Emotion Recognition Using IG-based Feature Compensation and Continuous Support Vector Machines

Chung-Hsien Wu and Ze-Jing Chuang

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
{chwu, bala}@csie.ncku.edu.tw

Abstract

This paper presents an approach to feature compensation for emotion recognition from speech signals. In this approach, the intonation groups (IGs) of the input speech signals are firstly extracted. The speech features in each selected intonation group are then extracted. With the assumption of linear mapping between feature spaces in different emotional states, a feature compensation approach is proposed to characterize the feature space with better discriminability among emotional states. The compensation vector with respect to each emotional state is estimated using the Minimum Classification Error (MCE) algorithm. For the final emotional state decision, the IG-based feature vectors compensated by the compensation vectors are used to train the Continuous Support Vector Machine (CSVMs) for each emotional state. The emotional state with the maximal output probability is determined as the final output. The kernel function of CSVM model is experimentally decided as Radial basis function and the experimental result shows that IG-based feature extraction and compensation can obtain encouraging performance for emotion recognition.

1. Introduction

Human-machine interface technology has been investigated for several decades. Recent research made more emphasis on the recognition of nonverbal information, especially on the topic of emotion reaction. Scientists have found that emotional skills can be an important component of intelligence, especially for human-human communication. Although human-computer interaction is different from human-human communication, some theories have shown that human-computer interaction is essentially following the basics of human-human interaction [1]. In our study, an emotion recognition approach from speech signals is proposed. This method consists of the definition and extraction of intonation groups (IGs), IG-based feature extraction, and feature compensation.

In the past years, many researchers have paid their attention to emotion recognition via speech signals. Several important recognition models have been applied to the emotion recognition task, such as Neural Network (NN) [2], Hidden Markov Model (HMM) [3], and Support Vector Machine (SVM)[4][5]. Besides the generally used prosodic and acoustic features, some special features are also applied for this task, such as TEO-based features [6]. Although lots of features and recognition models have been tested in these works, large overlaps between the feature spaces for different emotional states is rarely considered. Besides, the pre-trained emotion recognition model is highly speaker-dependent.

To solve the above questions, this paper proposes an approach to emotion recognition based on feature compensation. The block diagram of the approach is shown in Figure 1. The feature extraction process is shared by training and testing phase and is divided into two steps: intonation group identification and IG-based feature extraction. In order to identify the most significant segment, the intonation groups (IGs) of the input speech signals are firstly extracted. Following the feature extraction process [7], the prosodic feature sets are estimated for the IG segments. Then in training phase, the extracted feature vectors are applied for compensation vector estimation. In this process, the minimum classification error (MCE) training method [8] is adopted to iteratively estimate all compensation vectors until reach a best classification result. Finally, the compensated vectors are used to train the CSVM model. In the testing phase, the extracted feature vectors are directly compensated using the compensation vectors. Then the final emotional state is decided using the CSVM model.

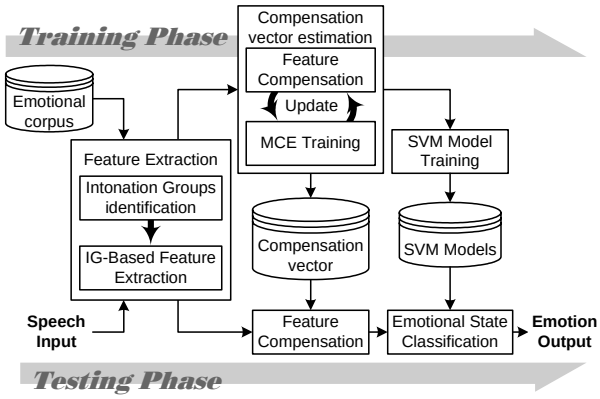

Figure 1: *Block diagram of the proposed emotion recognition approach*

2. Feature Extraction

2.1. Intonation Group Extraction

The intonation group, also known as breath-groups, tone-groups, or intonation phrases, is usually defined as the segment of an utterance between two pauses.

