min \frac{1}{2} \alpha^T H \alpha + c^T \alpha \quad (9.3)$$

$$H(i,j) = y_i y_j k(x_i x_j) \quad (9.4)$$

$$b = -\frac{1}{2}(w.x^+ + w.x^-) \quad (9.5)$$

The x value in the function above states the vector input, while b is a scalar which is referred to as bias. Assume that there are two data class y labels that have members of the value [+1, -1] which state the negative and positive classes where the value i = 1 to N with the value N is the length of the size of the matrix. To calculate the weighting point w used equation (9.1) where αi is the weight value of each data point, xi is a vector, and yi is a data class. To obtain the weight value of each data (αi) is obtained by obtaining the maximum value from the Lagrange Multipier (Ld) Duality of equation (9.2).

To calculate Lagrange Multipier Duality Quadratic Programming is used in equation (9.3). In Quadratic Programming the resulting value is the minimum value so that to fulfill the Lagrange Multipier Duality the value of the Lagrange Multipier Duality is negated (-Ld) so that the maximum value can be obtained from the Quadratic Programming equation. To obtain Quadratic Programming, the Hessian Matrix is calculated using equations (9.4) and c which are vectors with members of the value constant 1. The value of bias (b) will be obtained by equation (9.5) where the bias is half the product of the multiplication (w) vector x+ and x-. Then the classification function will be obtained by equation (9) where the weight point value (w) is multiplied by the kernel function K (xi xj) then added to the value of bias (b).

B. Neural Network

Neural networks can be called methods of information processing driven by biological nerve cell mechanisms. Artificial Neural Networks resemble the human brain in two ways, namely understanding obtained by the network through the learning process and the strength of the nerve cell (neuron) bond used to store information. The Artificial Neural Network takes the foundation of the biological nervous system, gets feedback from information or nerve cell production in the nerve tissue.

Each input emerges through an current connection and is processed in neurons for each input and output data pattern provided to the Artificial Neural Network. In layers called neuron layers, these neurons accumulate.

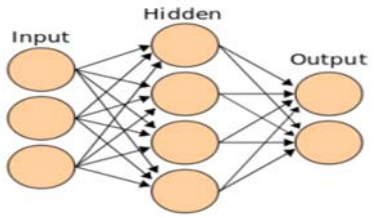

Fig. 5. Neural Network Structure

1. Network Device

Artificial Neural Networks consist of a number of different layers and vertices per layer.

1. Input Layer: consists of unit node units that act as input data processing processes on the neural network.
2. Hidden Layer: consists of unit node units which are analogous to hidden layers and act as layers that pass the response from the input.
3. Output Layer: consists of node units that play a role in providing solutions from input data.

2. Classification of Artificial Neural Networks

Artificial Neural Networks can be distinguished based on the level of activation of the output:

1. Singel Layer Perceptrons(SPL)

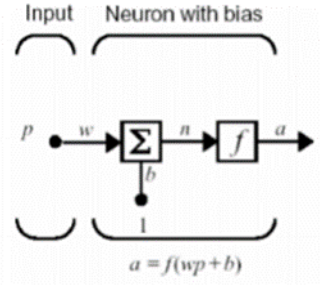

Fig. 6. Single Layer Perceptrons

SPL or Single Layer Perceptrons are artificial network groups using one layer of input and one output. The feature used is difficult to limit, i.e. if the input weight exceeds the bias value, the output unit is worth one.

2. Multi Layer Perceptrons(MLP)

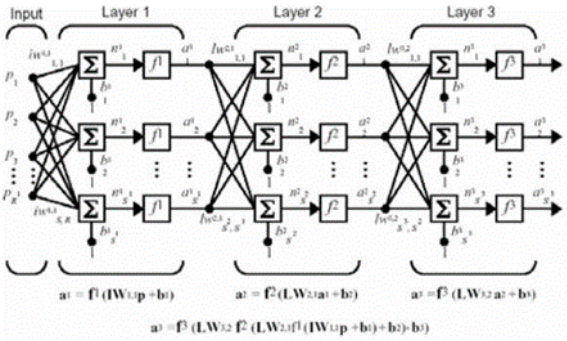

Fig. 7. Multi Layer Perceptrons

This form is a kind of of neural artificial network that utilizes an sophisticated route with a concealed or hidden layer of at least one layer. The issue with using this form is how many layers are used to obtain ideal outcomes or, in other words, to obtain the lowest mistake.

