PAPER

A Hybrid Speech Emotion Recognition System Based on Spectral and Prosodic Features

Yu ZHOU[†a)], **Junfeng LI**[††b)], **Yanqing SUN**[†], **Jianping ZHANG**[†], **Yonghong YAN**[†], ***Nonmembers***, ***and*** **Masato AKAGI**[††], ***Member***

SUMMARY In this paper, we present a hybrid speech emotion recognition system exploiting both spectral and prosodic features in speech. For capturing the emotional information in the spectral domain, we propose a new spectral feature extraction method by applying a novel non-uniform subband processing, instead of the mel-frequency subbands used in Mel-Frequency Cepstral Coefficients (MFCC). For prosodic features, a set of features that are closely correlated with speech emotional states are selected. In the proposed hybrid emotion recognition system, due to the inherently different characteristics of these two kinds of features (e.g., data size), the newly extracted spectral features are modeled by Gaussian Mixture Model (GMM) and the selected prosodic features are modeled by Support Vector Machine (SVM). The final result of the proposed emotion recognition system is obtained by combining the results from these two subsystems. Experimental results show that (1) the proposed non-uniform spectral features are more effective than the traditional MFCC features for emotion recognition; (2) the proposed hybrid emotion recognition system using both spectral and prosodic features yields the relative recognition error reduction rate of 17.0% over the traditional recognition systems using only the spectral features, and 62.3% over those using only the prosodic features.

key words: *speech emotion recognition, non-uniform subband processing, spectral feature, prosodic feature*

1. Introduction

As one of the most natural and important means in "human-human" interaction, speech communication consists of two channels, the explicit channel carrying the linguistic information (i.e., the content of the conversation) and the implicit channel carrying the non-linguistic information (e.g., gender, age, emotion, dialect) [1], [2]. Both linguistic and non-linguistic information play crucial roles in human speech communication. Many studies in linguistic information processing (e.g., speech recognition and speech synthesis) have been done in the past several decades; however, the research on non-linguistic cues has only recently become popular [3].

Non-linguistic cues generally include gender, age, emotion, stress and nervousness, dialect, and so on [1]. Among these properties, emotion plays a key role in many applications, for example, in text-to-speech systems to synthesize emotional speech [4]. So far, different approaches have been presented to model emotions: one approach is the definition of discrete basic emotions, for example, anger, disgust, fear, happiness, sadness, and surprise, as proposed by Ekman [5]; another approach is the utilization of continuous emotional dimensions, for instance, the three-dimensional emotional space: arousal (activation), potency, and valence, as proposed by Schlosberg [6].

In this paper, we present our recent study on automatic emotion recognition from speech, which has received much attention for building more intuitive "human-machine" interfaces [7]. Emotion recognition is basically a statistical pattern classification problem, which consists of two major steps, feature extraction and classification. While the theory of classification is pretty well developed [2], the extraction of distinctive features is a highly empirical issue [2]. Therefore, our main focus in this research is to extract more effective features and apply them in one emotion recognition system.

So far, many speech emotion recognition systems have already been reported [7]. In these existing systems, however, the features exploited for emotion recognition are generally the mel-frequency cepstrum coefficients (MFCC) [7], [8], which is often used in automatic speech recognition (ASR) systems. For speech recognition, the features should emphasize the content of speech (i.e., linguistic information). In contrast, the features to be used in emotion recognition systems should be able to highlight the discriminative cues among different emotions (i.e., non-linguistic information), rather than linguistic information of speech. This essential difference means that the MFCC features that are suitable for ASR systems do not satisfy the requirements for emotion recognition from speech. As a result, the systems using only MFCC features achieve very limited emotion recognition results, as reported in [7]–[9]. Furthermore, prosodic features of speech, which were found to be useful cues for representing speech emotion information in phonetics and linguistics [1], were also used for speech emotion recognition systems [2], [7]. However, the recognition systems using prosodic features alone demonstrated much poorer recognition performance than those using spectral features [2]. Many algorithms of utilizing both spectral and prosodic features have been proposed. In [10], prosodic features, MFCC, and formant features were investigated, and only the mean and standard deviation of per-frame MFCC features were extracted for each utterance, then the spectral

Manuscript received February 2, 2010.
Manuscript revised May 5, 2010.

[†]The authors are with Institute of Acoustics, Chinese Academy of Sciences, China.
[††]The authors are with School of Information Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan.
a) E-mail: zhouyu@hccl.ioa.ac.cn
b) E-mail: junfeng@jaist.ac.jp

DOI: 10.1587/transinf.E93.D.2813

and prosodic features were combined directly at the feature-level, which might cause a loss of information. Many other systems use the per-frame MFCC features, and then integrate them with prosodic features, such as the phoneme-level modeling in [11]. In [11], per-frame based MFCC features was suggested to be complementary to the suprasegmental prosodic features, however, the improvement is not obvious after combining the spectral features with prosodic features. Besides, the prosodic features used in this study mainly come from F0 and the performance could be further improved including much wider range of prosodic features [11].

