Journal of Southeast University (English Edition) Vol. 29, No. 2, pp. 129 - 133 June 2013 ISSN 1003—7985

Novel feature fusion method for speech emotion recognition based on multiple kernel learning

Jin Yun[1,2] Song Peng[1] Zheng Wenming[3] Zhao Li[1]

([1]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2]School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, China)
([3]Research Center for Learning Science, Southeast University, Nanjing 210096, China)

Abstract: In order to improve the performance of speech emotion recognition, a novel feature fusion method is proposed. Based on the global features, the local information of different kinds of features is utilized. Both the global and the local features are combined together. Moreover, the multiple kernel learning method is adopted. The global features and each kind of local feature are respectively associated with a kernel, and all these kernels are added together with different weights to obtain a mixed kernel for nonlinear mapping. In the reproducing kernel Hilbert space, different kinds of emotional features can be easily classified. In the experiments, the popular Berlin dataset is used, and the optimal parameters of the global and the local kernels are determined by cross-validation. After computing using multiple kernel learning, the weights of all the kernels are obtained, which shows that the formant and intensity features play a key role in speech emotion recognition. The classification results show that the recognition rate is 78.74% by using the global kernel, and it is 81.10% by using the proposed method, which demonstrates the effectiveness of the proposed method.

Key words: speech emotion recognition; multiple kernel learning; feature fusion; support vector machine
doi: 10.3969/j.issn.1003 - 7985.2013.02.004

The task of detecting emotions in speech utterances has become an active field in human-computer interaction and communication[1]. A speech emotion recognition system mainly includes feature extraction and classification. And how to extract suitable features that efficiently characterize different emotions is an important issue.

Prosodic features and voice quality features have been widely used in speech emotion recognition and obtained good performance in emotion classification[2-3]. Of these, pitch, formant, energy, and speaking rate are widely observed to be most significant characteristics[4]. Additionally, spectral features are also effective features used to discriminate emotional states[5], such as linear predictive cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCC). In some research, low level feature modeling on a frame level is pursued. Furthermore, linguistic features are often added these days[6]. In the first audio/visual emotion challenge, the audio baseline feature set consists of many low-level descriptors with statistical functionals[7]. Different kinds of features provide complementary information. So most of the existing methods concatenate different kinds of features (the local features) on high dimensional feature vectors (the global features) for training and classification.

The traditional feature fusion method only utilizes the global information while the local information is missing. Different kinds of local features are of different dimensions and different space distributions, and they contain their own local information. If such information is missing, the recognition rate will be decreased. To solve this limitation, the global features and the local features are combined together to obtain comprehensive information. Additionally, in order to improve the recognition performance, multiple kernel learning (MKL)[8] is adopted in our method. The global feature is mapped through a global kernel and each kind of local features is mapped through a different local kernel. Then the global kernel and the local kernels are combined together with various weights for SVM classification.

1 Proposed Feature Fusion Method Based on MKL

In this section, the MKL method and a novel speech feature fusion method based on MKL are introduced.

1.1 Multiple kernel learning

The goal of MKL is to learn a kernel machine with multiple kernel functions[9]. Specifically, suppose that M base kernels $k_m (m = 1, 2, \ldots, M)$ are given, and the ensemble kernel function k is denoted by

$$k(x_i, x_j) = \sum_{m=1}^{M} \beta_m k_m(x_i, x_j) \qquad \beta_m \geqslant 0, \ \sum_{m=1}^{M} \beta_m = 1 \tag{1}$$

In this paper, the major task of MKL is to learn the co

Received 2013-03-06.
Biographies: Jin Yun (1979—), male, graduate; Zhao Li (corresponding author), doctor, professor, zhaoli@seu.edu.cn.
Foundation items: The National Natural Science Foundation of China (No. 61231002, 61273266), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).
Citation: Jin Yun, Song Peng, Zheng Wenming, et al. Novel feature fusion method for speech emotion recognition based on multiple kernel learning[J]. Journal of Southeast University (English Edition), 2013, 29(2): 129 - 133. [doi: 10.3969/j.issn.1003 - 7985.2013.02.004]

efficients β_m ($m = 1, 2, \ldots, M$), such that the classifier $f(x)$ denoted by

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b = \sum_{i=1}^{N} \alpha_i y_i \sum_{m=1}^{M} \beta_m k_m(x_i, x) + b \quad (2)$$

can well classify the data points $\{x_i, y_i \in \pm 1\}_{i=1}^{N}$.