IV. RESULT AND DISCUSSION

A. *Dataset*

This paper uses 639 datasets taken from audio recordings of human voices with sound samples with emotional characteristics such as disgust, happy, angry, sad.

B. *Preprocessing*

Emotional audio recording datasets still need to be processed by preprosessing method using the following methods:

1. Remove noise
2. Normalize the duration of each recording (839 ms)
3. Match the bit rate (705 kbps)
4. Equate sampling frequency (44.1 khz)
5. Matching the Channel (Mono)

C. *Feature Extraction*

From the audio dataset, the feature retrieval process is done through MATLAB using MFCC.
The results of this extraction feature are stored in CSV format file and labeled emotions, which will be used as input to the classification process.

D. *Classification*

The classification process is carried out using 4 methods, namely Support Vector Machine, Random Forest, Neural Network and Naïve Bayes.

E. *Evaluation*

Based on the classification results using SVM, RF, NN and NB with stratified 10-fold cross validation, the test evaluation results are as follows:

TABLE 1. RESULT COMPARISON

Method	Accuracy	Precision	Recall	F1 Mesure
SVM	1.000	1.000	1.000	1.000
RF	0.970	0.970	0.970	0.970
NN	0.997	0.997	0.997	0.997
NB	0.941	0.945	0.941	0.940

The evaluation results of 4 methods in table 1 show that the Support Vector Machine method has the best level of accuracy with CA (Classification Accuracy) of 1.000, Precision of 1.000, Recall of 1.000 and F1 Measure of 1.000.

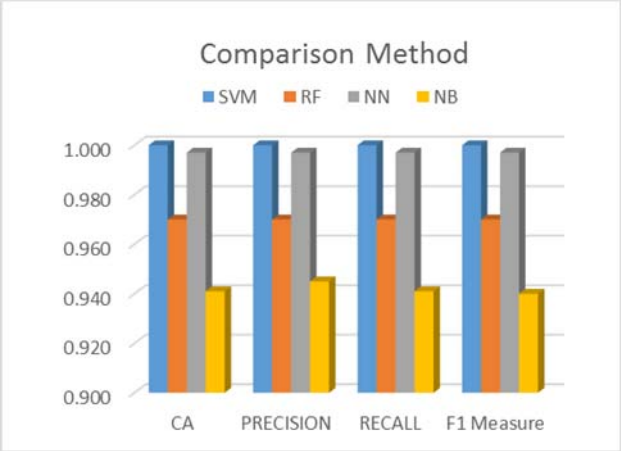

Fig. 8. Comparison of Classifier Graph

The results of the comparison of measurements of accuracy, precision, recall and F1 from the recognition and classification of emotional speech are shown in the graph above where SVM has the best results of accuracy, precision and recall compared to RF, NN and NB. The evaluation results are also indicated by the table confussion matrix as follows:

TABLE 2. SVM CONFUSION MATRIX

Actual \ Predicted	DISGUST	ANGRY	SAD	HAPPY	Σ
DISGUST	159	0	0	0	159
ANGRY	0	160	0	0	160
SAD	0	0	160	0	160
HAPPY	0	0	0	160	160
Σ	159	160	160	160	639

The table 2 above shows that with the SVM method obtained correct predictions of 159 disgust, 160 angry, 160 sad dan 160 happy with 639 data so that the prediction accuracy rate was 100%.

