Yongming HUANG[1], **Guobao ZHANG**[1], **Fei DONG**[1], **Yue LI**[1], **Feipeng DA**[1]

Southeast University (1)

Speech Emotion Recognition Using Hybrid Generative and Discriminative Models

***Abstract**. In this paper, we use Sequential Forward Selection to select 8 dimensional frame-level features from the total 69 dimensional features, and we reduce the dimensions of utterance-level eigenvectors from 63 to 12 by fisher discriminant. Then, two kinds of GMM multidimensional likelihoods are proposed for hybrid generative and discriminative models. Experimental results on Berlin emotional speech databases show that the GMM-MAP/SVM series hybrid model is the optimal Hybrid Generative and Discriminative Models, with the recognition rate up to 85.1%.*

***Streszczenie.** W artykule zaprezentowano system wykrywania emocji w głosie na podstawie modelu dyskryminacyjnego. Zaprezentowano badania skuteczności system na przykładzie bazy danych Berlin. (**System wykrywania emocji w głosie na podstawie modelu dyskryminacyjnego**)*

Keywords: Speech emotion recognition, Generative models, Discriminative models, Hybrid models.
Słowa kluczowe: wykrywanie emocji w głosie, model dyskryminacyjny.

Introduction

To convey information through speech is one of human's important abilities. Emotion information in speech has crucial influence on people's communication state. With the rapid development of pattern recognition and affective computing, speech emotion recognition using computer is attracting more and more attention from the researchers. However, performance of machine in speech emotion recognition is still far behind human beings, and there is still a long way for speech emotion recognition to enter human's everyday lives. There are mainly two issues for speech emotion recognition, the first is how to find features that describe emotions efficiently, and the second is how to build a proper model for emotion classification. In this study we mainly focus on the second one.

So far the approaches of speech emotion recognition can be divided into two classes: (1) approaches based on generative model, such as GMMs and HMMs. Frame-level features are usually used in generative model; (2) discriminative model based approaches， such as K-nearest neighbours (KNN), multi-layer perceptrons, artificial neural networks (ANN) and support vector machine (SVM), of which the key point is to seek for the optimal classification boundaries between classes and to reflect differences between different distributions. Utterance-level features are always used in discriminative model. As for the fusion of discriminative models, an approach to the fusion of GMM classifiers was introduced in [1], two GMM models based on different parameter sets were fused using dot product, with the mean recognition rate up to 74.25%. As for the fusion of generative models, classifiers such as SVM and KNN were fused by means of unweighted vote and stacked generalization in [2], and a mean recognition rate up to 72.18% was achieved. In [1, 2], since the fusion was limited within the same model, which are essentially the same, the improvement of recognition rate was rather limited.

As the fusion strategy of literature [3] is merely limited to simple fusion strategies just like voting, generative model's advantage of modeling data's internal distribution and discriminative model's ability to separate data of different classes effectively were not well combined. In this paper, we introduce SVM likehoods and GMM/HMM likehoods to build Parallel hybrid model, and then, fusion strategy is used to combine the outputs of the two models. Besides, since the GMM trained for each utterance is a representation of utterance characteristics, redundant information such as speaker identity and linguistic information are included in the emotion recognition processing, thus affecting the system performance. Unlike literature [4,5], to build the more effective GMM-SVM models, in this paper we propose GMM-MAP/ SVM series hybrid model.

1 Hybrid generative and discriminative models

1.1 Parallel hybrid model

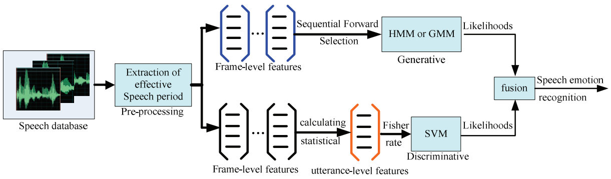

Fig.1. Parallel hybrid model

For a speech fragment, two kinds of features are extracted. Classifiers are built using both of the models for analysis, so that a more comprehensive result can be obtained and the generalization ability of the speech emotion recognition system would be improved. Based on this idea, a parallel "generative/discriminative" hybrid model approach is proposed in this paper. Only a generative model (GMM or HMM) and a discriminative model (SVM) are used. The generative approach, GMM or HMM, models the frame-level features, while the discriminative approach SVM models the utterance-level features. Then the probability outputs of SVM and generative model GMM or HMM are fused by a certain fusion strategy. Note that unlike GMM, the output of SVM is either +1 or -1, but not probability or likelihood. To unify the outputs to the same scale, the output of SVM is converted into probability[10] in this work, and then the experiments are conducted.