1.2 Proposed feature fusion method

The flowchart of the proposed feature fusion method in speech emotion recognition is shown in Fig. 1. In this paper, four kinds of features are adopted as the emotional features and they are pitch-related features, formant-related features, intense-related features and MFCC-related features. Let $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}$ and $\boldsymbol{x}^{(4)}$ denote the feature vectors consisting of the aforementioned four speech features extracted from the same speech utterance. They contain complementary discriminative information for speech emotion classification. Hence, fusing the four feature vectors will lead to a more powerful discriminative feature vector. One of the most popular feature fusion approaches is to simply concatenate the feature vectors into a global feature vector $\boldsymbol{x}$, i.e.,

$$\boldsymbol{x} = [\boldsymbol{x}^{(1)^{\mathrm{T}}}, \boldsymbol{x}^{(2)^{\mathrm{T}}}, \boldsymbol{x}^{(3)^{\mathrm{T}}}, \boldsymbol{x}^{(4)^{\mathrm{T}}}]^{\mathrm{T}} \quad (3)$$

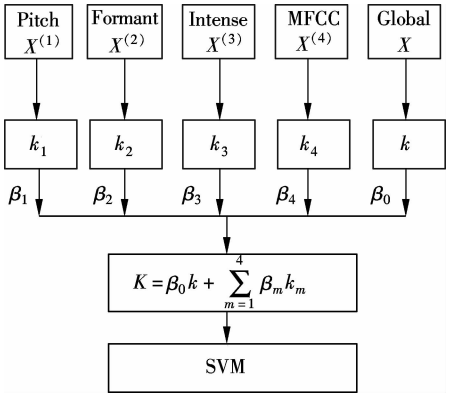

Fig. 1 Flowchart of proposed feature fusion method in speech emotion recognition based on MKL

The feature vector $\boldsymbol{x}$ includes the global speech information, while the four feature vectors $\boldsymbol{x}^{(i)}$ ($i = 1, 2, \ldots, 4$) contain the local speech information.

In order to make the features more classifiable, MKL is adopted for nonlinear mapping, where each kind of feature vector is associated with a kernel. For simplicity, the kernel corresponding to the global feature vector $\boldsymbol{x}$ is called the global kernel, whereas those corresponding to the local features are called local kernels. The global kernel K_0 is defined by

$$K_0 = k(x_i, x_j) \quad (4)$$

The m-th local kernel K_m is defined by

$$K_m = k_m(x_i^{(m)}, x_j^{(m)}) \qquad m = 1, 2, \ldots, 4 \quad (5)$$

The global kernel and the four local kernels are combined together with various weights. And then we obtain

$$K(x_i, x_j) = \beta_0 k(x_i, x_j) + \sum_{m=1}^{4} \beta_m k_m(x_i^{(m)}, x_j^{(m)}) = \sum_{m=0}^{4} \beta_m K_m$$
$$0 \leqslant \beta_m \leqslant 1;\ \sum_m \beta_m = 1;\ m = 0, 1, \ldots, 4 \quad (6)$$

where the weights β_m for each kernel indicate the importance of the associated features. Eq. (6) is the key of our proposed feature fusion method. Based on the idea of MKL, the feature fusion is made at the kernel level. $K(x_i, x_j)$ combine the global kernel with the local kernels according to their importance.

After obtaining the mixed kernel $K(x_i, x_j)$, we can predict the test sample using the SVM classifier. Given a test data $\boldsymbol{x}$, we first compute the global kernel function $k(x_i, x)$ and the local kernel functions $k_m(x_i^{(m)}, x^{(m)})$ ($m = 1, 2, 3, 4$). And then we obtain $K(x_i, x)$. The decision function for the test data is obtained as

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b \right) \quad (7)$$

2 Experiments

In this section, some experiments are conducted on the Berlin dataset[10] to evaluate the performance of our proposed method in speech emotion recognition. The proposed feature fusion method includes two steps. First, we should search for the optimal kernel parameters σ to compute the local kernels k_1, k_2, k_3, k_4 and the global kernel k. Secondly, we search for a set of optimal weights β_m ($m = 0, \ldots, 4$) to obtain the mixed kernel for SVM classification.