As shown in Figure 2, the intonation group is identified by analyzing the smoothed pitch contour (the gray-thick line in Figure 2). Three types of smoothed pitch contour patterns are defined as the intonation group:

- **Type 1**: a complete pitch segment that starts from the point of a pitch rise to the point of the next pitch rise,

- **Type 2**: a monotonically decreasing pitch segment,
- **Type 3**: a monotonically increasing pitch segment.

For all identified IG segments, only those IGs that match the following criterion are selected for feature extraction:

- the complete IGs with the largest pitch range or duration,
- the monotonically decreasing or increasing IGs with the largest pitch range or duration,
- the monotonically decreasing or increasing IGs at the start or end of a sentence.

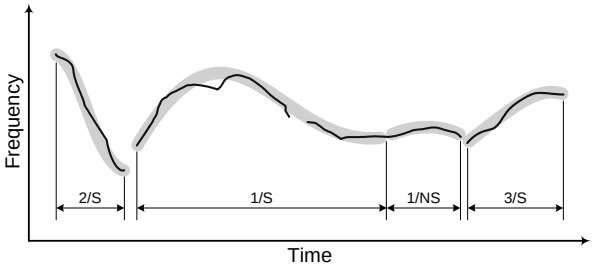

Figure 2: *An illustration of the definition and extraction of Intonation Groups. Four IGs are extracted from the smoothed pitch contour (the gray-thick line), but only three IGs (the first, second, and forth IGs) are selected for feature extraction*

In Figure 2, the numbers before the slash symbol indicate the type of IG, and the symbol S and NS indicate "Selected" and "Not-Selected", respectively. Although there are four IGs extracted, only three IGs are selected for feature extraction.

2.2. IG-based Feature Extraction

Emotional state can be characterized by many speech features, such as pitch, energy, or duration [9]. In this paper, we use the following 64 prosodic features as the input features for emotion recognition:

- Speaking rate and relative duration (2 values),
- Pause number and relative pause duration (2 values),
- Average, standard deviation, maximum, and minimum of pitch, energy, zero-crossing-rate, and F1 (formant one) values (16 values),
- Average and standard deviation of jitter (for pitch) and shimmer (for energy) (4 values),
- Relative positions where the maximal and minimal pitch, energy, zero-crossing-rate, and F1 value occur (8 values),
- Fourth-order Legendre parameters of pitch, energy, zero-crossing-rate, and F1 contours of the whole sentence and the "rapidest part" (32 values), which is the segment between the positions with the maximum and minimum values.

Jitter is a variation of individual cycle lengths in pitch-period measurement, and shimmer is the same measurement for energy [10].

3. Compensation Vector Estimation Using MCE

The goal of feature compensation is to move the feature space of an emotional state to a feature space more discriminative to other emotional states. Given a sequence of the training data $\mathbf{X}^e = \left\{x_n^e\right\}_{n=1}^N$, where x_n^e indicates the *n*-th feature vector that belongs to emotional state E_e. The feature vector extracted for each intonation group contains the prosodic features mentioned above. With the assumption of linear mapping between feature spaces in different emotional states, the vector compensation function is defined as:

$$\tilde{x}_n^{e\to f} = x_n^e + p\left(E_e \middle| x_n^e\right) r_{e\to f}, \tag{1}$$

where $r_{e\to f}$ is a compensation vector of emotional state E_e with respect to the reference emotional state E_f. The conditional probability of the emotional state E_e given the input feature vector x_n^e is estimated as:

$$p\left(E_e \middle| x_n^e\right) = \frac{p\left(x_n^e \middle| E_e\right) p\left(E_e\right)}{\sum_i p\left(x_n^e \middle| E_i\right) p\left(E_i\right)}. \tag{2}$$