To address the problems of traditional emotion recognition systems, In this paper, we propose a new spectral feature extraction approach using a novel non-uniform subband processing technique that is designed based on the speech emotion production mechanism. Moreover, to make use of the complementary information prosodic features may provide, a much wider range of prosodic features closely related to emotion states in speech are selected from the previous studies [1], [2] compared with [11]. Both the newly extracted spectral features and the selected prosodic features are exploited in the proposed emotion recognition system. Because of the differences in size of the feature vectors between the non-uniform spectral features and the prosodic features, two different classifiers are applied in our proposed emotion recognition system, namely, the Gaussian Mixture Model (GMM) classifier for the non-uniform spectral features, and the Support Vector Machine (SVM) classifier for the prosodic features. The proposed hybrid emotion recognition system yields the final decision through combining the results from both classifiers. Experimental results show that (1) the proposed non-uniform spectral feature is more effective than the traditional MFCC feature for emotion recognition; (2) the proposed emotion recognition system using both spectral and prosodic features yields the relative recognition error reduction rates of 17.0% over traditional recognition systems using only spectral features and 62.3% over using only the prosodic features.

The remainder of this paper is structured as follows. In Sect. 2, we show the overview of the proposed speech emotion recognition system. In Sect. 3, we describe the GMM-based emotion recognition subsystem using the spectral features extracted by a new non-uniform subband processing technique. In Sect. 4, we introduce the SVM-based emotion recognition subsystem using prosodic features. In Sect. 5, the proposed hybrid emotion recognition system is described by combining the GMM-based subsystem and the SVM-based subsystem. In Sect. 6, experiments are performed to evaluate the proposed non-uniform spectral features and the proposed hybrid emotion recognition system. Finally, Sect. 7 draws some conclusions.

2. Overview of the Proposed Hybrid Emotion Recognition System

As one pattern recognition system, the proposed emotion recognition system includes: a training procedure to train the models based on the extracted features, and a testing procedure to recognize emotions from speech using the trained models. More specifically, the proposed hybrid emotion recognition system consists of a GMM-based subsystem using spectral features and a SVM-based subsystem using prosodic features. The block diagram of the proposed system is shown in Fig. 1.

In the GMM-based subsystem, the new spectral features based on a novel non-uniform subband processing technique are applied to each frame of input signal. In the SVM-based subsystem, the prosodic features are exploited for each utterance, rather than each frame signal, which means that little data is available for prosodic features in comparison with the spectral features. This difference further leads to different modeling approaches being utilized for individual feature sets. That is, in the training process, spectral features are modeled using GMM, while prosodic features are modeled using SVM. Those trained models are further used for emotion classification in the recognition procedure. The final decision on emotion recognition of the proposed system is realized by combining the recognition scores of two recognition sub-systems.

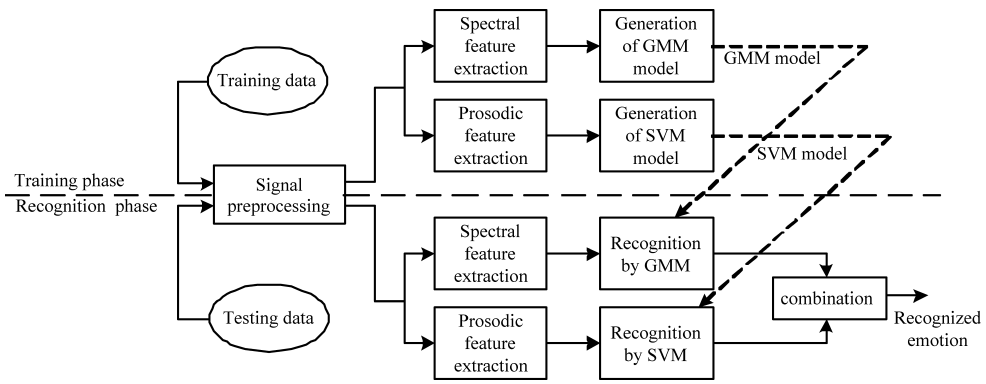


Fig. 1 Block diagram of the proposed hybrid emotion recognition system.

3. GMM-Based Emotion Recognition Subsystem with Non-uniform Spectral Features

In this section, a new non-uniform subband processing technique is proposed on the basis of the emotion production mechanism [12], from which the new spectral features are extracted to highlight the emotion information in speech. Finally, a GMM emotion recognition subsystem is constructed using the new non-uniform spectral features.

3.1 Novel Non-uniform Subband Processing

Recent research on emotion in speech production showed that horizontal shift of tongue tip is observed when emotion changes [12], which concerns the third formant of vowels (i.e., around 3000 Hz) [13], and glottal information (e.g., 100 ∼ 400 Hz) also contributes greatly to emotion expression in speech [2]. That is, emotion information in speech is encoded unevenly in frequency regions, which provides the basic motivation of the proposed non-uniform sub-band processing.

