Speech Emotional Classification Using Texture Image Information Features

Kun-Ching Wang
Department of Information Technology and Communication, Shih Chien University, 200 University Road, Neimen, Kaohsiung 84550, Taiwan
Email: kunching.wang@gmail.com

***Abstract*—Emotion recognition is one of the latest challenges for human-computer interaction (HCI). In general, a system approach for recognition of human emotional state is usually from two inputs: audio/speech and visual expressions. So, the emotional recognition system needs two audio-based and image-based kernels to process audio and visual modules. In order to cost down the requirement of the emotional recognition containing two-kernel two-module (TKTM) system, the speech-based kernel can be regarded as an image-based processing. In this paper, we present a novel speech emotional feature extraction based on visual signature extraction derived from time-frequency representation. First, we transform spectrogram as a recognizable image. Next, we use cubic curve to enhance the contrast of speech spectrogram image. Then, the texture image information (TII) derived from spectrogram image can be extracted by using Laws' masks to characterize emotional state. Finally, we use the support vector machine (SVM) to classify emotion.**

***Index Terms*—emotional feature extraction, speech emotional recognition, spectrogram, texture image information**

I. Introduction

Human emotion recognition can obtain emotional information from speech, facial expressions, gestures, and so forth. Among the information, Speech provides a most nature and fundamental interface for human-computer interaction (HCI). With the exponential growth in available computer power and signification progress in speech technologies, speech emotion recognition (SER) system is one of important roles in HCI. The SER has several potential applications, such as the interface with robots [1]-[3], call center environment [4], and enhancement of speech and speaker recognition performance.

In general, the SER is computational tasks consisting of two major parts: feature extraction and emotion machine classification. In fact, the emotional feature extraction is a crucial issue in the SER system and is an emphasis in this paper. All features have to carry sufficient information about the emotional states of a speaker. These popular features are prosodic features (e.g., pitch-related features, energy-related features [5], [6], and speaking rate [7]) and spectral features (e.g., linear predictive cepstral coefficients (LPCCs) [8], [9] and Mel-frequency predictive cepstral coefficients (MFCCs) [6], [9]-[11]). So far, a variety of acoustic features have also been explored.

In [5], the twenty pitch and energy related features are selected to recognize 7 discrete emotions (anger, disgust, fear, surprise, joy, neutral and sad). The accuracy in the speech corpus consisting of acted and spontaneous emotion utterances in German and English exceeded 77.8%. In [6], pitch, log energy, formant, band energies and MFCCs were selected as base features. The velocity/acceleration of pitch was added to form feature streams. The average classification accuracy achieved was 40.8% in a SONY AIBO database. In [7], the authors used pitch, formant, intensity, speech rate and energy related features to classify neutral, anger, laugh and surprise. The accuracy was about 40% for a 40-sentence corpus. In [9], various speech features, namely, energy, pitch, zero crossing, phonetic rate, LPCCs and their derivatives, were tested and combined with MFCCs. The authors also used a simple but efficient classifying method, Vector Quantization (VQ), for performing speaker-dependent emotion recognition. The average recognition accuracy achieved was about 70%. In [11], the short time log frequency power coefficients (LFPC) along with MFCCs were adopted as emotion speech features to recognize 6 emotions in a 60-utterance corpus. The average accuracy was 78%. In [12], fundamental frequency, energy and audible duration features were extracted to recognize sadness, boredom, happiness and anger in a corpus recorded by eight professional actors. The overall accuracy was only about 50%, but the discrimination of anger and other basic emotions can be successfully separated by these features. In [13], the prosodic features, derived from pitch, loudness, duration and quality features were extracted to recognize 400-utterance database. However, the accuracy in recognizing five emotions (anger, happy, neutral, sad and bored) was only 42.6%.

In this paper, we will propose a novel emotional feature extraction based on texture image information (TII) derived from speech spectrograms and apply into an emotion sensing in speech. First, we transform speech signal into gray-level spectrogram image. Next, we use cubic curve to enhance the contrast of spectrogram image

Manuscript received June 20, 2014; revised September 18, 2014.

