

Emotion Recognition from Speech Using IG-Based Feature Compensation

Chung-Hsien Wu*, and Ze-Jing Chuang*

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 extracted first. 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 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 compensated IG-based feature vectors are used to train the Gaussian Mixture Models (GMMs) and Continuous Support Vector Machine (CSVMs) for each emotional state. For GMMs, the emotional state with the GMM having the maximal likelihood ratio is determined as the final output. For CSVMs, the emotional state is determined according to the probability outputs from the CSVMs. The kernel function in CSVM is experimentally decided as a Radial basis function. A comparison in the experiments shows that the proposed IG-based feature compensation can obtain encouraging performance for emotion recognition.

Keywords: Emotional Speech, Emotion Recognition, Intonation Group, Feature Compensation

1. Introduction

Human-machine interface technology has been investigated for several decades. Recent research has put 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

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

have shown that human-computer interaction essentially follows the basics of human-human interaction [Reeves *et al.* 1996; Picard 1997; Cowie *et al.* 2001]. Scientists have found that emotion technology can be an important component in artificial intelligence, especially for human-human communication [Salovey *et al.* 1990]. Although the human-computer interaction is different from human-human communication, some theories show that human-computer interaction basically follows the fundamental forms of human-human interaction [Reeves *et al.* 1996]. In this 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 past years, many researchers have paid attention to emotion recognition via speech signals. Several important recognition models have been applied to the emotion recognition task, such as Neural Network (NN) [Bhatti *et al.* 2004], Hidden Markov Model (HMM) [Inanoglu *et al.* 2005], Support Vector Machine (SVM) [Kwon *et al.* 2003; Chuang *et al.* 2004], and others [Subasic *et al.* 2001; Wu *et al.* 2006; Silva *et al.* 2000]. Besides the generally used prosodic and acoustic features, some special features are also applied for this task, such as TEO-based features [Rahurkar *et al.* 2003]. 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 [Chuang *et al.* 2004; Wu *et al.* 2004].

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 the training and testing phase and is divided into two steps: intonation group (IG) identification and IG-based feature extraction. In order to identify the most significant segment, the intonation groups (IGs) of the input speech signals are first extracted. Following the feature extraction process [Deng *et al.* 2003], the prosodic feature sets are estimated for the IG segments. Then, in training phase, the extracted feature vectors are applied for compensation vector estimation. All the feature vectors compensated by compensation vectors are modeled by a Gaussian Mixture Model (GMM). Finally, the minimum classification error (MCE) training method [Wu *et al.* 2002] iteratively estimates all the model parameters. As a comparison, the compensated vectors are also used to train the Continuous Support Vector Machine (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.

The rest of the paper is organized as follows. Section 2 describes the definition of Intonation Group and the extraction of the prosodic features. Then the feature compensation technique and MCE training is provided in Section 3. The model description of CSVM is shown in Section 4. Finally, experimental results and conclusions are drawn in Section 5 and 6,

respectively.