3.1.1 Calculation of Mutual Information

To design the non-uniform sub-band processing technique, the different contribution of each frequency band in emotional speech production is proposed to be quantified through investigating the dependency between the output of each frequency band and emotional states. Specifically, the input emotional speech signal is first divided into several subbands by a group of triangle-shaped band-pass filters with linear frequency scale. For an emotional speech feature X and emotional state Y, the dependency of the emotional state on each frequency band is then formulated using the mutual information measure, defined in [14]

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{1}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, and $H(X,Y)$ is the joint entropy of X and Y. The entropy of X is defined as

$$H(X) = -\sum_{x\in X} p(x)\log p(x). \tag{2}$$

where x is a value of an emotional speech feature X, which is the output of each frequency band, and y is a label value of emotional classes. Specific to our study, as x is a continuous stochastic variable, its probability distribution function (PDF) is estimated by discretizing x and represented by a histogram [15], and then the marginal entropy is calculated using Eq. (2). The detailed steps are as follows:

- Find the minimum and maximum values of x, i.e., x_{min}, and x_{max}, as the boundary.
- Divide $[x_{min}, x_{max}]$ equally into I intervals, with the length of each segment equals Δ_x.
- Count the number of samples in the i-th interval, denote as k_i.
- With the total number of samples equals N, the marginal entropy of x is calculated as $H(X) = -\sum_{i=1}^{I}\left(\frac{k_i}{N}\log\frac{k_i}{N}\right) + \log\Delta_x$

As y is a discrete variable, the possible values of Y could be enumerated. Denote K is the number of emotion classes, the marginal entropy of Y is calculated as $H(Y) = -\sum_{j=1}^{K}\left(\frac{k_j}{N}\log\frac{k_j}{N}\right) + \log\Delta_y$, where k_j is the number of samples of the jth emotion class, and $\Delta_y = 1$. As in the training data, the samples of different emotion classes are approximately the same, k_j is approximated as $\frac{N}{K}$, and $H(Y) \approx \log K$.

The calculation of joint entropy of $H(X,Y)$ using histogram is similar to the calculation of $H(X)$.

- The same with the first two steps of the calculation of $H(X)$.
- Count the number of samples in the i-th interval for the j-th emotion class, denote as k_{ij}.
- With the total number of samples equals N, the joint entropy of X and Y is calculated as $H(X,Y) = -\sum_{i=1}^{I}\sum_{j=1}^{K}\left(\frac{k_{ij}}{N}\log\frac{k_{ij}}{N}\right) + \log(\Delta_x\Delta_y)$

With the entropies $H(X)$, $H(Y)$, and joint entropy $H(X,Y)$, the mutual information $I(X;Y)$ is finally computed using Eq. (1).

3.1.2 Design of Non-uniform Subband Processing

Given each emotional speech utterance, the speech is framed and windowed by a hamming window, then the FFT is carried on each frame. Using uniform spaced filter bands, the output of each frequency band could be obtained. Then the mutual information between the output of each frequency band and the emotional state could be calculated using above methods on each frequency band. After the mutual information are obtained for each subband using Eq. (1), a frequency-dependent mutual information curve is obtained by plotting each mutual information value at the center of its frequency region. To exemplify the non-uniform distribution of emotion in speech according to frequencies, we used the CASIA Mandarin emotional speech corpus, which was designed and collected for emotion recognition study, provided by Chinese-LDC [16]. This database contains short utterances from four persons, covering five emotions (i.e., angry, happy, surprised, neutral and sad). For each person, there are 1500 utterances (i.e., 300 utterances for each emotion) with the sampling frequency of 16 kHz. The emotion discriminating ability of each frequency band is quantified using mutual information criterion with 2000 utterances (i.e., 100 utterances for each person and each emotion), as shown in Fig. 2.

Figure 2 indicates that different frequency bands are characterized by frequency-dependent mutual information indices, corresponding to their different contributions to

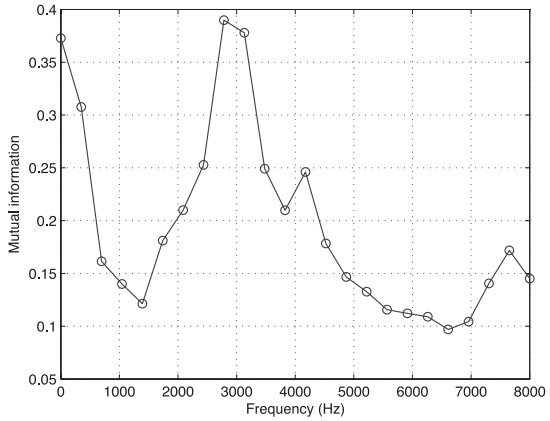


Fig. 2 Emotional speech discriminative score in frequency domain using mutual information.

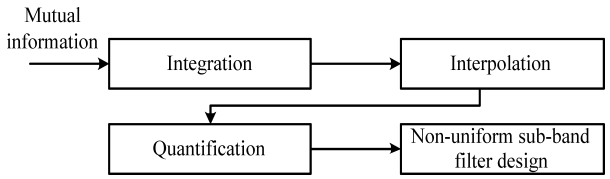


Fig. 3 The design procedure of the non-uniform sub-band processing.

emotion in speech. More specifically, two frequency regions, the region with frequencies less than 300 Hz (corresponding to glottal information) and that with frequencies around 3000 Hz (corresponding to the movement of tongue tip), play the important roles in emotion discrimination, which is also consistent with the study on emotional speech production [12], [13].