Minimum classification error (MCE) training based on the generalized probabilistic descent (GPD) method is applied in our study. We assume that the probability of a mapped feature vector $\tilde{x}_n^{e\to f}$ given an emotional state E_c follows the distribution of a mixture of Gaussian density function:

$$g_c\left(\tilde{x}_n^{e\to f}\right) = \sum_m w_m^c \cdot N\left(\tilde{x}_n^{e\to f}; \mu_m^c, \delta_m^c\right), \tag{3}$$

where $N\left(\cdot; \mu_m^c, \delta_m^c\right)$ denotes the normal distribution with mean μ_m^c and diagonal covariance matrix δ_m^c, and w_m^c is the mixture weight. To estimate the mapping coefficients and GMM parameters jointly by MCE training, the misclassification measure is defined as:

$$D_e = -g_e\left(\mathbf{X}_e\right) + \frac{1}{\eta}\log\left[\frac{1}{\mathbf{C}-1}\sum_{c\neq e}\exp\left(\eta \cdot g_c\left(\mathbf{X}_e\right)\right)\right], \tag{4}$$

where $\mathbf{X}_e$ denotes a set of data compensated from the emotional state E_e, $\mathbf{X}_e = \left\{\tilde{x}_n^{e\to f}\right\}_{f\neq e}$, $\mathbf{C}$ is the number of emotional state, and η is a penalty factor. The function $g_c\left(\mathbf{X}_e\right)$ is the average likelihood estimated by the GMM of the emotional state $\boldsymbol{E}_c$ given $\mathbf{X}_e$. Based on the GPD iterative theory, the parameters will approximate the global optimization using the iterative equation:

$$\Theta_{t+1} = \Theta_t - \varepsilon \cdot \nabla l, \tag{5}$$

The loss function is defined as a sigmoid function of misclassification measure. And the gradient of loss function ∇l is the partial differential to the updated parameter. Using chain rule, the gradient of loss function can be divided into three components. The first component can be derived to a closed form $a \cdot l_e \cdot \left(1 - l_e\right)$, and the second component is assumed as:

$$\frac{\partial D_e}{\partial g_c} = \begin{cases} -1 & , e = c \\ 1 & , e \neq c \end{cases}. \tag{6}$$

Since there are four different parameters needed to be updated, the last component of the gradient with respect to each parameter is obtained as:

$$\frac{\partial g_e}{\partial r_{e\to f}} = -\mathbf{A}\sum_n\sum_m\left[\frac{w_m^e\left(\tilde{x}_n^{e\to f}-\mu_m^e\right)p\left(E_e\left|x_n^e\right.\right)}{\left(\delta_n^e\right)^2}N\left(x_n^e;\mu_m^e,\delta_n^e\right)\right] \quad (7)$$

$$\frac{\partial g_e}{\partial w_m^e} = \mathbf{A}\sum_n\sum_r\mathbf{B} \quad (8)$$

$$\frac{\partial g_e}{\partial \mu_m^e} = \mathbf{A}\sum_n\sum_r\left[w_m^e\left(\tilde{x}_n^{e\to f}-\mu_m^e\right)\left(\delta_m^e\right)^{-2}\mathbf{B}\right] \quad (9)$$

$$\frac{\partial g_e}{\partial \delta_m^e} = \mathbf{A}\sum_n\sum_e\left[w_m^e\left(\left(\tilde{x}_n^{e\to f}-\mu_m^e\right)^2-\left(v_m^e\right)^2\right)\left(v_m^e\right)^{-3}\mathbf{B}\right]. \quad (10)$$

where

$$\mathbf{A}=\frac{1}{\mathbf{N}(\mathbf{C}-1)} \quad , \quad \mathbf{B}=N\left(\tilde{x}_n^{e\to f};\mu_m^e,\delta_m^e\right)$$

Given an input feature vector y, the recognized emotional state is determined according to the following equation:

$$E_e^* = \arg\max_e\left[\frac{\sum_{i\neq e}g_e\left(y+p\left(E_e\left|y\right.\right)r_{e\to i}\right)}{\sum_{j\neq e}g_j\left(y+p\left(E_j\left|y\right.\right)r_{j\to e}\right)}\right]. \quad (11)$$

4. Experimental Results

In this experiment four kinds of emotional states: Neutral, Happy, Angry, and Sad were adopted. The emotional speech corpus was collected in 8KHz and 16bits. 40 sentences for each emotional state were recorded by 8 volunteers.