©2015 Engineering and Technology Publishing
doi: 10.12720/ijsps.3.1.1-7

signal. Then, we extract the texture image information using Laws' masks to characterize correct emotional state.

In order to demonstrate the high effectiveness of the suggested TII-based feature extraction for emotion sensing in speech, we provide results on two open emotional databases (EMO-DB and eNTERFACE) and one self-recording database (KHUSC-EmoDB). Finally, the proposed TII-based feature extraction combining with support vector machine (SVM) classifier is then presented. Experimental results show that the proposed TII-based feature information inspired by human perception can provide significant classification for emotional recognition.

II. Emotional Speech Database

To demonstrate effectiveness of the proposed TII-based feature extraction applied into SER system, we carried out experiments on the three emotional datasets: EMO-DB, eNTERFACE and KHUSC-EmoDB. In the following, we will first discuss and describe the quality of the three emotional datasets. Then, our experimental results are presented in later.

A. EMO-DB

The Berlin Speech Emotion Database (EMO-DB) [14] was recorded at the Technical University, Berlin. It contains seven classes of basic emotions (Anger, Fear, Happiness, Disgust, Boredom, Sadness, and Neutral). Ten professional German actors (five men and five women) spoke ten sentences in German language.

B. eNTERFACE Corpus

The eNTERFACE corpus is a further public, yet audio-visual emotion database. It consists of six emotion categories: Anger, Disgust, Fear, Happiness, Sadness, and Surprise [15]. The 42 subjects (eight women) from 14 nations were recorded in English at office environment.

C. Self-Recording Database (KHUSC-EmoDB)

The recording of the corpus of KHUSC-EmoDB comprises Mandarin language. Its members are all derived from the students of Shih-Chien University. The emotional voice of this corpus is recorded from 4 women and 13 men. Each speaker is recorded in all four emotions (Happiness, Fear, Sadness and Anger), so a total of 408 sentences for four emotions are presented.

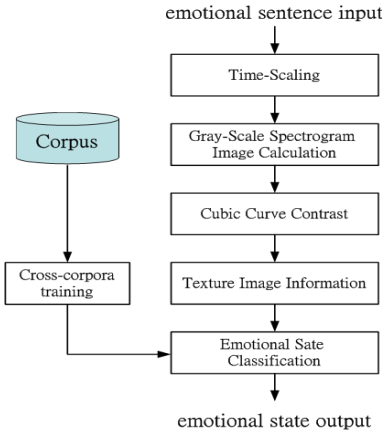

Figure 1. Diagram of the proposed TII-based SER algorithm

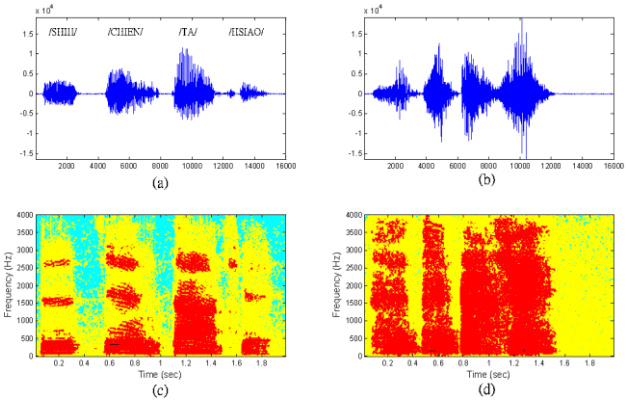

Figure 2. The input speech signals and the corresponding spectrograms. Speech uttered in Mandarin sentence "SHIH-CHIEN-TA-HSIAO."

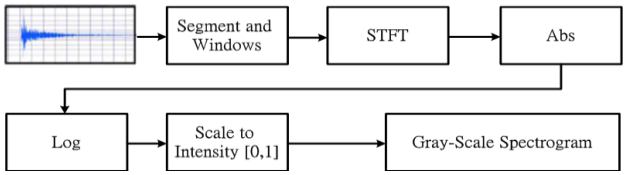

Figure 3. Gray-Scale spectrogram image calculation

III. The Proposed TII-Based SER System

The proposed TII based SER algorithm consists of five main parts: (i) time-scaling, (ii) gray-scale spectrogram image calculation, (iii) cubic curve contrast, (iv) texture image information, (v) emotional state classifier. Each part is shown in Fig. 1 and is described in the following subsections.