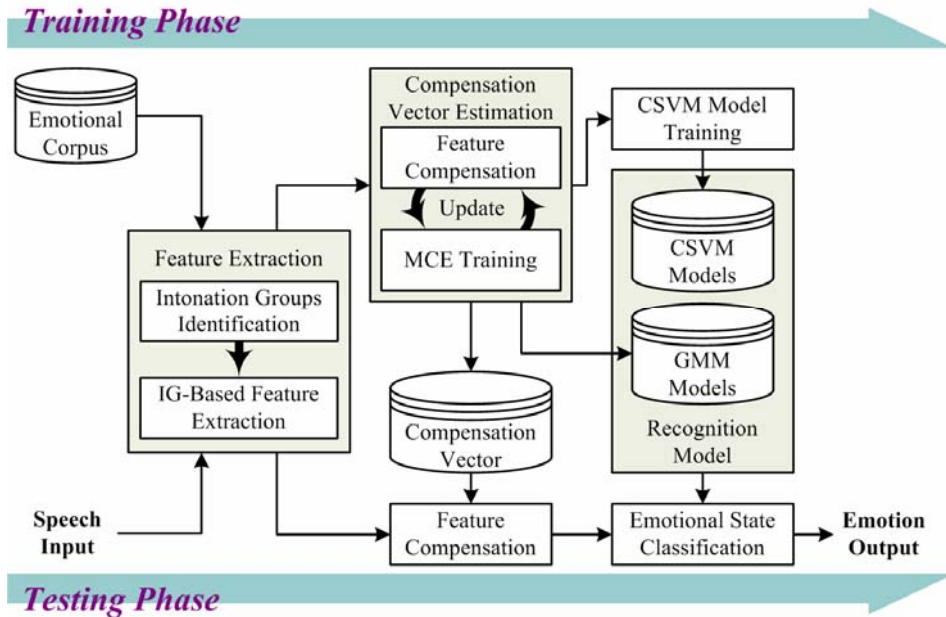


Figure 1. Block diagram of the proposed emotion recognition approach

2. IG-Based 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.

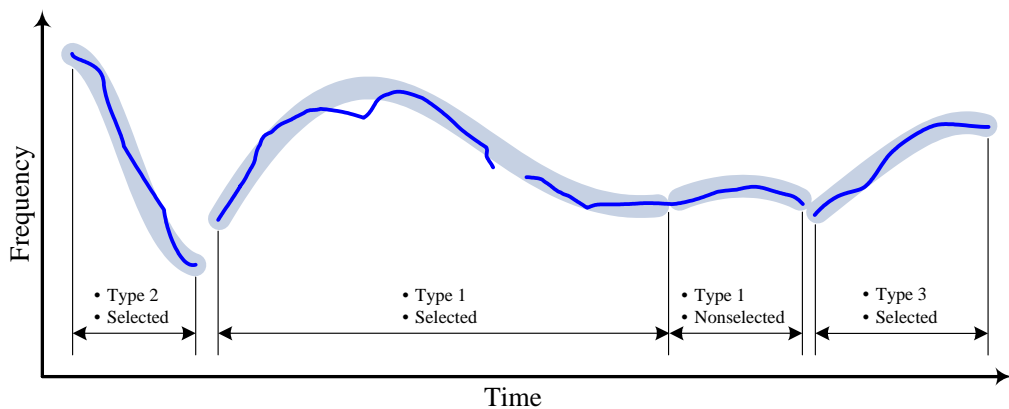


Figure 2. An illustration of the definition and extraction of Intonation Groups. Four IGs are extracted from the smoothed pitch contour (the gray-thick line), but only three IGs (the first, second, and fourth IGs) are selected for feature extraction

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.

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” IGs, 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 [Ververidis *et al.* 2005]. In this paper, the authors use the following 64 prosodic features as the input features for emotion recognition:

- Speaking rate and relative duration (2 values). The relative duration is normalized with respect to the length of the input sentence.
- Pause number and relative pause duration (2 values). The relative duration is normalized with respect to the length of intonation group. The same definition of relative position and relative duration are made in the following features.
- Mean and standard deviation of pitch, energy, zero-crossing-rate, and F1 values (8 values).
- Mean and standard deviation of jitter (for pitch) and shimmer (for energy) (4 values).
- Maximum and minimum of pitch, energy, zero-crossing-rate, and F1 values (8 values).
- Relative positions at which the maximal pitch, energy, zero-crossing-rate, and F1 value occur (4 values).
- Relative positions at which the minimal pitch, energy, zero-crossing-rate, and F1 value occur (4 values).
- Fourth-order Legendre parameters of pitch, energy, zero-crossing-rate, and F1 contours of the whole sentence (16 values).

- Fourth-order Legendre parameters of pitch, energy, zero-crossing-rate, and F1 contours inside the “significant segment,” which is the segment during the positions of maximum and minimum values (16 values).

