GMM SUPERVECTOR BASED SVM WITH SPECTRAL FEATURES FOR SPEECH EMOTION RECOGNITION

Hao Hu, Ming-Xing Xu, and Wei Wu

Center for Speech Technology, Tsinghua National Lab for Information Science and Technology,
Tsinghua University, Beijing, 100084, China
{huhao, wuwei}@cst.cs.tsinghua.edu.cn, xumx@tsinghua.edu.cn

ABSTRACT

Speech emotion recognition is a challenging yet important speech technology. In this paper, the GMM supervector based SVM is applied to this field with spectral features. A GMM is trained for each emotional utterance, and the corresponding GMM supervector is used as the input feature for SVM. Experimental results on an emotional speech database demonstrate that the GMM supervector based SVM outperforms standard GMM on speech emotion recognition.

Index Terms **—** Speech emotion recognition, SVM, GMM supervector, spectral features

1. INTRODUCTION

Speech emotion recognition is a challenging yet important speech technology. It can be applied to broad areas, such as human-computer interaction [1], call center environment [2], and enhancement of speech and speaker recognition performance.

Global prosodic and voice quality features (utterance level statistics) have been widely used in speech emotion recognition and demonstrated considerable recognition success [3][4]. Besides these prosodic and voice quality features, spectral features are another effective group of features for describing emotional states [5], such as linear predictive cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC). Gaussian mixture model (GMM) and hidden Markov model (HMM) with MFCC have achieved promising results on speech emotion recognition [6][7]. However, there is a problem when using GMM to model emotional states. Effective training of GMM demands a large amount of data, while emotional speech data is expensive to collect and therefore the available training data is usually limited.

Support vector machine (SVM) has proved to achieve better performance for recognition tasks than many other generative classifiers, while requiring a small amount of training data. However, spectral features can not be used directly for SVM learning, because the spectral features extracted from utterances of various lengths do not have a fixed dimension. An idea proposed in speaker recognition could solve this problem [8]. In this paper, we apply this idea to introduce a GMM supervector based SVM with spectral features for speech emotion recognition. For each emotional utterance, the MFCC features extracted from it are used to train a GMM, and then the GMM supervector is constructed as the input feature for SVM. The GMM KL divergence kernel [8] and some other commonly used SVM kernels are investigated in the GMM supervector based SVM. Experimental results demonstrate that the proposed approach outperforms the standard GMM when using spectral features for speech emotion recognition.

This paper is organized as follows. In Section 2, a description of the database used in the experiments is given. In Section 3, the proposed GMM supervector based SVM is described. In Section 4, experiments and analysis of the results are presented. In Section 5, conclusions are drawn and future work is suggested.

2. DATABASE DESCRIPTION

The emotional speech database contains five acted emotions, including four primary emotions [9] (anger, fear, happiness and sadness) and a neutral speaking style. 40 short sentences without any emotional tendency were selected as speech materials. To avoid exaggerated emotion expression, non-professional speakers were hired for the recordings. 8 native Chinese speakers (4 females and 4 males) uttered each sentence in five simulated emotional states, resulting in 1,600 utterances in total. These utterances were recorded in silent environment with 16-bit precision at a sampling rate of 16 kHz. Each of them contains around 4 seconds of valid speech.

To eliminate utterances whose emotional expression is ambiguous, a subjective assessment of this emotional speech database was carried out. Five listeners independently identified the emotion category of each utterance by their subjective perception. Finally, 1,309

This work is funded by National Natural Science Foundation of China under grant 60433030.

utterances with over four listeners' agreement on their emotion categories were selected. These selected utterances used for experiments are summarized in Table 1.

Table 1. Number of utterances in each emotion category

	Female	Male	**Total**
Anger	148	145	**293**
Fear	127	137	**264**
Happiness	124	91	**215**
Neutral	151	139	**290**
Sadness	137	110	**247**
Total	**687**	**622**	**1309**

There is another neutral speech database used as development data, which includes short paragraphs read by 16 speakers (gender balanced) in a neutral speaking style. These 16 speakers have no overlap with those in the above emotional speech database. This neutral speech database contains 30 minutes of valid speech.