A. Approaches to Time-Scaling of Speech

In order to unify the length of emotional voice signal, the emotional speech will be normalized. According the above cause, we will utilize an approach to time-scaling of speech. The time scale modification (TSM) algorithm uses overlap-addition (OLA) to synthesize the desired speech signal. Griffin *et al.* first proposed the OLA synthesis to realizing time-scale modifications using time domain operations only [16]. In order to improve the OLA algorithm, the synchronized OLA (SOLA) algorithm [17] was proposed to improve the shortcomings of the OLA. In this subsection, the normalized method of Overlap-and-Add (OLA) is utilized to solve the problem of inconsistent length of the sound spectrogram.

B. Spectrogram Image Calculation

In this subsection, a new method that extracts texture features from speech spectrogram is presented. It is well known that a 2-D narrowband speech spectrogram is a graphical display of the squared magnitude of the time-varying spectral characteristics of speech [18]. It is compact and highly efficient representation carrying information about parameters such as energy, pitch F0, formants and timing. These parameters are the acoustic features of speech most often used in emotion recognition systems [19]-[21]. The additional advantage is that by analyzing a speech spectrogram, it is more likely to

©2015 Engineering and Technology Publishing

preserve and take into account speech features that are caused by specific psycho-physiological effects appearing under certain emotions.

The input speech signals and the corresponding spectrograms are shown in Fig. 2. The two speech signals pronounce in Mandarin sentence "SHIH-CHIEN-TA-HSIAO", one uttered with "Neutral" emotion and the other with "Anger" emotion. Some visible differences can be observed in terms of signal duration and amplitude in Fig. 2(a) and Fig. 2(b). The speech uttered with anger emotion has a duration less than that uttered with neutral emotion. The average amplitude of the signal has a higher value in case of the speech signal uttered with anger emotion. Compared to the speech uttered with neutral emotion in Fig. 2(c) and Fig. 2(d), the spectrograms show that the frequencies have shifted upward or have higher values in the speech signal uttered with anger emotion.
With increasing level of stress, the spectrograms revealed increasing formant energy in the higher frequency bands, as well as clearly increasing pitch for strong level stress. Other acoustic information, such as the formants also vary under different levels of stress. These observations indicate that the representation of texture image on spectrogram usually contains distinctive patterns that capture different characteristics of neutral emotion and anger emotion signals. Furthermore, the TII features on spectrograms that can be used to discriminate the difference between various emotional levels in speech. In this subsection, inspired by the concept of spectrogram image feature [22] we generate the spectrogram images with time-frequency-intensity representation shown in Fig. 3.

First, the time-frequency-intensity representation, $X(k,t)$, is determined by applying to the input signal with the windowed Short-Time Fourier Transform (STFT), which is given by

$$X(k,t)=\sum_{n=0}^{N-1}x[n]w[n-t]e^{-2\pi ikn/N},\ k=0,.....,N\text{-}1 \quad (1)$$

where $x[n]$ in the input speech signal. N is the length of the window, $w[n]$ is the Hamming window function and k corresponds to the frequency $f(k)=kf_s/N$, where f_s is the sampling frequency in Hertz.

Owing to the logarithmic of the human perception of sound, the log-spectrogram defined as:

$$S_{\log}(k,t)=\log(|X(k,t)|) \quad (2)$$

Next, the spectrogram image representation, $R_{Spec\,\mathrm{Im}\,g}(k,t)$, is defined by the log-spectrogram is normalized into grayscale normalized image, within the range [0, 1]:

$$R_{Spec\,\mathrm{Im}\,g}(k,t)=S_{\log}(k,t)-S_{\min}/S_{\max}-S_{\min} \quad (3)$$

C. Image Contrast Enhancement Using Cubic Curve

In order to allow that the contrast of gray-scale spectrogram image is upgraded, the task of feature extraction can be smoothly processed. This paper uses a cubic curve for enhance the image to adjust its contrast [23]. The appropriate adjusting curve is shown in Fig. 4. This can be seen that the adjusting curve contains an inflection point. So, we can produce a variety of different curvature of the curve by controlling the inflection point. Based on the above motivation, adjusting the curve inflection point to change the image contrast is utilized motivation.