The definitions of jitter and shimmer are given in [Levity *et al.* 2001]. Jitter is a variation of individual cycle lengths in pitch-period measurement, while shimmer is the measure for energy values. The calculation of the jitter contour is shown as:

$$J_i = \frac{\sum_{u=0}^m f_i - ug_u}{f_i}, \quad g = \frac{1}{4}\{-1, 3, -3, 1\}, \quad (1)$$

where variable f_i indicates the i -th pitch value. The mean and standard deviation of jitter are calculated by equations 2 and 3, respectively.

$$J_{mean} = \frac{\sum_{J_i < J_\theta} J_i}{\# J_i},$$

$$J_{dev} = \sqrt{\frac{\sum_{J_i < J_\theta} (J_i - J_{mean})^2}{\# J_i}}. \quad (2)$$

where variable J_θ is a threshold of jitter used to avoid the outlier noise.

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 training data $\mathbf{X}^e = \{x_n^e\}_{n=1}^N$, 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:

$$\tilde{x}_n^{e \rightarrow f} = x_n^e + p \left(E_e \mid x_n^e \right) r_{e \rightarrow f}, \quad (3)$$

where $r_{e \rightarrow f}$ 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:

$$p(E_e | x_n^e) = \frac{p(x_n^e | E_e) p(E_e)}{\sum_i p(x_n^e | E_i) p(E_i)}. \quad (4)$$

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

$$g_c(\tilde{x}_n^{e \rightarrow f}) = \sum_m w_m^c \cdot N(\tilde{x}_n^{e \rightarrow f}; \mu_m^c, \delta_m^c), \quad (5)$$

where $N(\cdot; \mu_m^c, \delta_m^c)$ 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:

$$D_e \equiv D(\mathbf{X}_e) = -g_e(\mathbf{X}_e) + \frac{1}{\eta} \log \left[\frac{1}{\mathbf{C}-1} \sum_{c \neq e} \exp(\eta \cdot g_c(\mathbf{X}_e)) \right], \quad (6)$$

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

$$\Theta_{t+1} = \Theta_t - \varepsilon \cdot \nabla l, \quad (7)$$

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 $a \cdot l_e \cdot (1 - l_e)$, and the second component is assumed as:

$$\frac{\partial D_e}{\partial g_c} = \begin{cases} -1 & , e = c \\ 1 & , e \neq c \end{cases}. \quad (8)$$

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

$$\frac{\partial g_e}{\partial r_{e \rightarrow f}} = -\mathbf{A} \sum_n \sum_m \left[\frac{w_m^e (\tilde{x}_n^{e \rightarrow f} - \mu_m^e) p(E_e | x_n^e)}{(\delta_n^e)^2} N(x_n^e; \mu_m^e, \delta_n^e) \right], \quad (9)$$

$$\frac{\partial g_e}{\partial w_m^e} = \mathbf{A} \sum_n \sum_r \mathbf{B}, \quad (10)$$

$$\frac{\partial g_e}{\partial \mu_m^e} = \mathbf{A} \sum_n \sum_r \left[w_m^e \left(\tilde{x}_n^{e \rightarrow f} - \mu_m^e \right) \left(\delta_m^e \right)^{-2} \mathbf{B} \right], \quad (11)$$

$$\frac{\partial g_e}{\partial \delta_m^e} = \mathbf{A} \sum_n \sum_e \left[w_m^e \left(\left(\tilde{x}_n^{e \rightarrow f} - \mu_m^e \right)^2 - \left(v_m^e \right)^2 \right) \left(v_m^e \right)^{-3} \mathbf{B} \right]. \quad (12)$$

where

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

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

$$E_e^* = \arg \max_e \left[\frac{\sum_{i \neq e} g_e \left(y + p(E_e | y) r_{e \rightarrow i} \right)}{\sum_{j \neq e} g_j \left(y + p(E_j | y) r_{j \rightarrow e} \right)} \right]. \quad (13)$$

