Detecting Human Emotion via Speech Recognition by Using Ensemble Classification Model

Sathit Prasomphan(✉) and Surinee Doungwichain

Department of Computer and Information Science, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800, Thailand
ssp.kmutnb@gmail.com

Abstract. Speech Emotion Recognition is one of the most challenging researches in the field of Human-Computer Interaction (HCI). The accuracy of detecting emotion depends on several factors for example, type of emotion and number of emotion which is classified, quality of speech. In this research, we introduced the process of detecting 4 different emotion types (anger, happy, natural, and sad) from Thai speech which was recorded from Thai drama show which was most similar with daily life speech. The proposed algorithms used the combination of Support Vector Machine, Neural Network and k-Nearest Neighbors for emotion classification by using the ensemble classification method with majority weight voting. The experimental results show that emotion classification by using the ensemble classification method by using the majority weight voting can efficiency give the better accuracy results than the single model. The proposed method has better results when using with fundamental frequency (F0) and Mel-frequency cepstral coefficients (MFCC) of speech which give the accuracy results at 70.69%.

Keywords: Speech emotion recognition · Feature extraction
Ensemble classification · Weight majority vote · k-nearest neighbor
Neural Network · Support Vector Machines

1 Introduction

Speech Emotion Recognition (SER) is a challenging research area in the field of Human-Computer Interaction (HCI). The purpose of SER is to recognize emotion such as anger, disgust, fear, happiness, sadness, etc. from tonal variations in human speech [1, 2]. Several algorithms were introduced to make computer to be able to understand and to be able to classify several types of emotion in human speech. Some benefit of knowing this emotion from speech is to use with the application which requires a man-machine interaction such as computer tutorial, automatic translation, mobile interaction, health care, children education, etc. Emotion is an importance mental and physiological state. In natural, baby learns to recognize emotional information before understanding semantic information in his/her mother's utterance [3]. Reliable emotion detection in usability tests will help to prevent negative emotion [4]. Detecting emotion

J. J. Jung et al. (Eds.): BDTA 2017, LNICST 248, pp. 66–73, 2018.
https://doi.org/10.1007/978-3-319-98752-1_8

can help particularly for user opinion mining or stress prevention [3, 4]. Computer may not be able to exactly understand the natural of these emotions unless we employ the speech processing. Many researchers have used the statistics of difference speech attributes for being a representation of each sound such as pitch, formant, amplitude or power of the speech. Speech features can be classified to one of these three categories: *prosodic features* such as pitch (F0), intensity and duration, *voice quality* and *spectral features* such as Mel-Frequency Cepstral Coefficients (MFCC) or Linear Prediction Cepstral Coefficients (LPCC).

In case of classification model, researchers offer several model such as Support Vector Machines (SVM) [3, 8], Gaussian Mixture Model (GMM) [2], Hidden Markov Modeling (HMM) [7], k-Nearest Neighbor (k-NN) [10], and Neural Network (NN) [12, 15]. Although the above techniques provided the better classification accuracy, however these techniques are single model that resulted in a used data set in the study must include the parameter configuration step. Moreover, each parameter must be fixed which cause to bias and poor performance. Another way to reduce bias is to use common decision (ensemble), which can create diversity and minimize the errors caused by the variance [9]. Researchers have attempted to make ensemble decision applied to enhance the emotion classification. Anagnostopoulos et al. [11] presented a research on *ensemble majority voting classifier* for speech emotion by using a decision from the base classifier. The decision with majority voting using k-NN, C4.5, and SVM with polynomial kernel was used to find a suitable model to classify the speech in HUMAINE database [5]. These framework provided accuracy by 96%. Morrison [14] presented a technique for searching feature that combined ensemble model by using the base classifier SVM with RBF kernel, random forest, k-NN, K* and multilayer perceptron. The algorithm provided accuracy by 79.43% and 73.29% for NATURAL and ESMBS database in ordering. The results showed that if a dataset has different types of information and emotion, the feature selection methods and ensemble model will be different. Vasuki [17] focused on searching frame work to reject noisy and weak input file by using the weight factor ensemble model with SVM classifier to detect outliers. If input is unusual, it will be rejected from the training dataset. This framework showed the accuracy by 74.70%.

