Use of Neural Network Classifier for Detecting Human Emotion via Speech Spectrogram

Sathit Prasomphan

Department of Computer and Information Science,
Faculty of Applied Science,

King Mongkut's University of Technology North Bangkok, 10800, Thailand

*Corresponding Author: sathitp@kmutnb.ac.th

Abstract— The goal of this research is to present the algorithm for emotion recognition of speech based on neural networks. The input for neural network should be the partitions of speech spectrogram. The proposed algorithms are based on feature extraction which analysis of two dimensions time-frequency representation of speech signals. The Berlin Emotion Database with 7 emotion classes was tested. The experimental results show that the method works well in time-frequency domain. Moreover, the proposed framework can efficiently find the correct speech emotion compared to using the comparing method.

Keywords—*Speech Emotion Recognition; Neural Network; Spectrogram; Feature extraction;*

I. INTRODUCTION

Speech Emotion Recognition (SER) is a challenging research area in the field of Human Computer Interaction(HCI). Computer may not be able to exactly understand the natural of this unless we employ the speech processing. Several algorithms were introduced to make computer to be able to understand and to be able to classify the several type of emotion in human speech. The benefit of knowing emotion from speech is to use with the application which require a man-machine interaction such as computer tutorial, automatic translation, mobile interaction, health care, children education, *etc.* Some of researchers have used the statistics of the difference attributes of speech for being a representation of each sound such as pitch, formant, amplitude or power of the speech. These 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 Ceptral Coefficients (MFCC) or Linear Prediction Ceptral Coefficients (LPCC) [7].

Emotion feature extraction is an importance issue for the speech emotion classification. The most popular feature is *prosodic feature* which is pitch, energy and *spectral feature* such as MFCC, LPCC. The limitation of this system is that algorithms analyze each parameter separately and then combine them into a set of feature vector [9]. Another feature extraction is by using a two dimensions magnitude spectrogram which is a graphical display of the squared magnitude of the time-frequency of speech [9]. This magnitude spectrogram representation usually contains distinctive pattern that capture different characteristic of speech emotion signal [11].

Several algorithms have been developed for the emotional classification of speech signals such as Eun Ho Kim [1] presented a speech recognition using Eigen-FFT in clean and noisy environment. The accuracy of this algorithm is 90 percent. Qingli Zhang [2] used the technique based on Gaussian Mixture model for a classification of audio signals. Yazid Attabi and Pierre Dumouchel [3] presented an algorithm for speech classification which used a technique of Gaussian mixture model with support vector model to help for increasing accuracy of emotional classification. Yixiong Pan [4] presented a classification for emotion classification by using Support Vector Machine (SVM) for classification of audio signals. The limitation of each method is analyze each parameter separately in the feature extraction process and then combined these features into a set of feature space. From this limitation, a new novel for speech classification which does not use separately feature was proposed. A new research trend in emotional recognition/classification is to use a spectrogram for feature extraction processes. This method is based on the analysis of specific part of the spectrogram which called region of interest (ROI)[6]. Kun-Ching Wang[5][6] presented a novel texture image feature for emotion sensing in speech, the texture image information derived from the spectrogram image can be extracted by using laws mask to characterize emotional state. The experimental results show that the proposed method based on using spectrogram image can provide significant classification for emotional recognition in speech. Liang He, et al.[9] presented and tested an approach to feature selection based on analysis of two dimensions time-frequency representation of speech. The wavelet packet method combined with the log-Gabor filters was showed in the method. In their method, the highest classification accuracy was archived while using single vowels. J.Pribil and A.Pibilova[8] focused on application of emotion speech conversion to evaluation of sentence after spectral and prosodic modification for emotional speech conversion. Results of evaluation experiment are confirmed that spectrogram can be successfully used for visual comparison of different type of emotional synthetic speech. A. Harimi et al. [10] recognized human emotion by analyzing the acoustics of speech sound. Spectral Pattern features (SPs) and Harmonic Energy features (HEs) for emotion recognition was proposed. These features extracted from the spectrogram of speech signal using image processing techniques.

Results of these research confirm that a spectrogram can capture all of the characteristics of speech. Accordingly, in this paper, the spectrogram of speech signal for emotion recognition is proposed. The important feature will be extracted from spectrogram. We provide the new architecture for neural network to reduce the difficulty of the classification process and reduce number of feature to be sent to neural network compared with the other method.