2.1 Dataset and emotion feature selection

The Berlin dataset is one of the most popular datasets used by researchers for emotion recognition. This dataset contains the emotional utterances recorded by 10 German actors reading one of 10 pre-selected sentences which are typical of everyday communication. The utterances of the dataset cover the following seven emotions: anger, boredom, fear, disgust, joy, sadness, and neutral emotion, in which the utterances corresponding to boredom and joy emotions are removed in the experiments.

In this paper, two types of speech features , the traditional prosodic features and the spectral features[11], are extracted for emotion recognition. The prosodic features consist of pitch, intensity and the first four formant frequency profiles as well as their derivatives, while the spectral features are comprised of the statistics of mel-frequency cepstral coefficients (MFCCs) and their derivatives.

To extract the emotional speech features, the Praat

software[12] is adopted to estimate the fundamental frequency (F0) features, the formant frequencies, the voice intensity profiles and MFCCs. Then, the statistical features over the entire utterance are computed. In total, the set of utterance-level prosodic features consists of the following 60 features:

- Mean, std, min, max, range of F0 and its derivative;
- Mean, std, min, max, range of F1, F2, F3, F4 and their derivatives;
- Mean, std, min, max, range of voice intensity and its derivative.

Utterance-level spectral features are statistics of the MFCC computed over the entire utterances. For each utterance, 13 MFCCs (including log-energy) are computed using a 25 ms Hamming window at intervals of 10 ms and their derivatives. Then the mean value, standard deviation, minimum, maximum and range over the entire utterances are computed. In this case, the total number of utterance-level spectral features is 130, i. e., mean, std, min, max, range of MFCC and its derivative.

10 pitch-related features, 40 formant-related features, 10 intense-related features and 130 MFCC-related features are concatenated into a 190-dimensional feature vector to represent an utterance. So in our method, one global kernel and four local kernels are adopted, which are corresponding to the global features and four kinds of local features, respectively.

2.2 Optimal parameter selection for kernels

The Gaussian kernel is the most commonly used kernel function. So in this paper, only the Gaussian kernel is adopted. The parameter σ determines the distribution of the kernel mapping. So we first search for the optimal σ for the global kernel and the local kernels.

$X = \{x_i,\ i = 1, 2, \ldots, n\}$ denote the training data and $X^{(m)} = \{x_i^{(m)}, m = 1, 2, 3, 4\}$ denote the training data of the m-th kind of features. A 10-fold cross-validation strategy is adopted in the experiment. Specifically, the training data is split into 10 subsets. One subset is used for testing and the other nine subsets are used for training. A number of parameters σ are tested separately. The σ of the global kernel is tested using X, and the σ_m of the m-th local kernel is tested using $X^{(m)}$. Then the σ corresponding to the highest average correct rate are the optimal parameters, which are listed in Tab. 1.

Tab. 1 Optimal parameters for global kernel and local kernels

Feature	Pitch	Formant	Intense	MFCC	Global
σ	2 000	20 000	100	200	20 000

2.3 Speech emotion recognition based on proposed feature fusion method

To evaluate the effectiveness of our proposed feature fusion method for speech emotion recognition, we should first search for the optimal $\beta_m (m = 0, \ldots, 4)$ for the global kernel and the local kernels by using the exhaustion method. And then the classification result using the global kernel is compared with that using our proposed method.