3. GMM SUPERVECTOR BASED SVM FOR SPEECH EMOTION RECOGNITION

3.1. GMM Supervector

The density function of a GMM is defined as

$$p(x) = \sum_{i=1}^{N} w_i N(x;\ \mu_i, \Sigma_i) \tag{1}$$

where $N(;\ ,)$ is the Gaussian density function, w_i, μ_i and Σ_i are the weight, mean and covariance matrix of the i-th Gaussian component, respectively. The supervector of a GMM is formed by concatenating the mean of each Gaussian component, and it takes the form as

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix} \tag{2}$$

For each emotional utterance, a GMM is trained with the extracted spectral features, and the corresponding supervector is obtained. Instead of training the GMM via EM algorithm [10], we adapt the GMM from a universal background model (UBM) [11], which is widely used in speaker recognition. In our method, the UBM is a GMM trained via EM algorithm using neutral speech from a large number of speakers. The adaptation of each emotional utterance's GMM is performed with maximum *a posteriori* (MAP) algorithm [12], and only the means are adapted. This process is illustrated in Fig. 1.

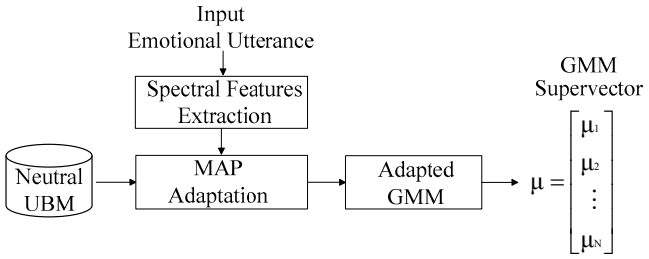

Fig. 1. Construction of the GMM supervector from an emotional utterance

There are two reasons for us to train the GMM for each utterance by adapting from a UBM. Firstly, effective training of a GMM via EM algorithm needs a large amount of data, while the speech in a single utterance is usually short, and thus it can not provide sufficient training data for the GMM. Since the UBM is trained with speech from a large number of speakers, it could represent the speaker-independent distribution of speech features. Hence, it can provide *a priori* knowledge for the training of each utterance's GMM, and thus compensates the lack of training data for the GMM. Secondly, since all the GMMs are adapted from the UBM by adapting means only, i.e., each of the corresponding Gaussian components has the same weight and covariance matrix, therefore the derived GMMs' supervectors are comparable in the supervector space. The reason we use the neutral speech to train the UBM is that the neutral speech is relatively easier to collect.

The GMM supervector can be considered as a mapping from the spectral features of an emotional utterance to a high-dimensional feature vector. This mapping allows the production of features with a fixed dimension for all the emotional utterances. Therefore, we can use the GMM supervectors as input for SVM learning.

3.2. SVM Kernel Selection

SVM performs a non-linear mapping from an input space to a high-dimensional space through a kernel, which is an important component for SVM learning.

In this work, we investigated four SVM kernels in the proposed GMM supervector based SVM, including:

- linear kernel
- polynomial kernel
- radial basis function (RBF) kernel
- GMM KL divergence kernel [8]

The first three kernels, which are commonly used, take the form as equation (3), (4) and (5), respectively.

$$K(x_i, x_j) = x_i^t x_j \tag{3}$$

$$K(x_i, x_j) = (x_i^t x_j + 1)^n \tag{4}$$

$$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma} \right)^2 \right] \quad (5)$$

where n is the order of polynomial, and σ is the width of the radial basis function.

The GMM KL divergence kernel is proposed in [8]. Given two GMM supervectors μ^a and μ^b, the GMM KL divergence kernel is defined as

$$\begin{aligned} K(\mu_a, \mu_b) &= \sum_{i=1}^{N} w_i \left(\mu_i^a\right)^t \Sigma_i^{-1} \mu_i^b \\ &= \sum_{i=1}^{N} \left(\sqrt{w_i}\, \Sigma_i^{-\frac{1}{2}} \mu_i^a \right)^t \left(\sqrt{w_i}\, \Sigma_i^{-\frac{1}{2}} \mu_i^b \right) \end{aligned} \quad (6)$$

where μ_i^a and μ_i^b are the means of the i-th mixture component in the two GMMs, w_i and Σ_i are the weight and covariance matrix (assumed diagonal) of the i-th Gaussian component in the UBM, respectively. Note that equation (6) is not the strict definition of KL divergence between two GMMs, but an approximation of it.