The processes for detecting the speech emotion based on spectrogram are discussed in Section II. Results and discussions are discussed in Section III, and the conclusions of the research are presented in Section IV.

II. PROPOSED ALGORITHMS

In this section, we first briefly review the fundamental of the speech signal. Next, the speech recognition system is introduced based on recognizing the spectrogram. Finally, neural network is shown for labeling the emotion of speech signal.

A. Speech features.

In speech signal there are basic speech attributes for the feature extraction. Example of speech signal is shown in Fig. 1.

Fig. 1. Example of speech signal.

There are some importance feature in every speech which are pitch(f_0), log-energy, formant frequency, mel-band energy and mel-frequency ceptral coefficient. More details of this feature can be described as follows.

1) Frequency or pitch

A pitch is the fundamental frequency (f_0) of the speech signal. It refers to the tone of speech. The pitch of each signal is different depends on the speed of the vibration of the object. Example of frequency or pitch is shown in Fig. 2.

Fig. 2. Frequency of low frequency wave and high frequency wave.

2) Wavelength

Wavelength refers to the distance between the two adjacent peaks which is 1 Lambda (λ), which occurs during the recording of the sound wave. Example of wavelength is shown in Fig. 3.

Fig. 3. The distance between two adjacent peaks which is 1 Lambda (λ).

3) Amplitude

Amplitude refers to the height of the waves. That represents the intensity of the sound or volume (Loudness). Example of amplitude is shown in Fig. 4.

Fig. 4. Amplitude.

4) Prosodic features

Prosodic features form the most widely used features in SER. In order to extract these types of features, statistical properties of pitch and energy tracking contours are commonly used. The statistical information of pitch and energy tracking contours were capture. These functions include: min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentile, interquartile range, average deviation, standard deviation, skewness and kurtosis.

5) Spectral features

Two types of spectral features: The Mel-Frequency Cepstral Coefficients (MFCCs) and formants are reported as effective spectral features for emotion recognition. The first 12 MFCCs and 4 formants are extracted from 20 ms Hamming-windowed speech frames every 10 ms, and so their contours are formed.

B. Speech emotion recognition system

1) *Emotional speech input*: The emotional speech data set from standard database is selected for increasing efficiency of the proposed framework.

2) *Pre-processing*: The continuous time signal is selected. The pre-emphasis for flattening speech spectrum is processed by High-pass Finite Impulse Response (FIR). After that normalize amplitude signal between -1 to 1.

3) *Framing*: Segment the speech signal into the small frame with time length 25 ms.

4) *Windowing*: The discontinuities of the speech signal at the edge of frame was reduced by using Hamming window.

5) *FFT*: The FFT is an abbreviation of the Fast Fourier Transform. Essentially, the FFT is still the DFT for transforming the discrete-time signal from time domain into its frequency domain. The difference is that the FFT is faster and more efficient on computation.

6) *Spectrogram*: After the speech signal in time domain in each segment is transformed to the frequency domain by FFT, rotate the amplitude – frequency by 90 degrees. Map spectral amplitude to gray level (0-255) value. Higher the amplitude, darker the corresponding region. Time vs frequency representation of a speech signal is referred to as spectrogram. This graph will be used for the classification step. Example of time and frequency representation of a speech signal is shown in Fig. 5.

Fig. 5. Time vs frequency representation of a speech signal.

7) *Feature extraction*: in this step the spectrogram which is time-frequency representation of speech signal is used to be input of neural network. The process for converting spectrogram to be input in neural network is as follows:

At first, convert the spectrogram into grayscale image.

Next, use a threshold to convert image to binary by set the value below threshold to be zero and the value greater than threshold to be one. After that, divide the spectrogram into 4x4 matrix. Sum all of number in each sub area. The total 16 areas will be used to neural network training process. Example of the total 16 areas which will be used in neural network training process is shown in Fig. 6.

Fig. 6. The process for converting spectrogram to be input in neural network.

8) *Neural network for emotion speech classification*.

