A new speech-based emotion identification approach for a modern technology education system

J-C. Wang, C-H. Lin, W-K. Liao & W-J. Liao

National Central University
Thongli City, Taiwan

ABSTRACT: Owing to the emotion deficiency problem in many of the conventional technology education systems, examining emotion sensing has become a new research trend that makes it possible to provide useful strategies to enhance the learning effectiveness of a student. Among the various modalities for emotion sensing, this paper addresses speech-based emotion identification technology. In addition, an emotion feature set, comprising mel-frequency cepstral coefficients (MFCCs) and five perceptual features, i.e. spectrum power, sub-band powers, brightness, bandwidth and pitch are presented and discussed in this paper. The proposed feature set was fed into a frame-based multiclass support vector machine (SVM) for emotion identification. The superiority of the proposed system has been demonstrated via a seven-class emotional database with a 78.5% accuracy rate.

INTRODUCTION

In recent years, identifying a student's emotions to facilitate the learning process has become a new trend in technology education learning. Addressing the emotion deficiency problem in many of the conventional technology education approaches, sensing emotion is able to provide useful cues about a student's learning effectiveness. Proper teaching strategy and content should be adopted to enhance the learning.

For example, D'Mello et al developed an agile learning environment that is sensitive to a learner's affective state [1]. This design augmented an existing intelligent tutoring system and promoted learning. Facial expressions, gross body movements, and conversational cues were utilised to sense emotions. Luo and Tan applied facial emotion and speech emotion identification technologies to a distance education system [2]. For a predefined state of emotion, corresponding emotion encouragement and compensation have to be created.

A similar emotion identification technology based on speech has also been developed for a Web-based education system [3]. Recently, Tsai et al brought together speech recognition, emotion inference and virtual agents to implement a system for student interaction in an educational environment [4]. Motivated by these studies, a new speech-based emotion identification approach to be used in the modern technology education systems is presented in this paper.

EMOTION IDENTIFICATION

The proposed emotion identification system is described here. With received speech, non-silent frames of input waveform are identified first and used to form feature vectors. In this paper, a feature set, which includes spectrum power, sub-band powers, brightness, bandwidth, pitch and MFCCs is presented. A frame-based multiclass SVM is then used to perform the emotion identification.

Emotion Feature Set

Total spectrum power. Denote f_0 as the half sampling frequency. The total spectrum power is computed by:

$$P = \log(\int_0^{f_0} |F(f)|^2 df) . \quad (1)$$

Sub-band powers. The sub-band powers are extracted from the following sub-band intervals: $[0,0.125 f_0]$, $[0.125 f_0,0.25 f_0]$, $[0.25 f_0,0.5 f_0]$ and $[0.5 f_0, f_0]$. The *i*-th sub-band power is computed using the following expression:

$$P_i = \log(\int_{L_i(i)}^{H(i)} |F(f)|^2 df), \tag{2}$$

where $H(i)$ and $L(i)$ are the upper and lower bounds of the i-th sub-band.

Brightness. The brightness is the gravity centre of the power spectrum. It describes whether the power spectrum is dominated by low or high frequencies. Denote p_i as the power associated with frequency f_i, the brightness is calculated as:

$$F_C = \int_0^{f_0} f_i \cdot p_i df \Big/ \int_0^{f_0} p_i df \cdot \tag{3}$$

Bandwidth. The bandwidth is the second moment of the power spectrum. It describes whether the shape of the power spectrum is concentrated near its centroid or spread out over the spectrum as follows:

$$F_B = \sqrt{\int_0^{f_0} (f_i - F_C)^2 p_i df \Big/ \int_0^{f_0} p_i df} \cdot \tag{4}$$

Pitch. A simple pitch detection algorithm, based on detecting the peak of the normalised autocorrelation function, was used. The pitch frequency is returned if the peak value is above a threshold, or the frame is labelled as non-pitched.

Besides the above mentioned perceptual features, mel-frequency cepstral coefficients, which model the human auditory perception system, were used. The derivation of MFCCs is based on the powers of the theses critical-band filters. The MFCCs can be found from logarithm and cosine transforms.

Emotion Classifier

This study proposes an emotion classifier using a frame-based multiclass SVM. The input waveform is segmented into separate frames. Passing through the procedure of feature extraction, each frame will be transformed into a feature vector. Assume a N_F - frame utterance, $\bar{x}^{(j)}$, $j = 1, \ldots, N_F$, is to be classified into emotion class C_m, $m \in \{1,2,\ldots,M\}$. The steps for emotion identification based on frame-based multiclass SVM follows. First, for each emotion class C_m, and for all the classes C_n ($n \neq m$), one can compute:

$$score_H(C_{m,n} | \bar{\mathbf{x}}^{(j)}) = \sum_{j=1}^{N_F} H\left(\overline{\mathbf{w}}\overline{\mathbf{x}}^{(j)} + b\right) - \sum_{j=1}^{N_F} H\left(-(\overline{\mathbf{w}}\overline{\mathbf{x}}^{(j)} + b)\right), \tag{5}$$

by the C_m-C_n 2-calss SVM. In Equation (5), $H(\cdot)$ is the Heaviside step function.

The accumulated score for each emotion class C_m is then computed using the following formula:

$$score(C_m | \bar{x}^{(j)}) = \sum_n score(C_{m,n} | \bar{x}^{(j)}) \cdot \tag{6}$$

Finally, the most possible emotion class C_{m^*} is chosen by:

$$m^* = \arg\max_m score(C_m | \bar{x}^{(j)}) \cdot \tag{7}$$

EXPERIMENTAL RESULTS

The German emotion speech database consisting of utterances with seven different emotions (anger, joy, sadness, fear, disgust, boredom and neutral) was used for the experiments. For each emotion class, half of the audio files were utilised for training and the others were used for testing.

The frame size is 512 samples (32 ms), with 50% overlap in each of the two adjacent frames. For assessing emotion identification results, the accuracy rate, which is defined as the ratio between correct-classified utterance number and the total testing utterance number, was used. With the proposed emotion classifier and feature set, the accuracy rate can achieve approximately 78.5%.

CONCLUSIONS

Identifying a student's emotions to facilitate the learning process is a new trend in technology education learning. This paper proposed and outlined an effective emotion feature set, which includes mel-frequency cepstral coefficients

(MFCCs) and five perceptual features. The proposed emotion identification system uses this feature set and a frame-based multiclass support vector machine. The experimental work has proved that the proposed system is able to identify 7-class emotions with about 78.5% accuracy rate.

It is envisaged that future work should endeavour to integrate the proposed emotion identification system into computer-based and Web-based technology education systems. The initial research work has already been undertaken in this respect.

REFERENCES

1. D'Mello, S.K., Craig, S.D., Gholson, B., Franklin, S., Picard, R. and Graesser, A.C., Integrating affect sensors in an intelligent tutoring system. *Proc. Inter. Conf. on Intelligent User Interfaces*, 7-13 (2005).
2. Luo, Q. and Tan, H., Facial and speech recognition emotion in distance education system, *Proc. Inter. Conf. on Intelligent Pervasive Computing*, 483-486 (2007).
3. Gong, M. and Luo, Q., Speech emotion recognition in web based education, *Proc. IEEE Inter. Conf. on Grey Systems and Intelligent Services*, 1,082-1,086 (2007).
4. Tsai, I-H., Lin, K.H-C., Sun, R-T., Fang, R.-Y., Wang, J-F., Chen, Y-Y., Huang, C-C. and Li, J-S., , Application of educational Emotion Inference via Speech and Agent Interaction, *Proc. Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, 129-133 (2010).