TABLE 3. NN CONFUSION MATRIX

Actual \ Predicted	DISGUST	ANGRY	SAD	HAPPY	Σ
DISGUST	159	0	0	0	159
ANGRY	1	159	0	0	160
SAD	0	0	160	0	160
HAPPY	0	0	0	160	160
Σ	160	159	160	160	639

The table 3 above shows that with the Neural Network method, correct predictions of 159 disgust, 159 angry, 160 sad dan 160 happy with 639 data so that the prediction accuracy rate was 99.8%.

V. CONCLUSION

Based on the results of testing on datasets about emotional speech by using Support Vector Machine, Random Forest, Neural Network and Naïve Bayes, it can be concluded that:

1. Testing 639 datasets for emotion speech with SVM has the best level of accuracy and precision from the other 3 methods.
2. The test results using SVM also show that this method has the best Recall level.
3. Calculation of F1-Measure (F1) shows that SVM has the best value.

REFERENCES

[1] Y. Pan, Peipei Shen And Liping Shen, 2012, Speech Emotion Recognition Using Support Vector Machine, International Journal of Smart Home, Vol. 6, No. 2, April.

[2] D.Ververidis, C.Kotropoulos, and I. Pitas, 2004, Automatic emotional speech classification, in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol.1, pp. 593-596, Montreal, May.

[3] Rong J, Li G, Chen Y-PP. Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag May 2009;45(3):315–28.

[4] Fairbanks G, Hoaglin LW. An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monogr 1941;8(1):85–90 .

[5] Sethu V. Automatic emotion recognition: an investigation of acoustic and prosodic parameters. The University of New South Wales; 2009 .

[6] Gharavian D, Sheikhan M, Nazerieh A, Garoucy S. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput Appl 2011;21(8):2115–26 .

[7] Mencattini A et al. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowl-Based Syst 2014;63:68–81.

[8] Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Commun Sep. 2006;48(9):1162–81 .

[9] Ayadi M. E., Kamel M. S. and Karray F., 'Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases', Pattern Recognition, 44 (16), 572-587, 2011.

[10] Zhou y., Sun Y., Zhang J, Yan Y., 'Speech Emotion Recognition using Both Spectral and Prosodic Features', IEEE, 23(5), 545-549, 2009.

[11] Schuller B., Rigoll G., Lang M., 'Hidden Markov Model Based Speech Emotion Recognition', IEEE ICASSP, 1-3, 2003.

[12] Shen P., Changjun Z. and Chen X., 'Automatic Speech Emotion Recognition Using Support Vector Machine', Proceedings of International Conference On Electronic And Mechanical Engineering And Information Technology, 621-625, 2011.

[13] D. Ververidis, C. Ko tropoulos, and I. Pitas, "Automatic emotional speech classification," IEEE International Conference on Acoustic, Speech, and Signal Processing , vol. 1, pp. I–593–596, 2004.

[14] M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. o n Speech and Audio Pro cessing , vol. 13, pp. 293–303, 2005.

[15] J. C ho, S. Kato, and H. Itoh, "Bayesian-based inference o f d ialogist's emotion f or sensitivity robots," IEEE International Conference on Robot & Human Interactive Communication , pp. 792–797, 2007.

[16] S. Kato, Y. Sugino, and H. I toh, "A bayesian approach to emotion detection in d ialogist's vo ice for human robot interaction," Lecture Notes in Computer Science, vol. 4252, pp. 961–968, 2006.

[17] Iliou, Theodoros and Anagnostopoulos, Christos-Nikolaos, 2009, Comparison Of Different Classifiers for Emotion Recognition, 13th Panhellenic Conference on Informatics, Mytilene, Lesvos Island

[18] Kumar S. S., T. Ranga Babu 2015, Emotion and Gender Recognition of Speech Signals Using SVM, International Journal of Engineering Science and Innovative Technology, Vol. 4, No. 3, h. 128-137, India.

[19] Rawat, Arti., 2015, Emotion Recognition through Speech Using Neural Network, Volume 5, Issue 5.