First, assuming that curve must pass through (0, 0) and (255, 255) two points, and the cubic curve as in (5) shown below:

$$y=f(x)=ax^3+bx^2+cx+d \quad (4)$$

where x is the pixel value in the original image, y is the pixel value of the image after adjusting the curve.

In Fig. 4, "A" represents a cubic curve inflection point in the x-coordinate. In order to determine the unknown variables at all three curves, the study uses parameter "A" to obtained compensation curve needed, wherein all variables in the cubic curve calculated by (5) to (8) below:

$$A=\min_{x\in I}\{x\}+0.7\{\max_{x\in I}\{x\}-\min_{x\in I}\{x\}\} \quad (5)$$

$$b^2=3\times a-(255)^2\times 3a^2-255\times 3\times a\times b \quad (6)$$

$$a=1/(255)^2-3\times 255\times A+3\times A^2 \quad (7)$$

$$c=1-a\times(255)^2-b\times 255 \quad (8)$$

where I is an image, x is the image pixel value at any point in the image. $\min_{x\in I}\{x\}$ is represented as a minimum pixel value. $\max_{x\in I}\{x\}$ is expressed as the maximum pixel value of the image.

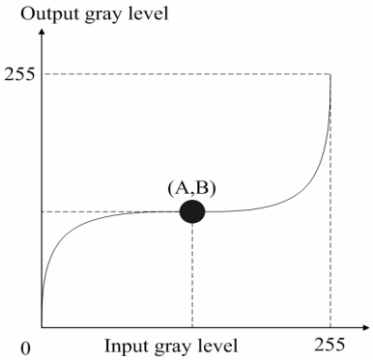

Figure 4. The different compensation curve with different turning point

In general, the texture information is the main resonance frequency of the speech signal. After contrast adjustment, the size of each frequency component of spectrogram for the original speech sound is highlighted. Then, the image can obviously reflect the texture information on spectrogram. We can see the original happy, scared, angry and sad emotions in four spectrograms without/with contrast adjustment. The performances have difference texture image information on the spectrograms. After compensated by image enhancement, each of emotional texture image information can effectively display.

©2015 Engineering and Technology Publishing

D. Extraction of Texture Image Information by Laws' Masks

In this subsection, the TII features derived from spectrogram image will be extracted by the Laws' masks based on the principle of texture energy measurement [24]. The laws masks are well described for texture energy variation in image processing. In general, the Laws' masks consist of five masks derived from one-dimensional vectors, such as edge (E_5), level (L_5), spot (S_5), ripple (R_5) and wave (W_5). All the masks were expressed in the (9)-(13).

$$E_5 = \text{Edge detection: [-1 -2 0 2 1]} \quad (9)$$

$$L_5 = \text{Level detection: [1 4 6 4 1]} \quad (10)$$

$$S_5 = \text{Spot detection: [-1 0 2 0 -1]} \quad (11)$$

$$R_5 = \text{Ripple detection: [1 -4 6 -4 1]} \quad (12)$$

$$W_5 = \text{Wave detection: [-1 2 0 -2 1]} \quad (13)$$

The two-dimensional filters of size 5×5 were generated by convoluting any vertical one-dimensional vector with a horizontal one. Finally, the 25 two-dimensional mask combinations are shown in Table I.