Besides the proposed prosodic features, we also evaluated the recognition rate for Mel-Frequency Cepstrum Coefficient (MFCC) features, which is generally used in speech recognition task. To investigate the performance of the proposed method, we tested both the proposed method and a baseline system, which is a CSVM emotion recognition system without any preprocessing before feature extraction.

4.1. Emotion recognition using CSVM models

The SVM has been widely applied in many research areas, such as data mining, pattern recognition, linear regression, and data clustering. Given a set of data belonging to two classes, the basic idea of SVM is to find a hyperplane that can completely distinguish two different classes. The illustration of SVM model is shown in Figure 3. The hyperplane is decided by the maximal margin of two classes, and the samples that lie in the margin are called "support vectors." The equation of the hyperplane is described as:

$$D(x)=\sum_{i=1}^{N}y_i k\left(x\cdot x_i\right)+w_0\,, \quad (12)$$

where $k\left(x\cdot x_i\right)$ is kernel function. Traditional SVMs can construct a hard decision boundary with no probability output. In this study, SVMs with continuous probability output are proposed. Given the test sample *x'*, the probability that *x'* belongs to class *c* is $P(class_c|x')$. This value is estimated based on the following factors:

- the distance between the test input and the hyperplane,

$$R=\frac{D(x')/\|w\|}{1/\|w\|}=D(x')\ ; \quad (13)$$

- the distance from the class centroid to the hyperplane,

$$R'=\frac{R}{D(\bar{x})}=\frac{D(x')}{D(\bar{x})}\ ; \quad (14)$$

  where $\bar{x}$ is the centroid of the training data in a class;
- the classification confidence of the class P_c, which is defined as the ratio of correctly recognized sentences number to total sentence number.

Finally, the output probability is defined as follows according to the above factors:

$$P\left(class_c\left|x'\right.\right)=\frac{P_c}{1+\exp(1-R')}=\frac{P_c}{1+\exp\left(1-\frac{D(x')}{D(\bar{x})}\right)}\ . \quad (15)$$

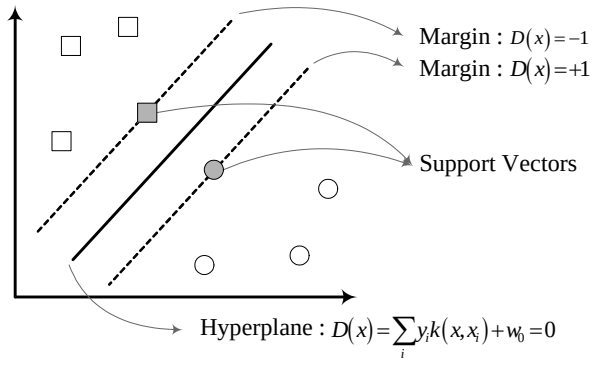

Figure 3: *An illustration of SVM. The vectors on the margins are so-called "Support Vectors"*

4.2. Experiments on SVM Kernel Function

The kernel function defined in CSVM model is used to transfer a vector in original vector space to a new space with higher dimension. There are several popularly used kernel functions:

- Simple dot

$$k(x,y)=x\cdot y \quad (16)$$

- Vovk's polynomial

$$k(x,y)=(x\cdot y+1)^p \quad (17)$$

- Radial basis function

$$k(x,y)=\exp\left(-\|x-y\|^2/2\delta^2\right) \quad (18)$$

- Sigmoid kernel

$$k(x,y)=\tanh\left(k(x,y)-\Theta\right) \quad (19)$$

In order to select a most appropriate kernel function, a primary test of emotion recognition using CSVM model with different kernel function is applied. The primary test used both prosodic and MFCC feature with feature compensation and intonation group. The result of emotion recognition is shown in Table 1.

Table 1: *The primary test of different kernel functions.*

	Prosodic	MFCC
Simple dot	59.00%	63.48%
Vovk's polynomial	60.83%	90.12%
Radial basis function	80.72%	95.12%
Sigmoid kernel	73.50%	81.46%

It is obviously that Radial basis function is the most suitable kernel function for our test.