As is shown in Fig.1, since the output of discriminative SVM model is the posterior probability score, utterance-level features are extracted from the testing utterance *X* and then processed by the SVM model. The output is formed by the posterior probability scores that the utterance belongs to each emotion class, denoted by $S_{dm}(X,1)$, $S_{dm}(X,2),\ldots,S_{dm}(X,N)$ (number *1* to *N* represents emotion classes, and *dm* is short for discriminative model). The output of GMM or HMM for the testing utterance *X* is also in probability form, denoted by $S_{gm}(X,1)$, $S_{gm}(X,2),\ldots,S_{gm}(X,N)$ (*gm* is short for generative model here). Finally, parallel fusion strategy is used to combine the outputs of the two models.

For a testing utterance *X*, the posterior probability score for emotion class *i* using generative model (GMM, HMM) is

$S_{gm}(X,i)$, *(i=1,2,…,N)*, and the posterior probability score for discriminative model SVM is $S_{dm}(X,i)$, then testing utterance *X*'s overall posterior probability score *S(X,i)* for emotion class *i* is given by:

$$(1) \qquad S(X,i) = S_{gm}(X,i) \times S_{dm}(X,i) \quad (1 \le i \le N)$$

The emotion recognition result i^* for testing utterance *X* is determined by the maximum overall posterior probability score:

$$(2) \qquad i^* = \arg\max_{1 \le i \le N} S(X,i)$$

1.2 Series hybrid model

In ordinary speech emotion recognition systems based on GMM and frame-level features, a GMM is trained for each emotion class. Obviously, we cannot do that here since during the process it is still unknown which emotion the emotional utterance belongs to, so the corresponding emotion model should not be used. In this model, a GMM-BM (Background Model, BM) model is trained firstly, and then MAP (Maximum A Posteriori, MAP) self-adaption algorithm is applied to each of the emotional utterances to build the GMM models (also known as GMM-Adapt models) for each utterance. Unlike GMM-UBM/SVM model, the GMM-BM model here is trained using all the training utterances in peaceful state. Since any emotion state can be regarded as a variation from the peaceful state, GMM model trained from peaceful speech provides a priori knowledge for training GMM models of each emotional utterance. To be exact, the GMM-BM model should be called GMM-Neutral model (Neutral is peaceful). The GMM-MAP/SVM serial hybrid model is illustrated in Fig.2.

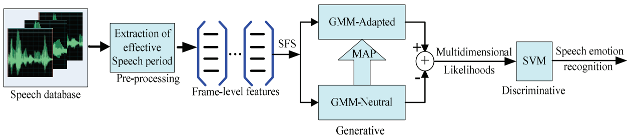

Fig.2. GMM-MAP/SVM series hybrid model

As is shown in Fig.2, after the self-adaptive GMM model is achieved, the frame-level features of a testing emotional utterance is extracted and used as input of the self-adaptive GMM model to obtain the multidimensional likelihood vector. However, the parameters passed on to SVM model are not the multidimensional probability vector, but the likelihood ratio computed by applying the testing emotional utterance to adaptive GMM and GMM-Neural model respectively. The ratio also can be represented as difference of the two log-likelihoods obtained. Therefore, the multidimensional probability of the d^{th} dimension for the frame-level features of the testing emotional utterance is given by:

$$(3) \qquad P_d = \log p(O_d \mid \lambda_{adapted}) - \log p(O_d \mid \lambda_{neutral})$$

Similarly, the likelihood of the i^{th} Gaussian component in Gaussian Mixture Model is given by:

$$(4) \qquad P_i = \log p(O_i \mid \lambda_{adapted}) - \log p(O_i \mid \lambda_{neutral})$$

Then the multidimensional likelihood vectors of the emotional utterances are trained and recognized using the SVM classifier. SVM's good discriminability is used for emotion classification to improve the recognition rate. The MAP self-adaptive algorithm is used to train the GMM serial hybrid model.