Specifically, the whole training dataset is equally partitioned into 10 subsets, and each time one subset is selected as testing data and the other nine subsets are used for training. $\beta_m (m = 0, \ldots, 4)$, subject to $\sum_{m=0}^{4} \beta_m = 1$, are given through a grid search using the range from 0 to 1 at a step size of 0.1. For each set of β_m, the experiments are repeated with ten-fold cross-validation, and then the average correct rate is obtained. The set of β_m corresponding to the highest average correct rate are the optimal kernel weights which are listed in Tab. 2. From the table, it is seen that the optimal weights β_m of pitch, formant, intense, MFCC and the global features are 0.1, 0.3, 0.3, 0.2 and 0.1, respectively, which means that the formant and the intense play the most important role in the speech emotion recognition. The MFCC features come second and the pitch features come last. The weight of the global features is only 0.1, so if only the global features are used, some important information on the local features will be missed.

Tab. 2 Optimal weights of global kernel and local kernels

Feature	Pitch	Formant	Intense	MFCC	Global
Weight	0.1	0.3	0.3	0.2	0.1

With the optimal weights β_m, the mixed kernel is computed by Eq. (6). And then the classification rate with the testing dataset is calculated. From Tab. 3, it is seen that the correct rate with the proposed method is 81.10%. For comparison, the classification rate by only using the global kernel with the SVM classifier is also calculated, and the best classification rate is 78.74%. The recognition rate is improved by 2.36% with the proposed method, which shows the effectiveness of the proposed method.

Tab. 3 Comparison results of only using global kernel and using proposed method

Method	The traditional method	The proposed method
Correct rate/%	78.74	81.10

The confusion matrices of only using the global kernel and of using the proposed method are shown in Fig. 2 and Fig. 3. From Fig. 2, it can be seen that the recognition rates of fear, disgust, neutral, sadness and anger are 65%, 47%, 77%, 86% and 95%, respectively. The recognition rate of disgust is very low. However, in Fig. 3, with the proposed method, the recognition rates are 83%, 87%, 65%, 95% and 81%, respectively. The recognition rate of disgust is greatly improved from 47% to 87%. Though the recognition rates of neutral and anger

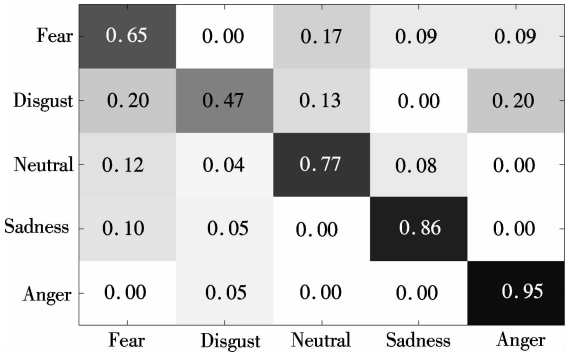

Fig. 2 Confusion matrix of using global features by SVM classifier

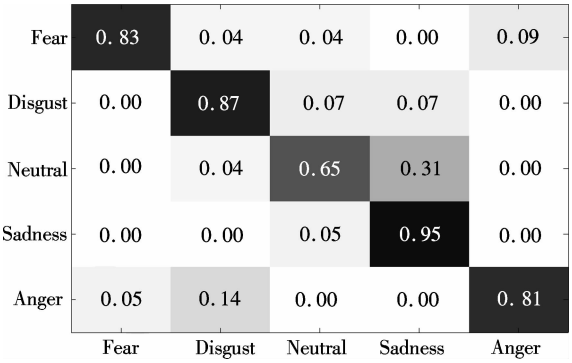

Fig. 3 Confusion matrix of using proposed feature fusion method by SVM classifier

are decreased from 77% to 65% and from 95% to 81%, respectively, the recognition rates of the other emotions are improved and the whole recognition rate is enhanced. So our modified multiple kernel feature fusion method can change the distribution of features in the high dimensional space, which makes it more classifiable.

2.4 Classification rate with only individual local kernel in speech emotion recognition

In this section, some experiments are conducted to show the classification performance with only individual local kernels. Each time, only one β_m is set to be 1 and the other β_m are set to be 0. The weights and the recognition rates are listed in Tab. 4. From the table, it is seen that the recognition rates are 33.07%, 17.32%, 44.88% and 70.87% with pitch-related features, formant-related features, intense-related features, and MFCC-related features, respectively. The recognition rate of our proposed method is better than any of those results obtained with only individual local kernels. As shown above, the proposed method is also better than the one using the global kernel method. So the novel feature fusion method is effective in speech emotion recognition.