3.1.3 Implementation of Non-uniform Subband Processing

The non-uniform distribution of emotion information of speech in the frequency domain is realized by a novel subband processing technique. The basic idea behind this proposed non-uniform subband processing technique is that: the frequency regions with high mutual information (i.e., contribute more to emotion discrimination) should be emphasized through more subband filters with narrower bandwidth (i.e., high frequency resolution); while those with low mutual information should be de-emphasized through using fewer subband filters with wider bandwidth (i.e., low frequency resolution).

The implementation of the proposed non-uniform subband processing technique is shown in Fig. 3. The bandwidth of each subband is determined according to the reciprocal of mutual information in that corresponding subband. It is not easy to calculate the bandwidth for each subband directly from the reciprocal value of mutual information when the boundary of the subband is not fixed, therefore, the number of subbands that is proportional to the mutual information is instead considered first. Since scaling does not change the distribution function, the normalized distribution function of the number of bands is the same as that of the mutual information. Assuming the mutual information value for each subband is MI_i, $(i \in 1 \ldots 24)$, first the cumulative sum of Md_i is calculated to obtain the distribution function of the mutual information as:

$$Md_i = \frac{\Sigma_{j=1}^{i} MI_j}{\Sigma_{j=1}^{24} MI_j}, i \in 1 \ldots 24 \tag{3}$$

Second, the distinct distribution function is interpolated from the frequency domain to the FFT space using cubic spline interpolation [17], which is the Mc_j. When the sampling rate is 16 kHz and the window size is 25 ms, the number of FFT is 512, then $j \in 1 \ldots 256$. Third, given the target number of subbands N_s, the distribution is mapped from $[0, 1]$ to $1..N_s$ by linear transformation and Rounding to ceiling integers as:

$$Cn_j = ceil(Mc_j * N_s), j \in 1 \ldots 256 \tag{4}$$

where $ceil(x)$ signifies the least integer which is no less than x. Then the new map from FFT point to the corresponding channel number is gotten as Cn_j. The data chosen for the mutual information analysis is the CASIA Mandarin emotional speech corpus, including 5 emotions from 4 speakers.

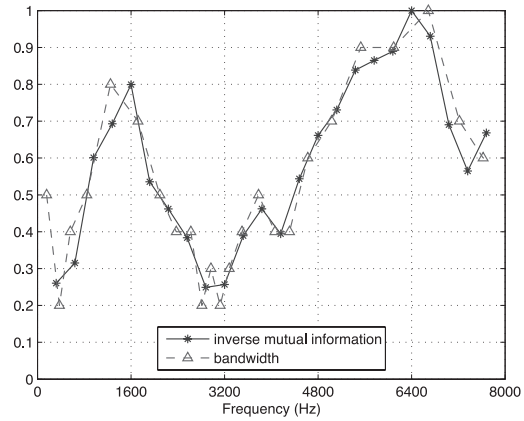


Fig. 4 Bandwidth of non-uniform sub-bands according to the reciprocal value of mutual information. The vertical axis shows the relative values of bandwidth and the reciprocal of mutual information.

The obtained bandwidths of all subbands are plotted with frequency in Fig. 4, where the ranges of curves have been normalized for comparison. Figure 4 demonstrates that the bandwidths of the designed subbands follow the changing tendency of the reciprocal value of mutual information. The designed non-uniform subband filters are shown in Fig. 5, along with the mel-frequency subband filters for comparison.

3.2 Novel Spectral Features Extraction Using Non-uniform Subband Processing

The non-uniform subband processing technique is further

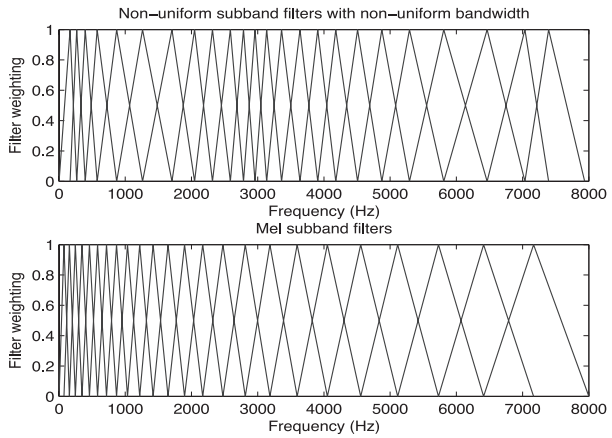


Fig. 5 The mel-frequency subband filters in MFCC features, and the proposed non-uniform subband filters in NUFCC features.

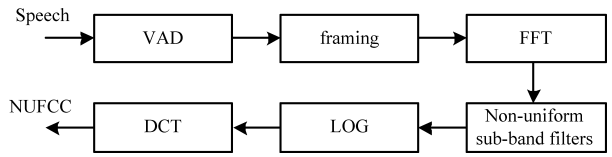


Fig. 6 The extraction diagram of the proposed spectral features, non-uniform frequency cepstral coefficient (NUFCC).

exploited to design the new spectral features, namely non-uniform cepstral coefficients (NUFCC). The NUFCC features are extracted using a similar procedure, shown in Fig. 6, to that in the calculation of mel-frequency cepstral coefficient (MFCC) which is generally used in speech recognition systems. Specifically, a voice activity detector (VAD) is first used to eliminate pauses in speech (while the pauses within one utterance are kept, since they contain some emotion information [18]); the hamming window with 25 ms frame length and 10 ms shift is applied to provide the windowed short-time frames. The 512-sample Fast Fourier Transform (FFT) is then exploited, followed by the proposed non-uniform subband processing to output the spectrum for each subband. Finally, logarithm and discrete cosine transform (DCT) are adopted to generate 12 order cepstral coefficients with energy. Note that the difference of calculation procedures between NUFCC and MFCC is that different subband design criterions are used, namely, mel-frequency subbands for MFCC and non-uniform subbands for NUFCC. We believe that this proposed NUFCC feature could be more suitable for emotion recognition task because of its frequency-dependent emotion discrimination ability (i.e., the non-uniform sub-band processing), which will be verified through an experiment.