2 Feature extraction

2.1 Frame-level features extraction

In this paper, MFCC[6] as well as its first and second order differences are extracted to form the 39-dimensional features, which make up a feature group of frame-level features for speech emotion recognition with a generative model. Besides the MFCC features, frame-level features that have little relationship with speech content, including fundamental frequency (F0), short-time energy (E), the first three formant frequencies (F1, F2, F3) and five Mel-frequency subband energies (MFSE1 to MFSE5), as well as their first and second order differences, sum to 30 features, are extracted in this paper. In other words, a series of 30-dimensional feature vectors is extracted for each emotional utterance. Additionally, the two feature groups are mixed together to build the third 69-dimensional feature group. Because of the redundant information existing in the frame-level features extracted, in this paper Sequential Forward Selection (SFS) algorithm is employed to select the best subset from the frame-level features. The feature set achieved is regarded to be optimal. For the 69-dimensional hybrid features, after feature selection, the 8-dimensional short-term feature selected is:

$$(5) \qquad TF_3 = \left[\frac{d^2F0}{dt^2}, \frac{dF2}{dt}, \frac{d^2MFSE2}{dt^2}, \frac{dMFCC1}{dt}, \frac{d^2E}{dt^2}, \frac{d^2MFSE4}{dt^2}, \frac{dF0}{dt}, \frac{d^2F1}{dt^2} \right]$$

2.2 Utterance-level features extraction

Utterance-level features are the statistical features obtained on the basis of frame-level prosodic features and voice quality features. The maximum, minimum, mean, standard deviation, range and their first and second order differences are typical statistics, besides, the median, the first quartile and third quartile, etc. are also taken into account by some researchers. In this paper, 63 features [8] are extracted as the utterance-level features based on the EMO-DB. There is also information redundancy in utterance-level features. For frame-level features, SFS algorithm is adopted for feature selecting in this paper. But for the utterance-level features extracted in this paper, considering the computational cost, Fisher criterion is used for dimension reduction instead. Fisher criterion is a classic linear discriminate method, which is widely used in pattern recognition [7]. After evaluation of the Fisher rate for the 63-dimensional utterance-level features extracted, 12 optimal utterance-level features are selected on the EMO-DB for 6 emotion classes: happy, anger, sad, fear, bored and peace. The features are listed in Table 1.

Table 1 Selected 12 best utterance-level features

Order of importance	utterance-level features	Fisher rate
1	Mean of fundamental frequency	1.43
2	Std of energy	1.26
3	Mean of first difference of fundamental frequency	1.11
4	Range of Log energy	1.05
5	Mean of first formant frequency	1.02
6	Mean of Log energy	0.89
7	Min of fundamental frequency	0.79
8	Mean of second formant frequency	0.73
9	first quartile of fundamental frequency	0.58
10	Max of third formant frequency	0.54
11	Max of second difference of energy	0.54
12	Median of energy	0.53

3 GMM Multidimensional likelihoods

The probability output of a *T*-frame testing utterance using GMM emotion model λ_n is represented by the following log-likelihood:

$$\ln P(O \mid \lambda_n) = \sum_{t=1}^{T} \ln P(o_t \mid \lambda_n) \tag{6}$$

where *1≤n≤N* are emotion labels. Using diagonal covariance matrix for GMM models, $P(o_t|\lambda_n)$ can be expanded as following:

$$\begin{aligned} P(o_t|\lambda_n) &= \sum_{i=1}^{M} P(o_t, i|\lambda_n) \\ &= \sum_{i=1}^{M} c_i \cdot P(o_t|i, \lambda_n) \\ &= \sum_{i=1}^{M} c_i \cdot \frac{1}{(2\pi)^{D/2} \prod_{d=1}^{D} \sigma_{id}} \cdot \exp\left[-\frac{1}{2} \sum_{d=1}^{D} \frac{(o_{td} - \mu_{id})^2}{\sigma_{id}^2} \right] \\ &= \sum_{i=1}^{M} c_i \cdot \prod_{d=1}^{D} \left\{ \frac{1}{\sqrt{2\pi}\sigma_{id}} \cdot \exp\left[-\frac{1}{2} \frac{(o_{td} - \mu_{id})^2}{\sigma_{id}^2} \right] \right\} \end{aligned} \tag{7}$$