TABLE I. THE LIST FOR MUTUAL COMBINATIONS OF 2-D LAWS' MASKS

$L_5^TL_5$	$E_5^TL_5$	$S_5^TL_5$	$W_5^TL_5$	$R_5^TL_5$
$E_5^TL_5$	$E_5^TE_5$	$S_5^TE_5$	$W_5^TE_5$	$R_5^TE_5$
$S_5^TL_5$	$E_5^TS_5$	$S_5^TS_5$	$W_5^TS_5$	$R_5^TS_5$
$W_5^TL_5$	$E_5^TW_5$	$S_5^TW_5$	$W_5^TW_5$	$R_5^TW_5$
$R_5^TL_5$	$E_5^TR_5$	$S_5^TR_5$	$W_5^TR_5$	$R_5^TR_5$

First, we convoluted the image with each two-dimensional mask to extract texture information from an image $I_{(i,j)}$ of size ($M \times N$). For example, if we used E_5E_5 to filter the image $I_{(i,j)}$, the result was a "texture image" ($TI_{E_5E_5}$), seen in the (14).

$$TI_{E5E5} = I_{ij} \otimes E5E5 \quad (14)$$

All the two-dimensional masks, except L_5L_5, had zero mean. According to Laws, texture image $TI_{L_5L_5}$ was used to normalize the contrast of all the texture images $TI_{(i,j)}$, seen in the (15).

$$Normalize(TI_{mask}) = TI_{mask} / TI_{mask} \quad (15)$$

Next, the outputs (*TI*) from Laws' masks were passed to "texture energy measurement" (TEM) filters. We can calculate non-linear interval by processing TI normalized and yield through "Texture Energy Measurements, (TEM)" filter. These consisted of a moving non-linear window average of absolute values (16).

$$TEM_{ij} = \sum_{u=-7}^{u=7} \sum_{v=-7}^{v=7} \left[Normalize(TI_{i+u, j+v}) \right] \quad (16)$$

However, not all mask energy is able to be used as the input basis of texture energy. Hence, we rotate the values within 25 masks from yielded TEM, and take out unchangeable TR values before and after rotation. The TR represents the calculation value of Laws' Mask, which are the values specially used to measure the texture energy as seen in the (17).

$$TR_{E5L5} = (TEM_{E5L5} + TEM_{L5E5})/2 \quad (17)$$

After the (17), we use this result to extract three texture feature values: Mean, Standard Deviance (SD) and Entropy. These three features are used to judge the variation of texture information. The (18) to the (20) are the calculation formula of three features values, where TR_{ij} represents the unchangeable values within 25 masks from TEM before and after rotation, $M \times N$ represents the size of whole image. Finally, each equation will have the 14-dimensional of feature vectors. A total of three feature vectors are 42-dimensional and the feature vectors will be used as the input for training the SVM classifier.

$$Mean = \sum_{i=0}^{M} \sum_{j=0}^{N} \left[TR_{ij} \right] / M \times N \quad (18)$$

$$SD = \sqrt{\sum_{i=0}^{M} \sum_{j=0}^{N} (TR_{ij} - Mean)^2 / M \times N} \quad (19)$$

$$Entropy = \sum_{i=0}^{M} \sum_{j=0}^{N} (TR_{ij})^2 / M \times N \quad (20)$$

E. Support Vector Machine (SVM)

After extracting the texture image information, the next is emotional state classification. The SVM, a supervised learning algorithm, is usually used for classification and regression. It is very popular in recent years due to its remarkable performance. In this paper, we adopt the support vector machine as our emotion classifier.

In the training phase, it needs to be given a set of samples belonging to two classes. The objective is to find a hyperplane that can completely distinguish these two classes [25]. We have the training data set:

$$\{x_i, y_i\},\ i = 1, 2, ..., n,\ x_i \in R^d,\ y_i \in \{1, -1\}$$

where d is the dimension of training set, y_i is the emotional class labeled. We use training data to find the best hyperplane, and then use it to classify the data. The hyperplane is defined and described as follows:

$$y_i(w \cdot x_i + b) > 0,\ \ i = 1, 2, ..., N \quad (21)$$

Rescaling w and b as:

$$\min_{1 \le i \le N} y_i(w \cdot x_i + b) \ge 1,\ \ i = 1, 2, ..., N \quad (22)$$

When the distance to the closest point is maximal, the hyperplane is called the optimal separating hyperplane, and $2/\|w\|$ is the margin. By introducing Lagrange multipliers α, the constrained problem is becomes:

©2015 Engineering and Technology Publishing

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \quad (23)$$

The final hyperplane decision function is:

$$F(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i^0 y_i x_i \cdot x + b_0\right) \quad (24)$$

IV. Experiments and Results

In this section, three types of corpora are used to compare and demonstrate the proposed TII-based algorithm for recognize emotional state. In order to reasonably evaluate the performance of the proposed TII-based algorithm in each corpus, this paper uses a common emotional category among the corpus for testing.