3.3 GMM-Based Emotion Recognition Subsystem

In our emotion recognition system, the newly designed NUFCC features are modeled using GMM, which has been widely used in state-of-the-art emotion recognition. GMM can provide a smooth approximation to the underlying distribution of the feature vectors of the speech signal. GMM assumes that the probability distribution of the mixture density used for the likelihood function of a D-dimensional feature vector, x, is defined as:

$$p(x|\lambda_j) = \sum_{i=1}^{M} \pi_i N(x : \mu_i, \Sigma_i) \text{ for } j = 1, 2..K \tag{5}$$

where K is the number of emotion classes, i.e., the number of emotion models, π_i are the mixture weights, $N(x : \mu_i, \Sigma_i)$ are the multivariate normal distribution, and M is the number of the components. Each component densities are parameterized as Eq. (6) with $\sum_i$ is a covariance matrix and μ_i is a mean vector.

$$N(x : \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{6}$$

The mixture weight, π_i, must satisfy the constraint $\Sigma_{i=1}^{M} \pi_i = 1$. The complete density model parameters can be written as Eq. (7)

$$\lambda_j = [\pi_1, \ldots, \pi_M, \mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M] \tag{7}$$

There are several methods available for estimating parameters of a GMM [19], and the most popular method is the Expectation Maximization (EM), which is based on the maximum likelihood (ML) criterion. The mixture weight, π_i, mean vector, μ_i, and covariance matrix Σ_i can be determined using the EM algorithm iteratively [20], [21]. To choose suitable orders of GMM, orders above 512 were not studied, considering insufficient training data for properly training the emotion models. For orders varying from 32 to 512, experiments which were not included here had shown that our best result is achieved for 512 mixture components with spectral features. The parameters of the GMM model are estimated during the training stage, and further used in the emotion recognition stage for the likelihood probability calculation.

Given an observation x, a spectral feature sample for emotion recognition, the posterior probability of being the j-th emotion model λ_j is defined as:

$$p(\lambda_j|x) = \frac{p(\lambda_j, x)}{p(x)} = \frac{p(x|\lambda_j)p(\lambda_j)}{\sum_{i=1}^{K} p(x|\lambda_i)p(\lambda_i)} \tag{8}$$

where $p(x|\lambda_j)$ is the likelihood of seeing the observation x given the j-th emotion model. Generally, all the emotion models are assumed to have equal priori, so this term could be removed from both the numerator and denominator, and the posterior probability $p(\lambda_j|x)$ can be obtained using Eq. (8) and Eq. (5).

4. SVM-Based Emotion Recognition Subsystem with Prosodic Features

As mentioned in the Introduction, in addition to spectral features, prosodic features have been found to be closely related to speech emotional states. In this paper, therefore,

prosodic features are also used in speech emotion recognition. According to the studies on prosodic features and emotional speech [1], [2], [8], the following prosodic features, namely, fundamental frequency, loudness features, voice source features and harmonicity features, are selected and implemented in the emotion recognition system. In the recognition system, more specifically, the prosodic features include the 11 dimensional fundamental frequency feature vector, 20 dimensional loudness feature vector, 42 dimensional voice source feature vector and 14 dimensional harmonicity feature vector. The detailed description of these prosodic features is shown in [1]. In total, 87 dimensional prosodic feature is extracted for each utterance and further used in the recognition system. For emotion recognition, the extracted prosodic features are modeled by support vector machine (SVM), which is motivated by the high robustness of SVM classifier in the scenario when only limited training data are available [22]. This is essentially because the prosodic features are extracted for each utterance, unlike the spectral features for each short-time frame.

C-Support Vector Classification was chosen as the SVM type, and the radial basis function (RBF) kernel was used in this study [22]. Crucial parameters for training a SVM are the value of variance (γ) in the RBF and the penalty parameter (C) allowing us how strictly we want the classifier to fit the training data [22]. However, LIBSVM provides a parameter selection tool: cross validation via parallel grid search [22]. First, the training data is separated to several folds. Sequentially a fold is considered as the validation set and the rest are for training. The average of accuracy on predicting the validation sets is the cross validation accuracy. Then, a possible interval of C (or γ) with the grid space was provided, and all grid points of (C, γ) are tried to see which one gives the highest cross validation accuracy. The best parameter was used to train the whole training set and generate the final model, which were used to predict the class label for a given utterance [22].