From all of these research shows that the ensemble model can increase the performance of emotion classification from speech. It also emphasizes that the effectiveness of methods for separating emotion depends on several factors such as the properties of the selected feature in the experiments, number of emotions; the quality of the audio data is also affected as well. Therefore, we have selected a set of features that are critical to the dataset and methodology to optimize performance of classification by using ensemble model.

The paper is organized as follows. Following this, Sect. 2 provides the details in speech emotion recognition. Section 3 discusses the experimental setup. Results and performance comparison are given in Sect. 4. Section 5 gives conclusions and discussions.

2 Speech Emotion Recognition System

The speech emotion recognition system is described in Fig. 1. In this section, the pre-processing of speech signal which is the pre-emphasis, frame blocking and Hamming windowing was described. The following speech features: energy, zero-crossing rate (ZCR), pitch, MFCC was used. The feature normalization was calculated for every windows of a specified number of frames by statistical method. Classifier was modeled to classify emotions. Finally, the ensemble model was applied to integrate the result of classifier with weighted majority vote. Details of each process can be described as followed:

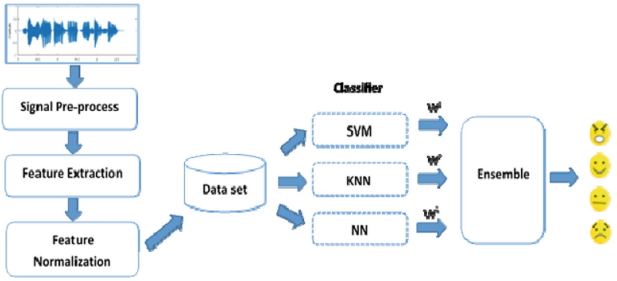

Fig. 1. Speech emotion recognition system.

A. Signal Pre-processing: The basic operations used in the speech pre-processing include the following: pre-emphasis, frame blocking and hamming windowing.

1. Pre-emphasis: The speech signal $s(n)$ is sent to a high-pass filter as show in Eq. (1).

$$s2(n) = s(n) - a * s(n-1) \tag{1}$$

where $s2(n)$ is the output signal and the value of a is usually between 0.9 and 1.0. The z-transform of the filter is given in Eq. (2).

$$H(z) = 1 - a * z - 1 \tag{2}$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can amplify the importance of high-frequency formants.

2. Frame blocking: The speech signal is divided into a sequence of frames where each frame can be analyzed independently and represented by a single feature vector. Frame shift is the time difference between the start points of successive frames, and the frame length is the time duration of each frame. The frame block is of length 10 ms to 40 ms from the filtered signal at every interval of 1/2 or 1/3 of frame length.
3. Hamming windowing: In order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by $s(n), n = 0, \ldots N-1$

Then the signal after Hamming windowing is $s(n) * w(n)$, where the Hamming window $w(n)$ is defined in Eq. (3).

$$w(n, a) = (1 - a) - a\cos(2pn/(N - 1)) \tag{3}$$

B. Feature Extraction: The speech feature extraction which is also called speech coding is a very important and it is basically part in the automatic speech processing systems. Features of the speech are generally obtained from the digital speech. Various methods are utilized that aim to extract speech features which are useful to classify the type of emotions. In this research, the extracted features are energy, ZCR, pitch, and MFCC.

C. Feature Normalization: The speech segments have different lengths. In order to obtain isometric speech segments and reduce redundancy of data, the statistical method [3] was adopted to normalize the states. For each coefficient, mean, variances, median, maximum and minimum across all frames are calculated.

D. Classifiers: Classifier is another component of a speech emotion recognition system. In this research, we used three classification methods: SVM, Neural Network and k-nearest neighbor.

E. Ensemble Classification Method: Ensemble classifier is the model which combines several classifiers' technique for solving the same problem by using the results from all of classifiers for decision in the final step. Ensemble model [13, 16] composed of several model for example, vote ensemble which uses the same training data with several classifier, bootstrap aggregating (Bagging) which uses the random training data and constructs the single ensemble model, and random forest which similar to bagging technique but instead of using random data, it randomly selects attribute from dataset and uses several decision tree for becoming classifier in the ensemble model. In this research, the vote ensemble with base classifiers which has low computational complexity and difference theoretical background was selected. The proposed model aims to reduce bias and redundancy [11] by using the combined model with weighted majority vote. If the classifier in the ensemble does not provide the identical classification result, then it is reasonable to attempt to give the more competent classifiers more power in making the final decision. We called this step is weighted majority vote. The formula for weighted majority vote is shown in Eq. (4).