4. EXPERIMENTAL RESULTS

In this work, two experiments were performed on the emotional speech database. Firstly, four SVM kernels were investigated in the proposed GMM supervector based SVM. Secondly, the performance of the GMM supervector based SVM system and the standard GMM system were compared.

The spectral features were 13-dimensional MFCC plus energy, together with their delta and acceleration coefficients. The features were extracted every 10 ms with a frame length of 25 ms. The Hamming windowing was applied and the pre-emphasis factor was 0.97. The neutral UBM was trained from the neutral speech database described in Section 2, which consisted of 64 Gaussian components. In the following experiments, 5 fold cross-validation is performed for error estimation. More precisely, the emotional database described in Table 1 is equally divided into 5 disjoint subsets, and classifiers were trained five times, each time with a different subset held out as a testing set [13]. The estimated classification error is the mean of these five errors for the testing data. Both gender-dependent and gender-independent experiments were performed for each classifier, i.e., gender-dependent indicates female and male data are considered separately for training and testing.

4.1. GMM Supervector based SVM with Different Kernels

LibSVM v2.8 function library [14] was used for the training and testing of SVM which use the one-against-one strategy for multi-class classification.

Table 2. Accuracy of GMM supervector based SVM using different kernels, where Linear, 3rd order Polynomial, RBF and GMM KL divergence (GMM KL) are the kernels defined in equation (3), (4), (5) and (6), respectively.

Accuracy (%)	GMM KL	Linear	3rd order Polynomial	RBF ($\sigma = 0.1$)
Female	**93.6**	91.6	91.3	89.3
Male	**91.4**	87.0	86.7	86.3
Mixed-gender	**82.5**	75.8	76.3	75.7

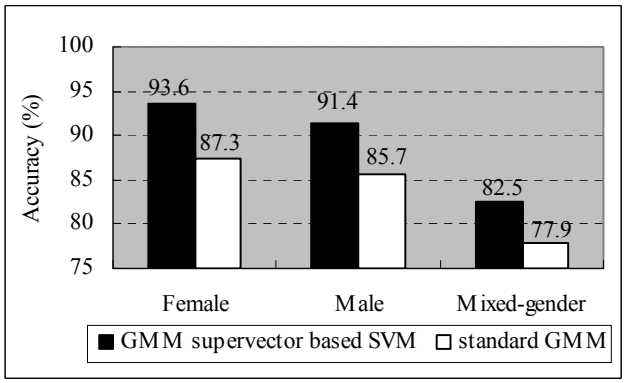

Fig. 2. Accuracy of GMM supervector based SVM using the GMM KL divergence kernel and standard GMM

Table 2 shows the accuracy of the GMM supervector based SVM using four different kernels. It can be seen that the GMM KL divergence kernel has the best performance in both gender-dependent and gender-independent experiments. The success of the GMM KL divergence kernel is attributed to its nature as an approximation to KL divergence between two GMMs [8]. Since the KL divergence is a measurement of the discrepancy between two distributions, thus using its approximation as the kernel fits well with the proposed approach comparing with other commonly used kernels, which considers GMM supervectors in context of SVM.

4.2. GMM Supervector based SVM v.s. GMM

With the results in Section 4.1, we used the GMM KL divergence kernel in our GMM supervector based SVM system to compare it with the standard GMM system. In the standard GMM system, each emotion was modeled by a GMM trained with the corresponding emotional utterances. The GMMs were trained via EM algorithm, each of which consisted of 64 Gaussian components. A maximum likelihood Bayes classifier is used for decision. The accuracy of these two systems is shown in Fig. 2.

From Fig. 2, it can be seen that the GMM supervector based SVM significantly outperforms the standard GMM for speech emotion recognition. More precisely, compared to the standard GMM, the accuracy of the GMM supervector based SVM is 6.3% higher for female subject,

5.7% higher for male subject and 4.6% higher for mixed-gender subject. The results also indicate that it is helpful to identify the speaker's gender in the utterance first, and then perform the emotion recognition on a gender-dependent system.