SVM was originally designed for binary classification. As emotion recognition is a multi-class decision problem, SVMs should be extended efficiently for this purpose. In [23], three methods of constructing a multiclass classifier by combining several binary classifiers were compared: 'one-against-all' (or '1-vs-rest'), 'one-against-one' and the directed acyclic graph SVM (DAGSVM), as well as other methods considering all classes at once. The results indicated that 'one-against-one' and DAG methods are more suitable for practical use than the other method, with less training time, and higher accuracy. So in this paper, the 'one-against-one' method is adopted to calculate the pairwise class probabilities [24], which were further used to get the class-dependent-probabilities for combination with the GMM system, with details given below.

For a multi-class decision problem, where the number of classes is k, given any input x, the posterior probability of being the i-th emotion model:

$$p_i = p(y = i|x), i = 1, 2, \ldots, k. \tag{9}$$

First, pairwise class probabilities r_{ij} was estimated using an improved implementation [25] of [26]

$$r_{ij} \approx p(y = i|y = i \text{ or } j, x) \approx \frac{1}{1 + e^{A\hat{f}+B}}, \tag{10}$$

Where A and B were estimated by minimizing the negative log-likelihood function using known training data and their decision value $\hat{f}$ [25]. Labels and decision values were obtained through conducting the above cross-validation.

Then the approach which was brought up in [24] is used to obtain p_i from r_{ij}'s. It solves the following optimization problem:

$$\begin{aligned} &\min_p \frac{1}{2} \sum_{i=1}^{k} \sum_{j:j\neq i} (r_{ji}p_i - r_{ij}p_j)^2 \\ &\text{subject to } \sum_{i=1}^{k} p_i = 1, p_i \geq 0, \forall i \end{aligned} \tag{11}$$

For further discussion, please refer to [25] and [22].

5. Proposed Hybrid Emotion Recognition System

With the GMM-based and SVM-based emotion recognition subsystems, described in Sects. 3 and 4, the proposed hybrid emotion recognition system finalizes the recognition results by combining these two subsystems.

Suppose $p_{i_{gmm}}$ and $p_{i_{svm}}$ are the posteriori probabilities for the i-th emotion class obtained by the GMM-based recognition subsystem in Sect. 3 and by the SVM-based recognition subsystem 4. The recognition score for the i-th emotion class by the proposed hybrid recognition system can be finalized by the linear combination of $p_{i_{gmm}}$ and $p_{i_{svm}}$, given by

$$p_{i_{hybrid}} = \alpha * p_{i_{gmm}} + (1 - \alpha) * p_{i_{svm}}, \tag{12}$$

where α is a weight constant between 0 and 1, and denotes a parameter controlling the contributions of both the GMM-based subsystem and the SVM-based subsystem to the hybrid recognition system. If α equals 0, then the classification result is entirely determined by the SVM likelihoods. On the contrary, if α equals 1, then it is entirely determined by the GMM likelihoods. This parameter was experimentally set to 0.8 in the following experiments, which means that the weight of the GMM-based posterior probability (α) is much larger than that of SVM-based posterior probability. It is reasonable as the recognition performance of GMM-based method is better than SVM-based one.

6. Experiments and Results

To evaluate the superiority of the proposed NUFCC features over the traditional MFCC features, the emotional recognition system with NUFCC features was compared with the system with MFCC features. In addition, the proposed emotion recognition system using both non-uniform spectral features and prosodic features was examined and further

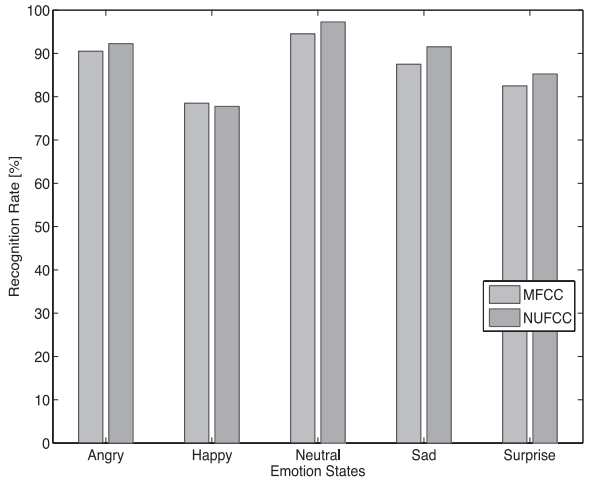


Fig. 7 Recognition results of the emotion recognition system with different spectral features, namely NUFCC and MFCC features. The GMM classifier is used in both recognition systems.

compared with the systems using only NUFCC features, and a system with only prosodic features.

In the following two experiments, the emotional speech corpus CASIA provided by Chinese-LDC was used. For training, 200 utterances from each emotion of each person were randomly selected, and the other remaining utterances were used for testing.

6.1 Evaluation of the NUFCC Features

In this experiment, we focus on investigation of the superiority of the newly proposed NUFCC features relative to the traditional MFCC features. In the experiment, both static and dynamic features were extracted and employed in emotion recognition. Specifically, 39 dimensional NUFCC (or MFCC) features, including 12 NUFCC (or MFCC) and their first and second order derivatives, along with the normalized power feature, were extracted from emotional speech as features of the emotion recognition system. The GMM with 512 mixtures were trained using the extracted NUFCC features or MFCC features for discriminating emotions in these two systems.