where *M* is the number of Gaussian components in GMM model, also called hidden state number. *D* is dimension of the feature vector, and σ_{id} is variance of the d^{th} dimension under the i^{th} hidden state (assumed diagonal). The mean in the d^{th} dimension under the i^{th} hidden state is denoted by μ_{id} . Note that in Eq.(33), the output probability for each mixture component is equal to the product of feature components in all the *D* dimensions. For a testing utterance with *T* feature vectors and for all of the GMM mixture components, only the likelihood in the d^{th} dimension is computed while the rest are ignored (the output probability is 1), then according to Eq.(33), the output probability P_d of GMM in the d^{th} dimension is given by:

$$\begin{aligned} P_d &= \ln P(O_d|\lambda) \\ &= \sum_{t=1}^{T} \ln \left\{ \sum_{i=1}^{M} c_i \cdot \frac{1}{\sqrt{2\pi}\sigma_{id}} \cdot \exp\left[-\frac{1}{2} \frac{(o_{td} - \mu_{id})^2}{\sigma_{id}^2} \right] \right\} \end{aligned} \tag{8}$$

where d=1,2,…,D.

After the processing above, for a *T*-frame testing emotional utterance with *D*-dimensional feature vectors, *D* output probabilities for each feature dimension are obtained under the GMM model *λ*. And a *D*-dimensional feature vector $[P_1,P_2,\ldots,P_D]^T$ is formed with these *D* output probabilities. The reason for this processing is that: for the original *D*-dimensional feature vectors, each dimension has a different contribution to the same emotion, and also has different influence on different emotions. After the processing of GMM model *λ*, the *D*-dimensional feature vectors of a *T*-frame testing emotional utterance are converted into *D*-dimensional output probabilities, which reflects likelihoods of the *D* feature components under the GMM model.

On the other hand, the probability density function of a *M*-order GMM model is the weighted sum of *M* Gaussian probability density functions. Similarly, for *T* feature vectors of a testing utterance, suppose only output probability of the i^{th} Gaussian component is computed while the rest are ignored, then according to Eq.(33), output probability of the i^{th} component in GMM model is given by:

$$\begin{aligned} P_i &= \ln P(O_i|\lambda) \\ &= \sum_{t=1}^{T} \ln \left\{ c_i \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_{id}} \exp\left[-\frac{(o_{td} - \mu_{id})^2}{2\sigma_{id}^2} \right] \right\} \end{aligned} \tag{9}$$

where *i=1,2,…,M*. In this way, for a testing emotional utterance, a *M*-dimensional feature vector $[P_1,P_2,\ldots,P_M]^T$ is formed by outputs of the *M* Gaussian components in the GMM model, and is then processed by SVM.

4 Experiments and Results

In this paper, EMO-DB[9] is used for the experiments. Ergodic HMM topology structure is used and each HMM emotion model has 5 states. Probability of the observations under each state has continuous distribution, and the distribution function is a linear combination of 4 Gaussian distributions. The number of GMM components is $M=8$. “Radial Basis Function” is used as the kernel function of SVM, and “one-versus-one” method is adopted as the multi-class classification strategy. A total of $C_5^2 = 15$ SVM classifiers are trained, and 4-fold cross validation is performed here.

4.1 Experiments based on parallel hybrid models

In the speech emotion recognition experiments based on generative model and SVM probability output, the selected frame-level features and utterance-level features are extracted respectively for the training emotional speech samples at first. After that, frame-level features are used to train two generative models, the HMM model and the GMM model, both including 6 emotion models. When recognizing the testing emotional speech samples, parallel hybrid models HMM/SVM and GMM/SVM are used respectively. For each hybrid model, the maximum overall posteriori criterion, denoted by Fusion-MAPmax, is used. The recognition results are shown in Table 2.