4. Emotion recognition using CSVM models

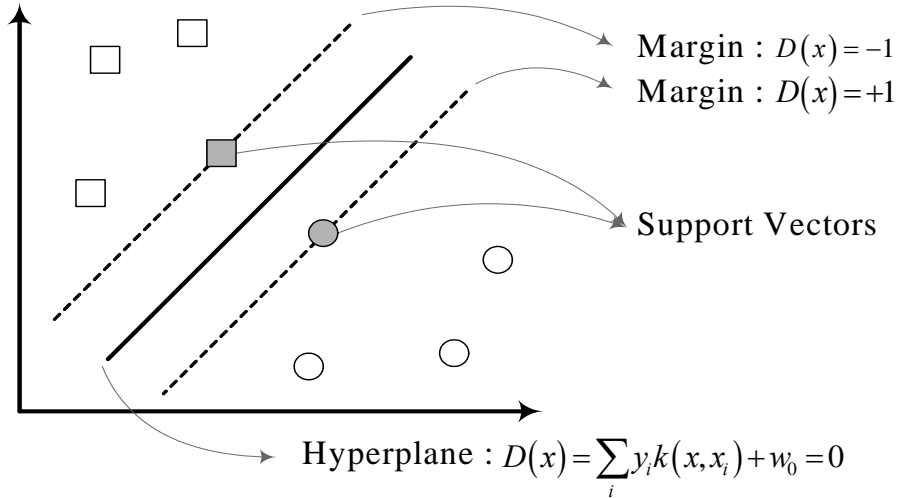


Figure 3. An illustration of SVM. The vectors on the margins are the so-called “Support Vectors”

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:

$$D(x) = \sum_{i=1}^N y_i k(x \cdot x_i) + w_0, \quad (14)$$

where $k(x \cdot x_i)$ 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,

$$R = \frac{D(x')/\|w\|}{1/\|w\|} = D(x') ; \quad (15)$$

- the distance from the class centroid to the hyperplane,

$$R' = \frac{R}{D(\bar{x})} = \frac{D(x')}{D(\bar{x})} ; \quad (16)$$

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 the total sentence number.

Finally, the output probability is defined as follows, in accordance with the above factors:

$$P(class_c|x') = \frac{P_c}{1 + \exp(1 - R')} = \frac{P_c}{1 + \exp\left(1 - \frac{D(x')}{D(\bar{x})}\right)}. \quad (17)$$

The CSVM model with the highest probability determines the emotion output.

5. Experimental Results

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

In addition to the proposed prosodic features, the researchers also evaluated the recognition rate for Mel-Frequency Cepstrum Coefficient (MFCC) features, which are generally used in a speech recognition task. Both GMMs and CSVMs are applied in the experiments to compare with the baseline system, which is a GMM emotion recognition system without any preprocessing before feature extraction.

5.1 Mixture Number Determination of GMM

The number of mixtures in the GMM is first determined for each emotional state. Assuming that the number of mixtures is greater than 3, the average likelihood of all training data given the GMM with different mixture number is calculated. Figure 4 shows the plot of GMM likelihoods using both prosodic and MFCC features. Accordingly, the numbers of mixtures using prosodic features is set to 6, 5, 12, and 9 for neutral, angry, happy, and sad emotion, respectively. To evaluate MFCC features, the number of mixtures was also evaluated using the same method. The contours of the likelihood using MFCC features is shown in the lower part of Figure 4, and the mixture numbers for neutral, angry, happy, and sad emotions are set to 13, 11, 18, and 20, respectively.