The recognition results for each emotion obtained by the two systems with different features (i.e., NUFCC or MFCC) are shown in Fig. 7. Figure 7 indicates that the system with NUFCC features provides higher emotion recognition rates for all emotional states except for happy, compared with the system with MFCC features. The relative error reduction rate averaged across all emotions amounts to 16.4%. The improved emotion recognition performance should be attributed to the fact that different contributions of different frequency bands to emotion discrimination in speech have been considered in the design of NUFCC features.

Table 1 Recognition results of three emotion recognition systems, the GMM-based system with NUFCC features, the SVM-based system with prosodic features, and the proposed hybrid system with both NUFCC features and prosodic features.

	GMM-NUFCC	SVM-Prosodic	Proposed Hybrid
Angry	92.25	71.25	90.25
Happy	77.75	59.75	82.25
Neutral	97.25	80.75	98.00
Sad	91.50	91.00	94.50
Surprise	85.25	73.50	88.50
Average	**88.80**	**75.30**	**90.70**

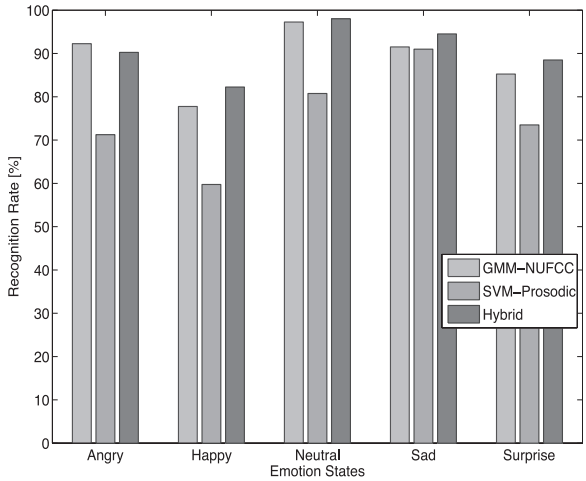


Fig. 8 Plots of the results of three emotion recognition systems, the GMM-based system with NUFCC features, the SVM-based system with prosodic features, and the proposed hybrid system with both NUFCC features and prosodic features.

6.2 Evaluation of the Proposed Hybrid Emotion Recognition System

In this section, experiments were conducted to evaluate the performance of the proposed hybrid emotion recognition system, which was further compared with that of the GMM-based recognition system with NUFCC features and the SVM-based recognition system with prosodic features.

In implementation, the GMM-based recognition system with NUFCC features was the same as that used in Experiment 1. The SVM-based recognition system involved the 87 dimensional prosodic features detailed in Sect. 4, which was trained and classified by SVM approach with radial basis function as kernel. The proposed hybrid system combined both subsystems in a linear way with the parameter $\alpha = 0.8$ in Eq. (12).

The recognition results of three emotion recognition systems, namely the GMM-based system with NUFCC features, the SVM-based system with prosodic features and the proposed hybrid system with both NUFCC and prosodic features, are listed in Table 1 and further plotted in Fig. 8. These experimental results show that the GMM-based system with NUFCC features results in much higher recognition accuracies for all emotion states in comparison with

the SVM-based system with prosodic features. The relative error reduction rate averaged across all emotions gets to 54.7%. Furthermore, the proposed hybrid emotion recognition system, by combining these two subsystems, results in more improved recognition rate (except for "angry"), which leads to the average relative recognition error reduction rates of about 62.3% and 17.0% compared to the SVM-based system and GMM-based system, respectively.

7. Conclusion

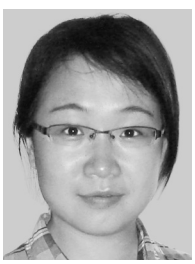
This paper introduced a hybrid speech emotion recognition system based on both spectral and prosodic features. For spectral features, we employed a new feature extraction method based on a novel non-uniform sub-band processing technique. For prosodic features, a set of prosodic features that are highly correlated with speech emotional states was chosen. The proposed hybrid emotion system combines a GMM-based subsystem that exploits the non-uniform spectral features and a SVM-based subsystem that exploits the prosodic significance. Experimental results show that the proposed hybrid emotion recognition system outperforms the systems using only spectral features or prosodic features.

Acknowledgements

This work is partially supported by The National Science & Technology Pillar Program (2008BAI50B03), National Natural Science Foundation of China (No.10925419, 90920302, 10874203, 60875014), and the China-Japan (NSFC-JSPS) Bilateral Joint Projects and the SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