A. The Corpora

First, we extract the common emotional category from three corpora: Emo-DB, eNTERFACE, and KHUSC-EmoDB. There are four types of emotional categories: Happiness, Fear, Sadness and Anger in Table II. In order to further evaluate the cross-corpus, the row labeled as "Mixed" is used to represent a mix of three abovementioned corpora. The total of mixed corpora is 1584 sentences.

B. The Evaluation of the Proposed TII-Based SER Using SVM on Three Corpora

To be fair its various emotional recognition rate, our experiments use minimum number of categories at various emotional speech corpuses as a test standard. The training set and test set are not both overlapped to achieve open test. For example, the 62-sentences is minima among four kind of emotions. We use 62-sentences as number of each emotional testing on EMO-DB speech database. Then, the number of test set and training set are 31, respectively. In addition, the 20-sentences is minima on "Happiness" emotion for eNTERFACE speech database. We use 103-sentences as test set and 104-sentences as training set, respectively. Because 102-sentences is same as each emotional category for KHUSC-EmoDB, test set and training set are both 51-sentences, respectively.

In Table III, the evaluation of TII-based SER using SVM on EMO-DB show that the recognition accuracy of "Anger" emotion can achieve 80.65% and outperform other emotional recognition accuracy. The Evaluation of TII-based SER using SVM can achieve 77.42% average accuracy of emotion recognition on EMO-DB. In other words, Table IV shows that the recognition accuracy of "Anger" emotion also can achieve 80.65%. The recognition accuracies of "Happiness" and "Fear", however, are lower than other emotional states. The average accuracy on eNTERFACE is about 73.06%, and it is lower than that on EMO-DB. In Table V, the evaluation results reveal that the proposed TII-based SER using SVM can model the emotional states on KHUSC-EmoDB except for in "Fear" emotional state. However, the performance on KHUSC-EmoDB is lower than that on EMO-DB. In conclusion, Table III-Table V show the confusion tables to evaluate the TII-based SER using SVM on EMO-DB, eNTERFACE and KHUSC-EmoDB. For the comparison among three corpora, the evaluation results of EMO-DB outperforms on other two corpora.

TABLE II. Description of the Collected Speech Database

Emotional category / Corpora	Happiness	Fear	Sadness	Anger	*Total*
EMO-DB	71	69	62	127	***329***
eNTERFACE	207	215	210	215	***847***
KHUSC-EmoDB	102	102	102	102	***408***
Mixed	***380***	***386***	***374***	***444***	***1584***

TABLE III. The Evaluation of TII-Based SER on EMO-DB

Target Emotion / Test Emotion	Recognition Rate (%)			
	Happiness	**Fear**	**Sadness**	**Anger**
Happiness	***77.42%***	3.23%	6.45%	12.90%
Fear	12.90%	***77.42%***	6.45%	3.23%
Sadness	12.90%	9.68%	***74.19%***	3.23%
Anger	12.90%	3.23%	3.23%	***80.65%***
Average Accuracy (%)	**77.42%**			

TABLE IV. The Evaluation of TII-Based SER on Enterface

Target Emotion / Test Emotion	Recognition Rate (%)			
	Happiness	**Fear**	**Sadness**	**Anger**
Happiness	***66.99%***	6.80%	9.71%	16.50%
Fear	5.83%	***64.08%***	19.42%	10.68%
Sadness	12.90%	9.68%	***74.19%***	3.23%
Anger	12.90%	3.23%	3.23%	***80.65%***
Average Accuracy (%)	**73.06%**			

TABLE V. The Evaluation of TII-Based SER on KHUSC-EMODB

Target Emotion / Test Emotion	Recognition Rate (%)			
	Happiness	**Fear**	**Sadness**	**Anger**
Happiness	***68.63%***	7.84%	5.88%	17.65%
Fear	13.73%	***56.86%***	19.61%	9.80%
Sadness	7.84%	21.57%	***66.67%***	3.92%
Anger	15.69%	13.73%	1.96%	***68.63%***
Average Accuracy (%)	**65.20%**			

