Fusion of Global Statistical and Segmental 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, xumx, wuwei}@cst.cs.tsinghua.edu.cn

Abstract

Speech emotion recognition is an interesting and challenging speech technology, which can be applied to broad areas. In this paper, we propose to fuse the global statistical and segmental spectral features at the decision level for speech emotion recognition. Each emotional utterance is individually scored by two recognition systems, the global statistics-based and segmental spectrum-based systems, and a weighted linear combination is applied to fuse their scores for final decision. Experimental results on an emotional speech database demonstrate that the global statistical and segmental spectral features are complementary, and the proposed fusion approach further improves the performance of the emotion recognition system.

Index Terms: speech emotion recognition, global statistical features, segmental spectral features, decision fusion

1. Introduction

With the growing need to give computers skills of emotional intelligence, automatically recognizing emotion from speech has received increasing attention recently. Speech emotion recognition is an interesting and challenging speech technology, which can be applied to broad areas, such as human-computer interaction [1][2], call center environment [3], and enhancement of speech and speaker recognition performance.

Global statistical features based on prosody and voice quality have been widely used in speech emotion recognition, and demonstrated considerable recognition success [4][5]. Besides these global statistical features, segmental spectral features are another effective group of features for describing emotional states [6], such as Mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC). Since the global and segmental features have played significant roles in speech emotion recognition, it is necessary to explore an effective way to complementarily fuse both features to further improve the performance of the emotion recognition system.

In this paper, we constructed two emotion recognition systems, the global statistics-based and segmental spectrum-based systems, and designed a method to fuse them at the decision level. The first system is based on an optimal set of global statistical features obtained from feature selection, and utilizes support vector machine (SVM) as the classifier. The second system is based on segmental spectral features using GMM supervector based SVM [7] as the classifier. A weighted linear combination is adopted to fuse the outputs of the two recognition systems for final decision. Experimental results demonstrate that the fusion system outperforms either of the two individual systems on speech emotion recognition.

The remainder of this paper is organized as follows. The global statistics-based and segmental spectrum-based systems are introduced in Section 2 and 3, respectively. The proposed fusion approach is described in Section 4. In Section 5, experiments and analysis of the results are presented. In Section 6, conclusions are drawn and future work is suggested.

2. Global statistics-based system

The global statistics-based system is briefly introduced in this section. We first focus on the extraction of global statistical features and then present a feature selection method for optimizing the initial feature set. Finally, we explain the applied classifier for emotion recognition.

2.1. Global statistical features

The global statistical features have been broadly used in speech emotion recognition. In this system, raw parameters of pitch, intensity (energy), first four formants, duration, harmonics-to-noise ratio (HNR), jitter, and shimmer are calculated with Praat [8]. The first derivative of the pitch and intensity contours is also estimated by smoothing with a moving window. Afterwards the global statistical features are derived via different statistical measurements, such as mean, median, standard deviation, minimum, maximum, and range. Since the characteristics and dynamics of these statistical features are different, the normalization (also called z-score) is subsequently performed on each statistical feature to make them on similar scale. The initial feature set includes a total of 85 global statistical features, which can not be described in detail here. Table 1 shows the distribution of these global statistical features.

Table 1. *Distribution of the global statistical features*

Feature type	Number
Pitch	21
Intensity	19
Formant	36
Duration	2
HNR	3
Jitter	2
Shimmer	2

2.2. Feature selection

It is expected that many of the initial global statistical features are redundant. Hence, we search for an optimal feature set by applying a SVM based sequential forward floating selection (SFFS) [9], which has been demonstrated an efficient technique to solve feature selection problems.

SVM-SFFS selects features via forward and conditional backward steps in a floating manner controlled by a criterion. The employed criterion is the emotion recognition accuracy of SVM with the selected features. The selection starts from an initial empty set, adds the feature which optimizes the criterion at forward step, and then possibly eliminates the least significant one at backward step. It stops until the required number of features is reached. The optimal feature set is determined according to the highest observed accuracy throughout the whole feature selection process.

2.3. Classification techniques

Apart from selection of the optimal set of features, choosing an effective classifier is also important for designing the emotion recognition system. In pervious studies, SVM has been shown to achieve better performance than many other classifiers on speech emotion recognition [10]. Therefore, we choose SVM with a linear kernel as the classifier in the global statistics-based system.