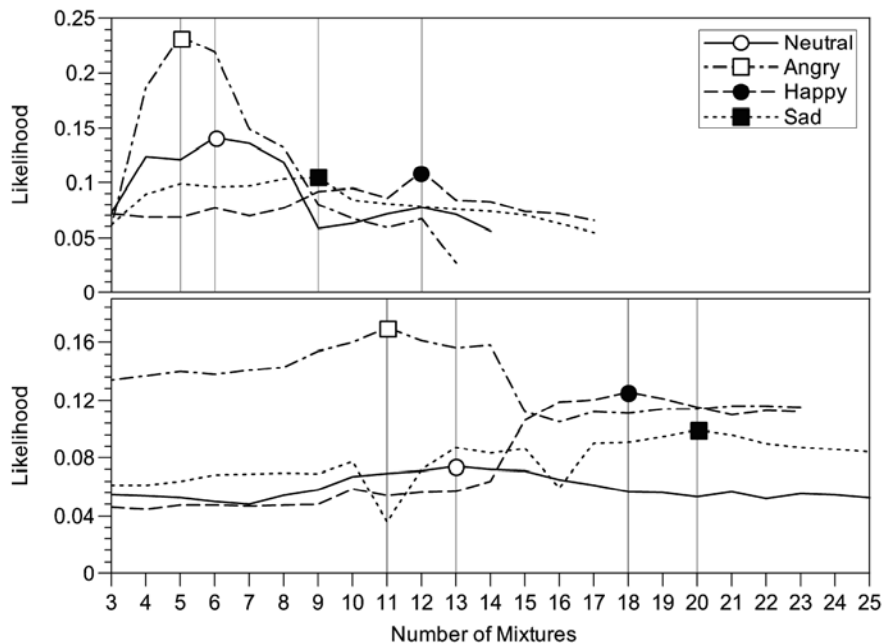


Figure 4. Likelihood contours of GMMs with increasing mixture number. The upper part is the plot using prosodic features, and the lower part is the plot using MFCC features

5.2 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

$$k(x, y) = x \cdot y \quad (18)$$

- Vovk's polynomial

$$k(x, y) = (x \cdot y + 1)^p \quad (19)$$

- Radial basis function

$$k(x, y) = \exp\left(-\|x - y\|^2 / 2\delta^2\right) \quad (20)$$

- Sigmoid kernel

$$k(x, y) = \tanh(k(x, y) - \Theta) \quad (21)$$

In order to select the most appropriate kernel function, a primary test of emotion recognition using CSVM model with different kernel functions was applied. The primary test used both prosodic and MFCC features with feature compensation and intonation group. The result of emotion recognition is shown in Table 1. It is obvious that the Radial basis function is the most suitable kernel function for this test.

Table 1. The primary test for different kernel functions.

	Prosodic Feature	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%

5.3 Experiments on Emotion Recognition using GMM

Table 2 shows the results for emotion recognition using GMM, including the proposed approach and the baseline system. In the first column, the abbreviations **FC**, **In**, **OO**, and **OC** indicate the methods 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 features from left to right: frame-based prosodic feature, frame-based MFCC feature, IG-based prosodic feature, and IG-based MFCC feature.

Table 2. Emotion recognition results using GMMs. The abbreviation FC indicates the method using Feature Compensation. The abbreviation In, OO, and OC indicate the results from Inside, Outside-Open, and Outside-Closed tests, respectively

	Prosodic Feature	MFCC	Prosodic Feature+IG	MFCC+IG
Without FC (In)	74.33%	99.96%	76.32%	99.24%
Without FC (OO)	49.78%	35.15%	51.07%	35.01%
Without FC (OC)	55.95%	37.22%	59.90%	42.13%
With FC (In)	80.72%	95.12%	83.94%	91.32%
With FC (OO)	55.19%	41.27%	60.13%	49.10%
With FC (OC)	61.03%	41.03%	67.52%	52.86%

Although MFCC feature outperformed prosodic features in the inside test, prosodic features achieved better performance in both outside-open and outside-closed tests. The reason for this result is that the MFCC features contain a considerable amount of information from the speech content and the speaker. Emotional state modeling using MFCC features is highly related to speech content and speaker. Therefore, the GMM model can better classify the emotional states of the trained MFCC features, but cannot well characterize the unseen features in the outside test. In the proposed approach, MFCC features retain their higher recognition rate in the 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 those without IG-based feature extraction. Furthermore, an improvement of 10% in recognition rate for the approaches with feature compensation is obtained compared to those without feature compensation.

5.4 Experiments on Emotion Recognition using CSVM

Table 3. Emotion recognition results using CSVMs.