$$\sum\nolimits_{t=1:T} w_t \;\; d_t, \, J(x) \; = \; max_{j=1,\ldots,c} \, m \sum\nolimits_{t=1:T} w_t \;\; d_{t,j} \tag{4}$$

The T classifiers are class-conditionally independent with accuracies $p_1, \ldots, p_T$. The optimal weights for the weighted majority voting rule can be shown to be

$$w_t \; = \; p_t/(1 - p_t) \tag{5}$$

3 Experimental Setup

Thai Emotional speech corpus [6] was used to classify emotion states. This corpus construction has been funded by National Electronics and Computer Technology Center (NECTEC). All emotional speech collected from conversations by professional actors and actresses in a Thai drama show that contains many background music and noise within the speech. There are two groups of emotion were used to annotate in this corpus. The first group consists of four basic emotions: neutral, happy, sad, and angry. The second group consists of twelve labels: happiness, satisfaction, fear, surprise, anger, jealousy, rage, doubt, hate, excitement, sadness, and fun. For this research, we firstly focused on detecting emotions from the first group. It is possible to recognize four real emotions of human. We used only 352 utterances from 2908 utterances in the corpus were utilized in this work. The details of each emotion are shown in Table 1.

Table 1. Number of emotions in Thai emotion speech corpus.

Emotion	Male	Female	Male + Female
Anger	47	72	119
Happy	42	40	82
Neutral	41	40	81
Sad	33	37	70
Total	163	189	352

In Thai emotion speech corpus, we have randomly selected 352 speeches for study which with and without noise, background music or one of them for the diversity of speech in the experiment. We use the cross-validation with holdout 1/3 to split the data into two sets for training and testing, 236 training speech and 116 testing speech.

In case of signal pre-process, the coefficient was set in the pre-emphasis step with 0.9375. In the framing process, frame has been segmented with size of 480 samples or approximately 30 ms, and the distance between the frames (frame overlap) is 240 samples or about 15 ms. After that we used Hamming window to emphasize the importance signal in the middle frame signal. The speech feature used in this research was energy, ZCR, F0 and MFCC. After feature extraction had been processed, the feature was normalized by using statistical methods. The important features were combined to analyze if it most affects to emotional classification. We used MFCC to combine with *prosodic feature* (energy, ZCR, F0) due to the MFCC feature give the highest accuracy compared to prosodic feature as shown in Table 2.

Emotion classification has been created by using the ensemble model from the same set of data. When each classifier gives the predicted class, these results will be weight for each classifier which is [2, 6, 7] for SVM with RBF-7 kernel function, KNN, and NN. The setting weight values depend on the prediction accuracy. After that, the sum of predicted class and predicted weight in each classifier was calculated for voting. The performance of proposed model was based on evaluation of data classification performance by using Eq. (6),

Table 2. The Classification accuracy in different features and models.

No.	Feature	No. feature	Accuracy (%)				
			SVM (RBF 7)	KNN	NN	Bagging	Weighted majority vote
1	F0	5	40.52	43.10	40.52	40.52	42.24
2	Energy	5	35.34	42.24	50.00	42.24	40.52
3	ZCR	5	41.38	37.93	40.52	41.38	41.38
4	MFC	105	62.93	56.03	58.62	60.34	68.97
5	MFCC + F0	110	**66.38**	56.90	61.21	62.93	**70.69**
6	MFCC + Energy	110	65.52	56.03	62.07	61.21	66.38
7	MFCC + ZCR	110	62.07	56.03	**65.52**	**63.79**	66.38
8	MFCC + F0 + Energy	115	**66.38**	57.76	62.07	60.34	68.10
9	MFCC + F0 + ZCR	115	65.52	52.59	58.62	58.62	69.83
10	MFCC + Energy + ZCR	115	61.21	**62.07**	59.48	**64.66**	65.52
11	MFCC + F0 + Energy + ZCR	120	63.79	59.48	57.76	62.07	66.38

$$Accuracy = (TP + TN)/(TP + FN + TN + TN) * 100 \quad (6)$$

where, TP is true positive, TN is true negative, FP is false positive, FN is false negative.

4 Experimental Result

In this research, 5 models were tested for the speech classification accuracy which is 3 single model: SVM with RBF-7 kernel function, k-NN, and NN, and 2 ensemble models: bagging which uses base classifier by using decision tree and weighted majority vote with 3 base classifier model: SVM with RBF-7 kernel function, k-NN and NN. In addition, the feature was also compared its classification accuracy.