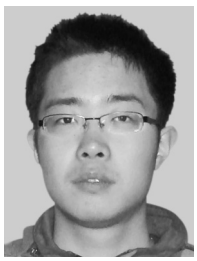
References

[1] R. Fernandez, A computational model for the automatic recognition of affect in speech, Ph.D. thesis, Massachusetts Institute of Technology, Supervisor-Picard, Rosalind W., 2004.
[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Commun., vol.48, no.9, pp.1162–1181, 2006.
[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., vol.18, no.1, pp.32–80, Jan. 2001.
[4] I. Campbell, A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A speech synthesis system with emotion for assisting communication," Proc. ISCA Workshop on Speech and Emotion, pp.167–172, Belfast, 2000.
[5] P. Ekman, "An argument for basic emotions," Cognition & Emotion, vol.6, no.3, pp.169–200, 1992.
[6] H. Schlossberg, "Three dimensions of emotion," Psychological Review, vol.61, no.2, pp.81–88, 1954.
[7] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.1, pp.39–58, Jan. 2009.
[8] T. Nwe, S. Foo, and L.D. Silva, "Speech emotion recognition using hidden markov models," Speech Commun., vol.41, no.4, pp.603–623, 2003.
[9] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Trans. Speech Audio Process., vol.9, no.3, pp.201–216, March 2001.
[10] Y. Wang and L. Guan, "An investigation of speech-based human emotion recognition," Proc. IEEE 6th Workshop on Multimedia Signal Processing, pp.15–18, 2004.
[11] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," Proc. ICSLP, pp.889–892, Oct. 2004.
[12] S. Lee, E. Bresch, and S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," Proc. 7th Int. Seminar Speech Production, 2006.
[13] http://hyperphysics.phy-astr.gsu.edu/hbase/Music/vowel.html
[14] Y. Zhou, Y. Sun, J. Li, J. Zhang, and Y. Yan, "Physiologically-inspired feature extraction for emotion recognition," Proc. Interspeech, pp.1975–1978, Brighton, U.K., Sept. 2009.
[15] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," Signal Process., vol.16, no.3, pp.233–246, 1989.
[16] "Mandarin emotional speech corpus." http://www.chineseldc.org/doc/CLDC-SPC-2005-010/intro.htm, 2005. Institute of Automation, Chinese Academy of Sciences.
[17] C. de Boor, A. Practical Guide to Splines, revised ed., Springer, Nov. 2001.
[18] L. Devillers, I. Vasilescu, and L. Vidrascu, "F0 and pause features analysis for anger and fear detection in real-life spoken dialogs," Speech Prosody 2004, Nara, Japan, 2004.
[19] G. McLachlan and K. Basford, Mixture Models: Inference and Applications to Clustering, Marcel Dekker, New York, 1988.
[20] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed., Wiley-Interscience, 2000.
[21] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," J. Royal Statistical Society, Series B, vol.39, no.1, pp.1–38, 1977.
[22] C.C. Chang and C.J. Lin, LIBSVM: A library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
[23] C.W. Hsu and C.J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw., vol.13, no.2, pp.415–425, 2002.
[24] T.F. Wu, C.J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," J. Machine Learning Research, vol.5, pp.975–1005, Aug. 2004.
[25] H.T. Lin, C.J. Lin, and R.C. Weng, "A note on platt's probabilistic outputs for support vector machines," Machine Learning., vol.68, no.3, pp.267–276, 2007. http://www.csie.ntu.edu.tw/˜cjlin/papers/plattprob.pdf
[26] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, pp.61–74, MIT Press, 1999.



Yu Zhou received the Bachelor's degree from the School of Electronic Information in Wuhan University in June, 2006. Currently she is a Ph.D. candidate of ThinkIT Speech Laboratory, IOA, CAS. Her research is focused on speech signal processing and speech emotion recognition.

Junfeng Li received the B.E. degree from Zhengzhou University and the M.S. degree from Xidian University, China, both in Computer Science, in 2000 and 2003, respectively. He received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology in March, 2006. From April 2006 to March 2007, he was a post-doctoral fellow at Research Institute of Electrical Communication (RIEC), Tohoku University. Since April 2007, he has been an Assistant Professor in the Graduate School of Information Science, JAIST. His research interests include speech signal processing and intelligent hearing aids. Dr. Li received the Best Student Award in Engineering Acoustics First Prize from the Acoustic Society of America in 2006, and the Best Paper Award from JCA2007 in 2007.



Yanqing Sun received the Bachelor's degree in Information Engineering from Nanjing University of Aeronautics and Astronautics (NUAA) in June, 2005. Currently he is a Ph.D. candidate of ThinkIT Speech Laboratory, IOA, CAS. His research interests include robust speech recognition and confidence measures.

Jianping Zhang graduated from the department of electronic engineering of Tsinghua University in July 1992. He got his Ph.D. in Electronic Engineering from Tsinghua University, May 1999. And he was a postdoctoral research fellow in Center for Spoken Language Research at the University of Colorado at Boulder From Nov 1999 to Dec 2001, and in the academy of Oregon From Jan 2002 to July 2002. Then he joined ThinkIT laboratory. He has a long record of research in speech, especially in spoken signal processing and understanding. Presently his research emphasizes language model, multi-language TTS.

Yonghong Yan received his B.E. from Tsinghua University in 1990, and his Ph.D. from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of the Center for Spoken Language Understanding. He worked in Intel from 1998–2001, chaired the Human Computer Interface Research Council, and worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Currently he is a professor and director of ThinkIT Lab. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. He has published more than 100 papers and holds 40 patents.

Masato Akagi received his B.E. degree in Electronic Engineering from Nagoya Institute of Technology in 1979, and his M.E. and Dr. Eng. degrees in Computer Science from Tokyo Institute of Technology in 1981 and 1984, respectively. In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT). From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with the School of Information Science, Japan Advanced Institute of Science and Technology, where he is currently a professor. His research interests include speech perception mechanisms of humans and speech signal processing.