Table 2 Speech emotion recognition rate based on parallel hybrid model

Hybrid model	Recognition rate (%)						
	Angry	Happy	Sad	Bore	Fear	Neutral	Average
HMM	83.4	81.1	83.2	64.8	69.1	79.6	76.9
GMM	73.4	74.1	78.2	68.5	74.7	78.3	74.5
HMM/SVM	87.2	73.9	86.6	67.6	75.1	76.3	77.8
GMM/SVM	76.8	80.2	78.0	66.1	69.5	84.2	75.8

According to the experiment result in Table 2, the recognition rates of HMM model and GMM model using TF3 feature set are 76.9% and 74.5%, respectively, and we can find that the parallel hybrid models have completely outperformed them.

4.2 Experiments based on parallel series hybrid models

Two types of multi-dimensional output vector are introduced in this paper. One is the *D*-dimensional GMM output vector that is decomposed according to the dimensions of the input feature vectors, here *D* is the dimension of the model’s input, and the features used are the ones selected from frame-level features, in the experiment *D*=8. The other is the *M*-dimensional GMM output vector decomposed according to different Gaussian mixture components. *M* is the number of gaussian mixture components in GMM and in the experiment *M*=8. The results of serial hybrid model are listed in Table 3 and 4.

Table 3 Speech emotion recognition rate based on D dimension GMM-MAP/SVM hybrid model

validation	Recognition rate (%)						
	Angry	Angry	Angry	Angry	Angry	Angry	Angry
First time	90.3	76.5	93.3	75.0	70.6	94.7	83.4
Second time	87.5	77.8	86.7	80.0	76.5	90.0	83.1
Third time	93.8	83.3	81.3	75.0	64.7	90.0	81.4
Fourth time	84.4	83.3	93.8	71.4	77.8	95.0	84.3
Average	89.0	80.2	88.8	75.4	72.4	92.4	83.1

Table 4 Speech emotion recognition rate based on M dimension GMM-MAP/SVM hybrid model

validation	Recognition rate (%)						
	Angry	Angry	Angry	Angry	Angry	Angry	Angry
First time	77.4	82.4	86.7	80.0	76.5	89.5	82.1
Second time	78.1	77.8	93.3	75.0	76.5	85.0	81.0
Third time	84.4	83.3	93.3	75.0	70.6	90.0	82.8
Fourth time	84.4	77.8	93.3	71.4	72.2	85.0	80.7
Average	81.1	80.3	91.7	75.4	74.0	87.4	81.7

According to data in Table 3 and 4, recognition rates up to 83.1% and 81.7% are obtained by the two types of GMM-MAP/SVM serial models, respectively. Compared with the recognition rates of 74.5% and 79.1% achieved by GMM and SVM alone, the serial hybrid models have better performances. The result shows that the fusion of GMM multi-dimensional probability output and SVM can effectively improve the accuracy rate of speech emotion recognition, since the hybrid models combine the modeling ability of GMM and the discriminability of SVM effectively. No matter for the GMM-UBM/SVM model or for GMM-MAP/SVM, *D*-dimensional models perform better than *M*-dimensional models, that is, the probability outputs decomposed in feature dimensions have better discriminability than that in the GMM component dimensions. At last, according to data in Table 3 and 4, the highest recognition rate of 83.1% is achieved by the *D*-dimensional GMM-MAP/SVM serial hybrid model.

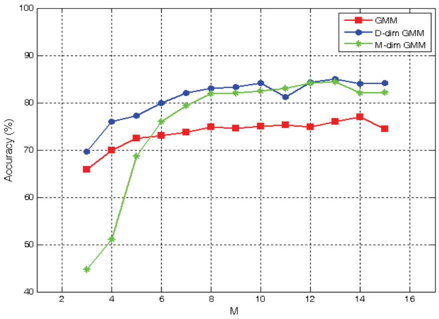

Fig.3. Recognition rate of GMM models under different mixtures M

In addition, the number of Gaussian components is also a crucial factor that influences the emotion recognition result. In this paper, another experiment is designed to improve the performance of GMM-MAP/SVM model for speech emotion recognition. For the same set of frame-level features, the GMM alone, GMM-MAP/SVM with *D*-dimensional output and GMM-MAP/SVM with *M*-dimensional output are used for speech emotion recognition respectively. Change of the recognition rate along with the number of Gaussian mixture components *M* is observed and shown in Fig.3. When the value of *M* (in the range of 3 to 15 here) changes, the GMM-MAP/SVM model with *D*-dimensional output always has the best recognition rate, and a highest recognition rate of 85.1% is achieved with *M*=13.