In this paper, neural networks have been created in order to distinguish for each spectrogram. The formulation of neural network can be shown in Equation (1),

$$n = \sum_{i=1}^z x_i w_i + b \quad (1)$$

where n is a weight summation.

x_i is an input vector i

w_i is a weight vector i

z is number of input attribute

b is bias

i is vector index from 1 to z

Use input attribute from windows sizing process become an input to neural network which is the summation of identity point from each sub-window. We use 16 attributes to be input attribute.

Fig. 7. Neural network architecture

Neural network architecture for learn the relationship of data is shown in Fig. 7.

Train on only the training set by setting the stopping criteria and the network parameters. Feed-forward multilayer neural network with back propagation learning algorithms was used. The network consists of one input layer, one hidden layer, and one output layer. Set of inputs and desired output belongs to training patterns were fed into the neural network to learn the relationship of data. The manifold of these data set are generated using only the set of nearest data which called the local training approach. The difference between our proposed

algorithms and traditional method is using the semi-supervised technique. The process in hidden layer is to adjust weight which connected to each node of input. The root mean square error is calculated from desired output and its calculated output. If the error is not satisfied with the predefined values, it will propagate error back to the former layer. This will be done from the direction of the upper layer towards the input layer. This algorithm will adjust weight from initial weight until it gives the satisfied mean square error.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Experimental setup.

The Berlin Emotion Database (German emotional speech) [11] is a well-known public database. This database contains 535 speech files for 7 emotion classes. These seven emotions are anger, joy, boredom, neutral, disgust, fear, and sadness. This database includes 10 different contexts expressed by ten professional actors (5 males and 5 females). Table 1 lists the numbers of samples for the emotion categories. In our experiments, we selected 5 speech emotions which are angry, sad, happy, neutral, and fear. These five emotions data can be represented for the emotion which shows the obviously different spectrogram. There are different characteristics. Details of emotion in speech database can be described as Table I.

TABLE I. NUMBER OF SAMPLES IN THE BERLIN DATABASE

Emotion	Female	Male
Anger	67	60
Joy	44	27
Boredom	46	35
Neutral	40	39
Disgust	35	11
Fear	32	37
Sadness	37	25
All	301	234

Tables II–IV are the confusion matrices, which are widely used graphical tools that reflect the performance of an algorithm. Each row of the matrix represents the instances of a predicted class, while each column represents the instances of an original class. Thus, it is easy to visualize the classifier's errors while trying to accurately predict each original class's instances. Percentage of the test data being classified to the original emotion was shown inside the table.

B. Experimental results.

The experimental results for speech emotion classification were shown in Table II- Table IV. Table II shows the confusion matrix of the proposed algorithms with male and female speech. Table III shows the confusion matrix of the proposed algorithms with male speech. Table IV shows the confusion matrix of the proposed algorithms with female speech.

The experimental results shown that by using the proposed method the following results was shown:

For the anger emotion the accuracy of proposed algorithms is 80.67 percent, 76.67 percent and 84.67 percent in both male and female speech, male speech, and female speech, respectively.

For the sad emotion the accuracy of proposed algorithms is 79.35 percent, 75.35 percent and 82.35 percent in both male and female speech, male speech, and female speech, respectively.

For the happy emotion the accuracy of proposed algorithms is 82.47 percent, 79.47 percent and 85.47 percent in both male and female speech, male speech, and female speech, respectively.

For the neutral emotion the accuracy of proposed algorithms is 76.80 percent, 73.80 percent and 78.80 percent in both male and female speech, male speech, and female speech, respectively.

For the fear emotion the accuracy of proposed algorithms is 83.28 percent, 74.28 percent and 85.28 percent in both male and female speech, male speech, and female speech, respectively.

From the experimental results, we confirmed that the spectrogram image can be successfully used for visual quality comparison of different approach to emotion speech. The method works well in the time and frequency domain. Moreover, by using neural network for being a classifier, the accuracy of classification was satisfying algorithm. Number of attribute can be reduced compared to the past method.