3. Segmental spectrum-based system

Different from the global statistical features, the segmental spectral features contain the time varying information in local level spectrum. Spectral features have been mostly used with hidden Markov model (HMM) or Gaussian mixture model (GMM) in emotion recognition systems [11][12]. However, effective training of GMM demands a large amount of data, while the available emotional speech data is usually limited. GMM supervector based SVM has proved to outperform the traditional GMM with the limited training data [7]. Therefore, we apply this approach to implement the segmental spectrum-based system.

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 segmental spectral features, and the corresponding supervector is obtained. Instead of training the GMM via EM algorithm, we adapt the GMM from a universal background model (UBM) [13], which is widely applied to speaker recognition. In our system, 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 [14], and only the means are adapted. This process is illustrated in Figure 1.

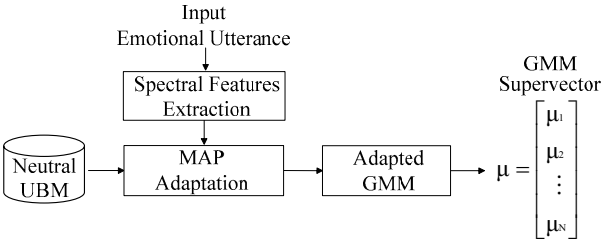

Figure 1: *Construction of the GMM supervector from an emotional utterance*

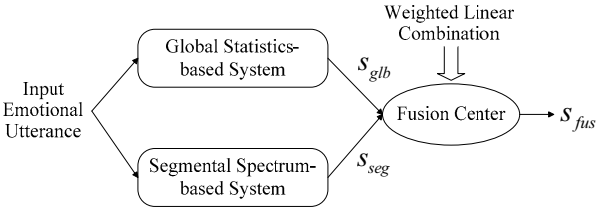

Figure 2: *Fusion of global statistics-based and segmental spectrum-based systems*

The process of constructing a 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

GMM KL divergence kernel has been shown to outperform other commonly used kernels in the GMM supervector based SVM [7]. Given two GMM supervectors μ^a and μ^b, the GMM KL divergence kernel is defined as

$$\begin{aligned} K(\mu^a, \mu^b) &= \sum_{i=1}^{N} w_i \left(\mu_i^a\right)^t \Sigma_i^{-1} \mu_i^b \\ &= \sum_{i=1}^{N} \left(\sqrt{w_i}\,\Sigma_i^{-\frac{1}{2}} \mu_i^a\right)^t \left(\sqrt{w_i}\,\Sigma_i^{-\frac{1}{2}} \mu_i^b\right) \end{aligned} \tag{3}$$

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 (3) is not the strict definition of KL divergence between two GMMs, but an approximation of it.

4. Decision fusion

The global statistics-based and segmental spectrum-based systems are fused at the decision level. Each emotional utterance is individually scored by either of the two systems. The output score of each individual system is a vector consisted of the estimated posterior probabilities of each emotional state $P(E_k|x)$, where x denotes the input emotional utterance. Afterwards, the two output scores are fused with a weighted linear combination to produce a fusion score for final decision. This process is shown in Figure 2.

Assuming S_{glb} and S_{seg} are the output scores of the global statistics-based and segmental spectrum-based systems, respectively. The fusion score S_{fus} is calculated by

$$S_{fus} = w \cdot S_{glb} + (1-w) \cdot S_{seg} \quad (4)$$

where w is a weight varying in the interval [0, 1]. Since the performances of the two recognition systems are different, it is not appropriate to use the equal weights for the linear combination. Therefore we further apply the logistic regression model [15] to estimate the weight w through a development set, which is described in Section 5.3. After obtaining the fusion score, the final decision is made by the maximum posterior probability of the fusion score as the following equation.

$$E_{rec} = \arg\max_{k} P(E_k \mid x) \quad (5)$$

5. Experiments

5.1. Experimental setup

Experiments in this work were performed on the emotional speech database introduced in [7]. This emotional speech database contains five acted emotions, including anger, fear, happiness, sadness and a neutral speaking style. 40 short sentences, the content of which does not have 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. To eliminate the utterances with ambiguous emotional expression, a subjective assessment was carried out by other five listeners. Finally, 1,309 utterances with over four listeners' agreement on their emotion categories were selected, which consisted of 687 and 622 utterances for female and male subjects respectively. Each of them contains around 4 seconds of valid speech. There is another neutral speech database used as development data, which contains 30 minutes of valid speech.