4.3. Experiments on Emotion Recognition

Table 2 shows the results for emotion recognition, including the proposed approach and the baseline system. In the first column, the abbreviation FC, In, OO, and OC indicate the method using feature compensation, the results from Inside, Outside-Open, and Outside-Closed tests, respectively. The first row in Table 2 shows four kinds of speech feature from left to right: frame-based prosodic, frame-based MFCC, IG-based prosodic, and IG-based MFCC feature.

Table 2: *The emotion recognition result.*

	P	M	P+IG	M+IG
-FC (In)	74.33%	99.96%	76.32%	99.24%
-FC (OO)	49.78%	35.15%	51.07%	35.01%
-FC (OC)	55.95%	37.22%	59.90%	42.13%
+FC (In)	80.72%	95.12%	83.94%	91.32%
+FC (OO)	55.19%	41.27%	60.13%	49.10%
+FC (OC)	61.03%	41.03%	67.52%	52.86%

Although MFCC feature outperforms prosodic features in the inside test, prosodic features achieved better performance in both outside-open and outside-closed tests. The reason of this result is that the MFCC features contain much information from speech content and speaker. To model the emotional state using MFCC features become to model the speech content and speaker. Therefore, the CSVM model can better classify the emotional states of trained MFCC features, but cannot well characterize the unseen features in outside test. In the proposed approach, MFCC features remain its higher recognition rate in inside test, and the prosodic features obtain the best overall performance. From the above experiments, an increase in recognition rate for the approaches with IG-based feature extraction is about 5% to 10% compared to that without IG-based feature extraction. Furthermore, an improvement of 10% in recognition rate for the approach with feature compensation is obtained compared to that without feature compensation.

5. Conclusion

In this paper, an approach to emotion recognition from speech signals is proposed. In order to obtain crucial features, the IG-based feature extraction method is used. After feature extraction, the feature vector compensation approach and MCE training method are applied to increase the discriminability among emotional states. The experiments show that it is useful to integrate IG-based feature extraction and feature compensation to emotion recognition. The result of emotion recognition using the proposed approaches is 83.94% for inside test and 60.13% for outside-open test. We also demonstrate that the prosodic feature is more suitable for emotion recognition than the acoustic MFCC features in speaker-independent task.

The future work of this research is to improve the recognition accuracy for outside data. Though the feature compensation is useful for emotion recognition, the compensation vector is still speaker-dependent. An adaptation method will be useful to adapt compensation vectors for emotional speech with different speaking styles.

6. References

[1] Reeves, B.; Nass, C., 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. University of Chicago Press.

[2] Bhatti, M.W.; Wang, Y.; Guan, L., 2004. A neural network approach for human emotion recognition in speech. *IEEE International Symposium on Circuits and Systems*, Vancouver, Canada, 181-184

[3] Inanoglu, Z.; Caneel, R., 2005. Emotive alert：HMM-based emotion detection in voicemail mes-sages. *IEEE Intelligent User Interfaces '05*, San Diego, California, USA, 251-253

[4] Kwon, O.W.; Chan, K.; Hao, J.; Lee, T.W., 2003. Emotion Recognition by Speech Signals. *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 125-128.

[5] Chuang, Z.J.; Wu, C.H., 2004. Multi-Modal Emotion Recognition from Speech and Text. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 1-18.

[6] Rahurkar, M.A.; Hansen, J.H.L., 2003. Frequency Distribution Based Weighted Sub-Band Approach for Classification of Emotional/Stressful Content in Speech. *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 721-724.

[7] Deng, L.; Droppo, J.; Acero, A., 2003. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Transactions on Speech and Audio*, 11(6), 568-580.

[8] Wu, J.; Huo, Q., 2002. An environment compensated minimum classification error training approach and its evaluation on aurora2 database. *7th International Conference on Spoken Language*, Denver, Colorado, USA, 453-456.

[9] Ververidis, D.; Kotropoulos, C.; Pitas, I., 2005. Automatic emotional speech classification. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Montreal, Canada, 593-596.

[10] Levity, M.; Huberz, R.; Batlinery, A.; Noeth, E., 2001. Use of prosodic speech characteristics for automated detection of alcohol intoxication. *Prosody in Speech Recognition and Understanding*, Molly Pitcher Inn, Red Bank, NJ.