From Table 2, it shows that model which can give the best classification accuracy for speech emotion classification is ensemble weighted majority vote by using F0 and MFCC.

The confusion matrix in Table 3 showed that ensemble weighted majority vote model with F0 and MFCC give the best accuracy with 70.69%.

Table 3. Confusion matrix for the feature set MFCC + F0 of ensemble weighted majority vote.

Emotion	Recognized emotions (%)			
	Anger	Happy	Neutral	Sad
Anger	**76.92**	7.69	12.82	2.56
Happy	14.81	**66.67**	14.81	3.70
Neutral	11.11	18.52	**62.96**	7.41
Sad	8.70	0.00	17.39	**73.91**

5 Conclusions

This research presents a novel algorithm for detecting human emotion via speech recognition by using ensemble classification model. The proposed algorithm aims to detect the emotional by using information with the combination of SVM classifier, Neural Network classifier and k-Nearest Neighbor with the weighted majority voting ensemble method with combine speech feature Fundamental Frequency (F0) and Mel Frequency Cepstral Coefficient (MFCC) for Thai emotional speech corpus. The experimental results show that the proposed framework can efficiently find the correct speech emotion compared to by using the comparing method. For the future work, the process for noise removal and background music should be considered. In addition, the feature selection and model selection for improve the classification accuracy should be focused.

Acknowledgment. This research was funded by King Mongkut's University of Technology North Bangkok. Contract no. KMUTNB-58-GEN-048.

References

1. Ayadi, M.M.H.E., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**, 572–587 (2011)
2. Xu, S., Liu, Y., Liu, X.: Speaker recognition and speech emotion recognition based on GMM. In: 3rd International Conference on Electric and Electronics (2013)
3. Seehapoch, T., Wongthanavasu, S.: Speech emotion recognition using Support Vector Machines. In: the 5th International Conference on Knowledge and Smart Technology (KST), pp. 219–223 (2011)
4. Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G., Holzinger, A.: Emotion detection: application of the valence arousal space for rapid biological usability testing to enhance universal access. In: Stephanidis, C. (ed.) UAHCI 2009. LNCS, vol. 5614, pp. 615–624. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02707-9_70
5. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of Interspeech (2005)
6. Kasuriya, S., Teeramunkong, T., Wutiwiwatchai, C.: Developing a Thai emotional speech corpus. In: International Conference on Asian Spoken Language Research and Evaluation (2013)
7. Kasuriya, S., Banchaditt, T., Somboon, N., Teeramunkong, T., Wutiwiwatchai, C.: Detecting emotional speech in Thai drama. In: 2nd ICT International Student Project Conference (ICT-ISPC) (2013)
8. Shen, P., Changjun, Z.: Automatic speech emotion recognition using support vector machine. In: International Conference on Electronic & Mechanical Engineering and Information Technology, pp. 621–625 (2011)
9. Thamsiri, D., Meesad, P.: Ensemble data classification based on decision tree, artificial neuron network and support vector machine optimized by genetic algorithm. J. King's Mongkut's Univ. Technol. North Bangk. **21**(2), 293–303 (2011)
10. Rieger Jr., S.A., Muraleedharan, R., Ramachandran, R.P.: Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In: 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 589–593 (2014)

11. Anagnostopoulos, T., Skourlas, C.: Ensemble majority voting classifier for speech emotion recognition and prediction. J. Syst. Inf. Technol. **16**(3), 222–232 (2014)
12. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion recognition in speech using neural networks. In: 6th International Conference on Neural Information Processing, vol. 2, pp. 495–501 (1999)
13. Mu, X., Lu, J., Watta, P., Hassoun, M.H.: Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. In: Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, 14–19 June, pp. 2168–2171 (2009)
14. Morrison, D., Wang, R., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. J. Speech Commun. **49**(2), 98–112 (2007)
15. Aha, D., Kibler, D.: Instance-based learning algorithms. Mach. Learn. **6**, 37–66 (1991)
16. Sharkey, A.J.C.: Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems. Springer, London (1999). https://doi.org/10.1007/978-1-4471-0793-4
17. Vasuki, P.: Speech emotion recognition using adaptive ensemble of class specific classifiers. Res. J. Appl. Sci. Eng. Technol. **9**(12), 1105–1114 (2015)