In the following experiments, 5-fold cross validation was employed for error estimation. The emotional speech database was equally divided into 5 disjoint subsets, and classifiers were trained five times, each time with a different subset held out as a testing set. The estimated classification error is the mean of these five errors for the testing data. Both gender-dependent and gender-independent experiments were carried out for each system, i.e., gender-dependent indicates female and male data are considered separately for training and testing.

5.2. Selection of global statistical features

In the global statistics-based system, SVM-SFFS was applied to eliminate the redundant features in the initial set of 85 global statistical features. LibSVM v2.8 function library [16] was utilized for training and testing of SVM, which used the one-against-one strategy for multi-class classification.

The selection results of global statistical features for both gender-dependent and gender-independent subjects are shown in Figure 3. It can be seen that the accuracy of three curves (female, male, and mixed-gender) increases until the size of feature set approaches 30, which indicates that the three accuracy curves reach a saturation level. After the saturation, adding more features to the feature set does not result in significant performance gain. The sizes of the optimal set of global statistical features are 29, 28, and 33 for female, male, and mixed-gender subjects, respectively.

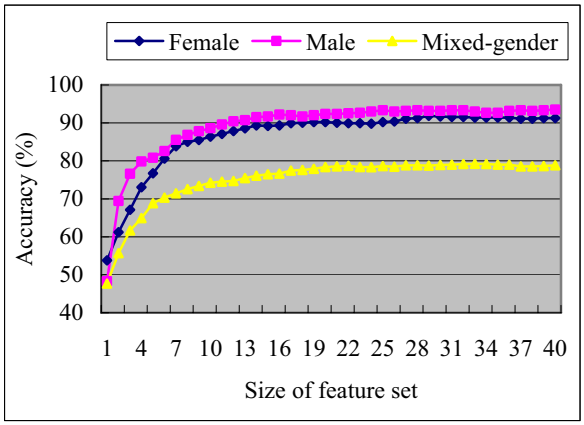

Figure 3: *Selection of global statistical features by SVM-SFFS for gender-dependent and gender-independent subjects*

Table 2. *Accuracy of global statistics-based, segmental spectrum-based, and fusion systems*

Accuracy (%)	Global statistics-based system	Segmental spectrum-based system	Fusion
Female	91.9	92.7	**95.3**
Male	93.5	91.9	**95.0**
Mixed-gender	79.2	82.5	**86.5**

From Figure 3, we can see that the accuracy of the recognition system approaches 90% and 80% for gender-dependent and gender-independent subjects, respectively. It confirms that the global statistical features are effective for speech emotion recognition.

5.3. Fusion of global statistics-based and segmental spectrum-based systems

To evaluate the proposed fusion approach, we first implemented the global statistics-based and segmental spectrum-based systems. In the global statistics-based system, SVM with a linear kernel was applied to recognize the emotional states, which used the optimal set of global statistical features described in Section 5.2. In the segmental spectrum-based system, the segmental spectral features were 13-dimensional MFCC plus energy, together with their delta and acceleration coefficients. The MFCC features were extracted using 25 ms frame length every 10 ms with Hamming window, and the pre-emphasis factor is 0.97. The neutral UBM was trained from the neutral speech database described in Section 5.1, which consisted of 64 Gaussian components.

In the decision fusion, the estimation of weight w in equation (4) requires a supervised development set. In the 5-fold cross validation, the development set was the same as the training data described in Section 5.1. The accuracy of the global statistics-based, segmental spectrum-based, and fusion systems is shown in Table 2.

From Table 2, it can be seen that the segmental spectrum-based system achieves higher accuracy than the global statistics-based system for female and mixed-gender subjects, but lower accuracy for male subject. These results indicate

that it is possible to fuse both systems to further improve the performance of the emotion recognition system. As shown in Table 2, the fusion system outperforms either of the two systems for both gender-dependent and gender-independent subjects. More precisely, when comparing with the better one of the two individual systems, the fusion system achieves 35.6%, 23.1%, and 22.9% error reduction for female, male, and mixed-gender subjects, respectively. The performance improvement can be attributed to the complementarity of the global and segmental features, for the global statistical features contain long term information about prosody and voice quality over the entire utterance, while the segmental spectral features describe time varying spectral characteristics at the local level.