C. Comparison with the Existing Method

In this subsection, a comparison of the proposed TII-based SER with the existing techniques will be presented. Table VI shows a comparison between the proposed method using TII feature set with SVM classifier and the two methods proposed by Ashish B. *et al.* for EMO-DB corpus [10]. The two methods use 1-D MFCC features with the HMM classifier and SVM classifier, respectively. Observing Table VI, it is found that the difference between the two methods is feature set only. In order to make a reasonable comparison, the common emotional categories: "Anger", "Happiness" and "Sadness" between the two methods is then extracted for comparing each other in next experiments.

©2015 Engineering and Technology Publishing

TABLE VI. A COMPARISON BETWEEN THE PROPOSED METHOD AND THE EXISTING METHOD [10]

Author	Corpus	Emotional State	Feature Set	Classifier
The proposed	EMO-DB	Happiness Fear Sadness Anger	TII (42-dimentional feature vectors)	SVM
Ashish B. *et al.*	EMO-DB	Happiness Sadness Anger Surprise neutral	MFCC (39-dimentional feature vectors)	HMM SVM

Table VII shows that the comparison evaluation of SER for three common emotional states. From the table, it is clear that the proposed feature, the TII feature set is visible for distinguishing "Happiness", which is superior to other two methods proposed by Ashish B., *et al*. In addition, an AHS Rate for three common emotional states: "Anger", "Happiness" and "Sadness" is 77.42% and is higher than other two methods.

TABLE VII. THE COMPARISON EVALUATION OF SER FOR THREE COMMON EMOTIONAL STATES

Emotional States / Methods	Anger	Happiness	Sadness	*AHS* Accuracy
Proposed method using TII with SVM	80.65%	77.42%	74.19%	***77.42%***
Ashish B. using MFCC with HMM	83.33%	57.14%	62.50%	***67.66%***
Ashish B. using MFCC with SVM	71.42%	57.14%	74.43%	***66.66%***

V. CONCLUSIONS

In this paper, we can find that the proposed SER system using TII feature set can provide significant performance for speech emotion sensing. Through spectrogram image calculation and Laws' masks, we can successfully extract texture image information to identify emotion status than the other conventional features. In addition, the proposed SER was verified using three corpora, uttered in three different languages.

It is critical to extract features that capture major temporal-spectral characteristics of signal to achieve a high accuracy in speech emotional classification. In conclusion, we find that the TII feature set derived from time-frequency representation can perform well for emotion classification. In future work, the TII-based feature extraction for SER can combine with facial features. This paper will apply the concept of texture image information into human emotional state from audiovisual signals. Therefore, the audiovisual signals, including speech and image, can be both processed with image processing to build one-kernel two-module (OKTM) system. Based on the OKTM system, we will significantly cost down for SER system. In addition, we find that the different language may cause variable performance in emotion recognition. This is worth for exploring the SER research directions.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funds was partially sponsored by the National Science Council, Taiwan, under contract number NSC 102-2221-E-158-006, and partially sponsored by the research project of Shih-Chien University, Taiwan, under contract number USC-102-08-01005 and USC-102-05-05014.

REFERENCES

[1] B. Adams, C. Breazeal, R. Brooks, and B. Scassellati, "Humanoid robots: A new kind of tool," *IEEE Intelligent Systems and Their Applications*., vol. 15, pp. 25-31, Jul. 2000.

[2] E. Kim, K. Hyun, S. Kim, and Y. Kwak, "Emotion interactive robot focus on speaker independently emotion recognition," in *Proc. Int. Conf. on Advanced Intelligent Mechatronics*., Sep. 2007, pp. 1-6.

[3] R. Cowie, *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32-80, Jan. 2001.

[4] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," in *Proc. of ICSLP 2000*, 2000, pp. 222-225.

[5] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2003, pp. 401-405.

[6] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," in *Proc. Eurospeech*, Sep. 2003, pp. 125-128.