5 Conclusions and Discussions

In this paper, the advantages and disadvantages of both generative models and discriminative models are analyzed along with their complementarily. And the "generative/discriminative" hybrid model is introduced and applied in speech emotion recognition. Based on two approaches for building the "generative/discriminative" hybrid model, two effective model fusion methods are introduced in this paper for speech emotion recognition. The first is parallel fusion method. The SVM is improved with probability output, thus the problem of classification uncertainty is solved. Then two fusion strategies are used to fuse SVM with the probability output generative models HMM and GMM to build the parallel hybrid model and improve the recognition rate. The second one is serial fusion method, of which the basic idea is to model the frame-level features with generative model GMM, and the multi-dimensional output probability of GMM is then used as input of the subsequent SVM model, aiming to improve the speech emotion recognition system's performance by combining the modeling ability of GMM with the discriminability of SVM. The serial hybrid models can be divided into the *M*-dimensional model and the *D*-dimensional model according to different patterns of GMM probability output. When *M* varies within the range of 3 to 15, the *D*-dimensional GMM-MAP/SVM serial hybrid model with *M*=13 is the optimal hybrid generative and discriminative model, with the highest recognition rate of 85.1%.

REFERENCES

[1] Vondra M, Vich R. Evaluation of Speech Emotion Classification Based on GMM and Data Fusion[C]. In: Proc. of the Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, 2009. 98-105.
[2] Morrison D, Wang R L, De Silva L C. Ensemble methods for spoken emotion recognition in call-centres[J]. Speech Communication, 49(2007), No.2, 98-112.
[3] Kim S, Georgiou P G, Lee S, et al. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. 2007 Ieee Ninth Workshop on Multimedia Signal Processing, 2007, 48-51.
[4] Schwenker, F., et al., The GMM-SVM Supervector Approach for the Recognition of the Emotional Status from Speech. Artificial Neural Networks - Icann 2009, Pt I, 2009, 894-903.
[5] Chang Huai, Y., L. Kong Aik, and L. Haizhou, An SVM Kernel With GMM-Supervector Based on the Bhattacharyya Distance for Speaker Recognition. Signal Processing Letters, IEEE, 16(2009), No1, 49-52.
[6] Yun S, Yoo C D. Speech Emotion Recognition Via a Max-Margin Framework Incorporating a Loss Function Based on the Watson and Tellegen's Emotion Model[C]. In: 2009 Ieee International Conference on Acoustics, Speech, and Signal Processing, 2009. 4169-4172.
[7] Guo Y F, Shu T T. Feature Extraction Method Based on the Generalised Fisher Discriminant Criterion and Facial Recognition. Pattern Analysis & Appl. 4(2001), No.1, 61–66.
[8] Platt J C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods[M]. ADVANCES IN LARGE MARGIN CLASSIFIERS. Cambridge: MIT Press, 1999, 61-74.
[9] Burkhardt F, Paeschek A, Rolfes M, et al. A Database of German Emotional Speech. In: Proc. INTERSPEECH 2005.
[10] Wu T-F, Lin C-J, Weng R C. Probability Estimates for Multi-class Classification by Pairwise Coupling [J]. The Journal of Machine Learning Research, 5(2004), 975-1005.

***Authors**: Yongming Huang. He is PHD student in Southeast University, His research interests are in the areas of Speech emotion recognition and Speech recognition E-mail:* huang_ym@163.com *; prof. Guobao Zhang. E-mail:* guobaozh@seu.edu.cn *; Master Fei Dong, E-mail:* zdhdf2008@163.com; Yue Li, *She is master student in Southeast University, her research area is automatic emotion recognition. E-mail:* kccyq@hotmail.com *. prof. Feipeng Da, E-mail:* dafp@seu.edu.cn