Tab. 4 Classification rate with only local kernel in speech emotion recognition

Feature	β_1	β_2	β_3	β_4	Correct rate/%
Pitch	1	0	0	0	33.07
Formant	0	1	0	0	17.32
Intense	0	0	1	0	44.88
MFCC	0	0	0	1	70.87

3 Conclusion

In this paper, a novel feature fusion method for speech emotion recognition based on MKL is presented. For traditional feature fusion methods, different kinds of features are concatenated into a high-dimensional vector. Such a concatenation only utilizes the global information, missing the local information. So, the local features are added into the global features. Additionally, the multiple kernel learning method is adopted to improve the performance. One global kernel and four local kernels are combined together with optimal weights to form a mixed kernel, which contains more comprehensive information. The classification rate is 78.74% by using the global kernel method and it is 81.10% by using the proposed method, which demonstrates the effectiveness of the proposed method in speech emotion recognition.

References

[1] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction [J]. *IEEE Signal Process Magazine*, 2001, **18**(1): 32 – 80.

[2] Ververidis D, Kotropoulos C, Pitas I. Automatic emotional speech classification[C]//*Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. Montreal, Canada, 2004: 593 – 596.

[3] Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: audio, visual, and spontaneous expressions [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(1): 39 – 58.

[4] Lee C M, Yildirim S, Bulut M, et al. Emotion recognition based on phoneme classes [C]//*8th International Conference on Spoken Language Processing*. Jeju Island, Korea, 2004: 889 – 892.

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

[6] Schuller B, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge [C]//*Proceedings of the Annual Conference of the International Speech Communication Association.* Brighton, UK, 2009: 312 – 315.

[7] Schuller B, Valstar M, Eyben F, et al. AVEC 2011—the first international audio/visual emotion challenge [C]// *Lecture Notes in Computer Science*. Springer, 2011, **6975**: 415 – 424.

[8] Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm [C]//*Proceedings of the Twenty-First International Conference on Machine Learning*. Banff, Canada, 2004: 41 – 48.

[9] Lin Yen-Yu, Liu Tyng-Luh, Fuh Chiou-Shann. Multiple kernel learning for dimensionality reduction [J]. *IEEE Transactions on Pattern Analysis*

and Machine Learning, 2011, **33**(6): 1147 - 1160.
[10] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech [C]//*9th European Conference on Speech Communication and Technology*. Lisbon, Portugal, 2005: 1517 - 1520.
[11] Bitouk D, Verma R, Nenkova A. Class-level spectral features for emotion recognition [J]. *Speech Communication*, 2010, **52**(7/8): 613 - 625.
[12] Boersma P. Praat, a system for doing phonetics by computer [J]. *Glot International*, 2001, **5**(9/10): 341 - 345.

一种新的基于多核学习特征融合方法的语音情感识别方法

金 赟[1,2] 宋 鹏[1] 郑文明[3] 赵 力[1]

([1] 东南大学信息科学与工程学院, 南京 210096)
([2] 江苏师范大学物理与电子工程学院, 徐州 221116)
([3] 东南大学学习科学与研究中心, 南京 210096)

摘要:为了提高语音情感识别率,提出一种新的特征融合方法. 在全局特征的基础上,利用各种不同特征的局部信息,把全局特征和局部特征结合起来,引入多核学习的方法,使整体的全局特征和每类局部特征都对应一个核函数,加权求和得到一个组合核进行非线性映射,使不同类别的情感特征在高维再生核 Hilbert 空间中变得更容易分开. 采用 Berlin 语音情感数据库,利用交叉验证的方法确定相应的全局核和局部核的参数,经过多核学习计算,得到所有核的权重,确定共振峰和强度是情感识别中相对重要的特征. 实验表明,采用传统的方法识别率为 78.74%,而采用所提出的方法,识别率为 81.10%. 因此,所提出的特征融合方法能够有效地提高语音情感的识别率.

关键词:语音情感识别;多核学习;特征融合;支持向量机
中图分类号:TN912.3