[7] C. H. Park, K. S. Heo, D. W. Lee, Y. H. Joo, and K. B. Sim, "Emotion recognition based on frequency analysis of speech signal," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 2, no. 2, pp. 122-126, 2002.

[8] T. L. New, F. S. Wei, and L. C. DE Silva, "Speech based emotion classification," in *Proc. IEEE Region 10 International Conf. on Electrical and Electronic Technology*, Aug. 2001, pp. 297-301.

[9] X. H. Le, G. Quenot, and E. Castelli, "Recognizing emotions for the audio-visual document indexing," in *Proc. the Ninth IEEE International Symposium on Computers and Communications*, Jul. 2004, pp. 580-584.

[10] A. B. Ingale and D. S. Chaudhari, "Speech emotion recognition using hidden Markov model and support vector machine ," *International Journal of Advanced Engineering Research and Studies*, vol. 1, pp. 316-318, 2012.

[11] T. L. Nwe, S. W. Foo, and L. C. De-Silva, "Speech emotion recognition using hidden Markov models speech communication," vol. 41, no. 4, pp. 603-623, 2003.

[12] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems," in *Proc. Eurospeech*, Sep. 2003, pp. 729-732.

[13] R. S. Tato, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition," in *Proc. International Conference on Spoken Language Processing*, Sep. 2002, pp. 2029-2032.

[14] (Jun. 22, 2009). Berlin database of emotional speech. [Online]. Available: http://pascal.kgw.tu-berlin.de/emodb/

[15] O. Martin, I. Kotsia, and B. Macq, "The eNTERFACE'05 audio-visual emotion database," in *Proc. the 22nd International Conference on Data Engineering Workshops*, Los Alamitos, 2006.

[16] D. W. Griffn and J. S. Lim, "Signal estimation from modified short-time Fourier transforms," *IEEE Trans. Acoust. Speech Signal Process*, vol. 32, no. 2, pp. 236-248, 1984.

[17] S. Roucos and A. Wilgus, "High quality time scale modification for speech," in *Proc. IEEE International Conference on ASSP*, 1985, pp. 493-496.

[18] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice Hall PTR, 2002.

[19] L. He, M. Lech, N. Maddage, and N. Allen, "Emotion recognition in speech of parents of depressed adolescents," in *Proc. IEEE 3rd International Conference on Bioinformatics and Biomedical Engineering*, 2009, pp. 1-4.

©2015 Engineering and Technology Publishing

[20] L. He, M. Lech, N. Maddage, S. Memon, and N. Allen, "Emotion recognition in spontaneous speech within work and family environments," in *Proc. IEEE 3rd International Conference on Bioinformatics and Biomedical Engineering*, 2009, pp. 1-4.
[21] L. He, M. Lech; S. Memon, and N. Allen, "Recognition of stress in speech using wavelet analysis and teager energy operator," in *Proc. 9th Annual Conference, International Speech Communication Association and 12 Biennial Conference, Australasian Speech Science and Technology Association*, Australia, 2008.
[22] J. Dennis, H. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Process. Lett.*, vol. 18, no. 2, pp. 130-133, 2011.
[23] C. T. Lin and C. L. Chin, "Using fuzzy inference and cubic curve to detect and compensate backlight image," *International Journal of Fuzzy Systems*, vol. 8, no. 1, Mar. 2006.
[24] K. I. Laws, "Textured image segmentation," PhD Dissertation, University of Southern California. Los angles, 1980.
[25] A. R. Ganapathiraju, J. E. Hamaker, and J. Picone, "Application of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, Aug. 2004.

Kun-Ching Wang was born in Kaohsiung, Taiwan in 1976. He received the B.S. degree in electric engineering from Southern Taiwan University of Technology in 1998 and M.S. degree in electric engineering from Feng Chia University in 2000, and the Ph.D. degree in control engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2005. From 2005 to Jan. 2008, He was R&D in ITRI. Since Feb. 2008, he has been with the Department of Information Technology and Communication, Shih Chien University, Kaohsiung, Taiwan, where he is currently an Associate Professor. His research interests include speech and audio processing, speech and image recognition, speech enhancement, wavelet analysis and applications.

©2015 Engineering and Technology Publishing