Since the gender-dependent emotion recognition system is preferred, we analyzed the confusion between different emotions in condition of separate-gender subject. Confusion matrices of the GMM supervector based SVM for female and male subjects are shown in Table 3 and Table 4, respectively. From the results, we can see that fear, happiness and anger are the most frequently confused emotions for both female and male subjects. This might be attributed to the similar arousal level [1] when speakers are in these three emotional states. It can be also found that sadness and fear are easily misclassified for male subject. The reason could be that the valence level of these two emotional states is close [1].

Table 3. Confusion matrix of GMM supervector based SVM for female subject (A: anger; F: fear; H: happiness; N: neutral; S: sadness)

	Classified Emotion (%)				
Intended Emotion	A	F	H	N	S
A	**97.9**	0.7	1.4	0.0	0.0
F	8.0	**78.4**	9.6	0.0	4.0
H	1.7	7.5	**90.8**	0.0	0.0
N	0.0	0.0	0.0	**100.0**	0.0
S	0.0	0.7	0.0	0.7	**98.6**

Table 4. Confusion matrix of GMM supervector based SVM for male subject (A: anger; F: fear; H: happiness; N: neutral; S: sadness)

	Classified Emotion (%)				
Intended Emotion	A	F	H	N	S
A	**96.5**	0.0	2.8	0.7	0.0
F	3.0	**91.1**	3.7	0.0	2.2
H	8.9	7.8	**80.0**	3.3	0.0
N	0.0	0.0	2.2	**96.3**	1.5
S	0.0	10.0	0.0	1.8	**88.2**

5. CONCLUSIONS

In this paper, we propose to apply the GMM supervector based SVM with spectral features to speech emotion recognition. The GMM KL divergence kernel was shown to yield better performance than other commonly used kernels in the proposed system. The results suggest that the gender information should be considered in speech emotion recognition, and demonstrate that the GMM supervector based SVM system significantly outperforms standard GMM system. For the frequently confused emotional states, other type of features, such as prosodic and voice quality features can be fused with our proposed method to enhance the emotion recognition performance in future work.

6. REFERENCES

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G., "Emotion Recognition in Human-Computer Interaction", *IEEE Signal Processing Magazine*, 18(1), pp. 32-80, Jan. 2001.

[2] Petrushin, V.A., "Emotion recognition in speech signal: experimental study, development, and application", in *Proc. of ICSLP 2000*, pp. 222-225, 2000.

[3] Ververidis, D., Kotropoulos, C., Pitas, I., "Automatic Emotional Speech Classification", in *Proc. of ICASSP 2004*, pp. 593-596, Montreal, Canada, 2004.

[4] Fernandez, R., Picard, R.W., "Classical and Novel Discriminant Features for Affect Recognition from Speech", in *Proc. of INTERSPEECH 2005*, pp. 1-4, Lisbon, Portugal, 2005.

[5] Banse, R., Scherer, K., "Acoustic profiles in vocal emotion expression", *J. Personality Social Psych.*, 70(3): 614-636, 1996.

[6] Lee, C.M., Yildirim, S., E., Bulut, M., Kazemzadeh, A., Busso, C., Deng Zh.g., Lee, S., Narayanan, S., "Emotion Recognition based on Phoneme Classes", in *Proc. of ICSLP 2004*, Korea, 2004.

[7] Luengo, I., Navasm E., Hernaez, I., Sanchez, J., "Automatic Emotion Recognition using Prosodic Parameters", in *Proc. of INTERSPEECH 2005*, pp. 493-496, Lisbon, Portugal, 2005.

[8] Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", in *Proc. of ICASSP 2006*, pp. 97-100, France, 2006.

[9] Cowie, R., Cornelius, R., "Describing the emotional states that are expressed in speech", *Speech Communication*, (40): 5-32, 2003.

[10] Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Series B, 39: 1-38, 1977.

[11] Reynolds, D.A., Quatieri, T.F., Dunn, R., "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 10(1-3): 19-41, 2000.

[12] Gauvain, J. L., Lee, C.H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. Speech Audio Process*, 2(2): 291-298, 1994.

[13] Duda, R., Hart, P., Stork, D., *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[14] Chang, Ch., Lin, Ch., *LIBSVM: a Library for Support Vector Machines, 2005*, http://www.csie.ntu.edu.tw/~cjlin/libsvm.