With the promising results achieved by the proposed fusion approach, we further analyzed the performance gain of different emotional states for gender-independent subject, which is shown in Figure 4. From the results, we can see that the fusion system outperforms both global statistics-based and segmental spectrum-based systems for all the emotional states except anger. For anger, the accuracy of the fusion system is a little lower than the segmental spectrum-based system. The reason for this is that the decision made by the segmental spectrum-based system was dominated over that from the global statistics-based system, which got lower accuracy in identification of anger. This problem might be solved by applying other techniques to decision fusion.

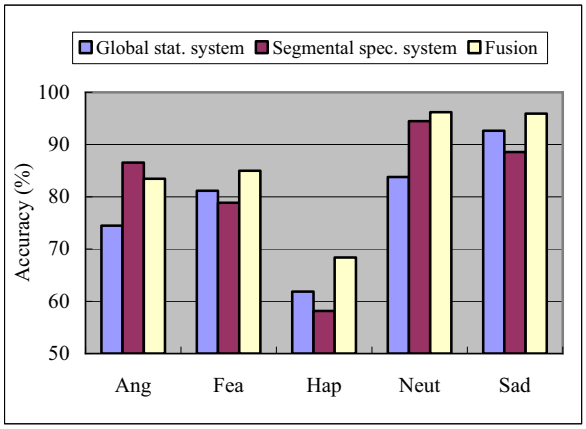

Figure 4: *Performance gain of different emotional states for gender-independent subject*

6. Conclusions

In this paper, we propose to fuse the global statistical and segmental spectral features at the decision level for speech emotion recognition. Each emotional utterance is individually scored by two recognition systems, the global statistics-based and segmental spectrum-based systems, and a weighted linear combination is applied to fuse their scores for final decision. Experimental results demonstrate that the global statistical and segmental spectral features are complementary, and the fusion system outperforms either of the two individual systems on speech emotion recognition. To further improve the performance of the fusion system, some other fusion techniques, such as SVM or multi-layer perceptron (MLP) should be investigated in future work.

7. Acknowledgements

This work has been partially funded by National Natural Science Foundation of China under grant No. 60433030.

8. 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): 32-80, Jan. 2001.

[2] Pantic, M., Rothkrantz, L.J.M., "Toward an Affect-Sensitive Multimodal Human-Computer Interaction", *Proceedings of the IEEE*, 91(9): 1370-1390, Sep. 2003.

[3] Petrushin, V.A., "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application", in *Proc. of ICSLP 2000*, pp. 222-225, 2000.

[4] Hozjan, V., Kacic, Z., "Improved Emotion Recognition with Large Set of Statistical Features", in *Proc. of EUROSPEECH 2003*, pp. 133-136, 2003.

[5] 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.

[6] Banse, R., Scherer, K., "Acoustic Profiles in Vocal Emotion Expression", *J. Personality Social Psych.*, 70(3): 614-636, 1996.

[7] Hu, H., Xu, M.-X., Wu, W., "GMM Supervector based SVM with Spectral Features for Speech Emotion Recognition", in *Proc. of ICASSP 2007*, pp. 413-416, Honolulu, USA, 2007.

[8] Boersma, P., Weenink, D., *Praat: doing phonetics by computer (Version 4.4.20)*, http://www.praat.org.

[9] Pudil, P., Novovicova, J., Kittler, J., "Floating Search Methods in Feature Selection", *Pattern Recognition Letters*, 15(11): 1119-1125, Nov. 1994.

[10] Schuller, B., Rigoll, G., "Timing Levels in Segment-Based Speech Emotion Recognition", in *Proc. of ICSLP 2006*, pp. 1818-1821, Pittsburgh, USA, 2006.

[11] Lee, C.M., Yildirim, S., 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.

[12] 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.

[13] Reynolds, D.A., Quatieri, T.F., Dunn, R., "Speaker Verification using Adapted Gaussian Mixture Models", *Digital Signal Processing*, 10(1-3): 19-41, 2000.

[14] Gauvain, J.L., Lee, C.H., "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, 2(2): 291-298, 1994.

[15] Hosmer, D.W., Lemeshow, S., *Applied Logistic Regression*, John Wiley, New York, 1989.

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