	Prosodic Feature	MFCC	Prosodic Feature+IG	MFCC+IG
Without FC (In)	78.23%	98.12%	81.19%	99.15%
Without FC (OO)	50.83%	32.76%	53.79%	34.91%
Without FC (OC)	58.10%	38.48%	60.94%	40.33%
With FC (In)	83.10%	92.71%	86.08%	91.73%
With FC (OO)	57.36%	40.08%	62.12%	47.60%
With FC (OC)	62.55%	41.98%	68.00%	53.25%

Table 3 shows the results for emotion recognition using CSVM. General speaking, the results of emotion recognition using CSVM are similar to the results using GMM. The proposed feature extraction method, IG-based prosodic feature with feature compensation, obtains the

best overall performance. As seen in the result shown in Table 2, the MFCC feature attains a better recognition rate than the prosodic feature in the inside test, but worse in the outside test. The difference between using CSVM and GMM is the suitability of CSVM in a data sparse situation. Since CSVM constructs the classification hyperplane using only few support vectors, it attains better classification results than other classification methods when the training corpus is insufficient. In the proposed method, the MFCC feature is extracted frame by frame, while the prosodic feature is extracted by a single sentence or IG segment. It is obvious that, under the same number of training corpora, the number of MFCC features is larger than the number of prosodic features. Accordingly, with the prosodic features, most results using CSVM are better than that using GMM.

6. 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 with GMM is 83.94% for an inside test and 60.13% for an outside-open test, and the result with CSVM is 86.08% for an inside and 62.12% for an outside-open test. This result shows that the CSVM classification model is more suitable than GMM when performing the emotion recognition task. The authors 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.

References

- Bhatti, M.W., Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," In *Proceedings of the 2004 IEEE International Symposium on Circuits and Systems*, Vancouver, Canada, pp. 1811-1184.
- Chuang, Z.J., and C.H. Wu, "Multi-Modal Emotion Recognition from Speech and Text," *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 2004, pp. 45-62.

- Chuang, Z.J., and C.H. Wu, "Emotion Recognition using Acoustic Features and Textual Content," In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, 2004, Taipei, Taiwan, pp. 53-56.
- Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, 18(1), 2001, pp. 32-80.
- Deng, L., J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Transactions on Speech and Audio*, 11(6), 2003, pp. 568-580.
- Inanoglu, Z., and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," In *Proceedings of IEEE Intelligent User Interfaces*, 2005, San Diego, California, USA, pp. 251-253.
- Kwon, O.W., K. Chan, J. Hao, and T.W. Lee, "Emotion Recognition by Speech Signals," In *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland, pp. 125-128.
- Levity, M., R. Hubertz, A. Batlinery, and E. Noeth, "Use of prosodic speech characteristics for automated detection of alcohol intoxication," In *Proceedings of the Workshop on Prosody and Speech Recognition 2001*, Red Bank, NJ, pp. 19-22.
- Picard, R.W., *Affective Computing*, Cambridge, MA: MIT Press, 1997.
- Rahurkar, M.A., and J.H.L. Hansen, "Frequency Distribution Based Weighted Sub-Band Approach for Classification of Emotional/Stressful Content in Speech," In *Proceedings of 8th European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland, pp. 721-724.
- Reeves, B., and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, University of Chicago Press, 1996.
- Reeves, B., and C. Nass, *The Media Equation*, Center for the Study of Language and Information, 1996.
- Salovey, P., and J. Mayer, "Emotional Intelligence," *Imagination, Cognition and Personality*, 9(3), 1990, pp. 185-211.
- Silva, L.C.D., and P.C. Ng, "Bimodal Emotion Recognition," In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, Grenoble, France, pp. 332-335.
- Subasic, P., and A. Huettner, "Affect Analysis of Text Using Fussy Semantic Typing," *IEEE Transactions on Fussy System*, 9(4), 2001, pp. 483-496.
- Ververidis, D., C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, Montreal, Montreal, Canada, pp. 593-596.
- Wu, J., and Q. Huo, "An environment compensated minimum classification error training approach and its evaluation on aurora2 database," In *Proceedings of 7th International Conference on Spoken Language*, 2002, Denver, Colorado, USA, pp. 453-456.

- Wu, C.H. and Z.J. Chuang, "Intelligent Ear for Emotion Recognition: Multi-Modal Emotion Recognition via Acoustic Features, Semantic Contents and Facial Images," In *Proceedings of The 8th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2004, Orlando, Florida, USA.
- Wu, C.H., Z.J. Chuang, and Y.C. Lin, "Emotion Recognition from Text using Semantic Label and Separable Mixture Model," *ACM Transactions on Asian Language Information Processing*, 5(2), 2006, pp. 165-183.