TABLE II. CONFUSION MATRIX OF THE PROPOSED ALGORITHMS (MALE AND FEMALE)

Emotion	Emotion Recognition				
	Angry	Sad	Happy	Neutral	Fear
Angry	80.67	4.43	7.35	2.18	5.37
Sad	4.07	79.35	3.29	5.67	7.62
Happy	7.43	3.79	82.47	3.98	2.33
Neutral	2.15	7.65	5.31	76.80	8.09
Fear	3.45	7.63	5.31	0.33	83.28

TABLE III. CONFUSION MATRIX OF THE PROPOSED ALGORITHMS (MALE)

Emotion	Emotion Recognition				
	Angry	Sad	Happy	Neutral	Fear
Angry	76.67	2.83	9.95	3.64	6.91
Sad	6.07	75.35	4.29	3.67	10.62
Happy	8.68	5.79	79.47	3.97	2.09
Neutral	3.56	9.15	3.02	73.80	10.47
Fear	2.07	9.52	3.69	10.44	74.28

TABLE IV. CONFUSION MATRIX OF THE PROPOSED ALGORITHMS (FEMALE)

Emotion	Emotion Recognition				
	Angry	Sad	Happy	Neutral	Fear
Angry	84.67	2.72	5.15	7.02	0.44
Sad	5.56	82.35	6.02	6.01	7.62
Happy	9.25	4.69	85.47	2.34	0.06
Neutral	1.59	8.64	3.01	78.80	7.96
Fear	7.05	4.05	2.78	0.84	85.28

IV. CONCLUTIONS

This research presents a novel algorithm for detecting human emotion via speech recognition by using speech spectrogram. The spectrogram of speech signal for emotion recognition is proposed. We extract the important feature from spectrogram and provide the new architecture for neural network to reduce the difficulty of the classification process and reduce number of feature to be sent to neural network compared with the other methods. We have presented a new approach to feature extraction based on analysis of two dimensions time-frequency representation of a speech signal. The algorithm was tested with EMO-Database. The experimental results show that the proposed framework can efficiently find the correct speech emotion compared to using the traditional method.

ACKNOWLEDGMENT

This research was funded by King Mongkut's University of Technology North Bangkok.
Contract no. KMUTNB-GEN-58-28.

REFERENCES

- [1] Eun Ho Kim , Kyung Hak Hyu, Soo Hyun Kim and Yoon Keun Kwak, "Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments", 16th IEEE International Conference on Robot & Human Interactive Communication August 26 -29, 2007 / Jeju, Korea, 2007.
- [2] Qingli Zhang, Kunxia Wang, Fuji Ren, "Speech emotion recognition using combination of features", 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 9-11 June 2013, pp. 523 – 528.
- [3] Yazid Attabi and Pierre Dumouchel, "Anchor Models for Emotion Recognition from Speech", IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 4, NO. 3, JULY-SEPTEMBER 2013.
- [4] Yixiong Pan, Peipei Shenand Liping Shen, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home Vol.6, No.2, April, 2012/10/1.
- [5] Kun-Ching Wang, "Speech Emotion Classification using Texture Image Information Feature.", International Journal of Signal Processing System. Vol.3, No.1, June, 2015.
- [6] Kun-Ching Wang, "Time-Frequency Feature Representation Using Multi-Resolution Texture Analysis and Acoustic Activity Detector for Real-Life Speech Emotion Recognition", Sensors 2015, 15(1), pp.1458-1478.
- [7] Bhoomika Panda, Debananda Padhi, Kshamamayee Dash, Sanghamitra Mohanty, "Use of SVM Classifier & MFCC in Speech Emotion Recognition System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012, pp.225-230.
- [8] J. Pribil, and A. Pribilova, "An Experiment with Evaluation of Emotional Speech Conversion by Spectrograms", MEASUREMENT SCIENCE REVIEW, Volume 10, No. 3, 2010, pp.72-77.
- [9] Ling He, M. Lech, M.C.Maddage, N.B.Allen, "Time-frequency feature extraction from spectrograms and wavelet packets with application to automatic stress and emotion classification in speech Information", 7th International Conference on Communications and Signal Processing, 2009. ICICS 2009, Macau, 8-10 Dec. 2009, pp. 1 – 5.
- [10] A. Harimi1, A. Shahzadi1, A.R. Ahmadyfard and Kh.Yaghmaie, "Classification of emotional speech through spectral pattern features", Journal of AI and Data Mining, Vol. 2, No .1, 2014, pp. 53-61.
- [11] Berlin Emotion Database (German emotional speech).