LETTER

Speaker-Independent Speech Emotion Recognition Based on Two-Layer Multiple Kernel Learning

Yun JIN[†,††a)], **Peng SONG**[†], **Wenming ZHENG**[†††], **Li ZHAO**[†], *Nonmembers*, *and* **Minghai XIN**[†††], *Member*

SUMMARY In this paper, a two-layer Multiple Kernel Learning (MKL) scheme for speaker-independent speech emotion recognition is presented. In the first layer, MKL is used for feature selection. The training samples are separated into n groups according to some rules. All groups are used for feature selection to obtain n sparse feature subsets. The intersection and the union of all feature subsets are the result of our feature selection methods. In the second layer, MKL is used again for speech emotion classification with the selected features. In order to evaluate the effectiveness of our proposed two-layer MKL scheme, we compare it with state-of-the-art results. It is shown that our scheme results in large gain in performance. Furthermore, another experiment is carried out to compare our feature selection method with other popular ones. And the result proves the effectiveness of our feature selection method.

key words: *emotion speech recognition, multiple kernel learning, feature selection, speaker-independent*

1. Introduction

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 selection and classification. And how to select suitable features that efficiently characterize different emotions and how to choose an effective classifier are two important issues.

Feature selection is often utilized in speech emotion task to speed up the learning process and minimize "the curse of dimensionality" problem. In recent speech emotion recognition research works, the feature vector extracted from utterances becomes larger and larger (sometimes more than 6000 features from an utterance), so feature selection become more and more important. Popular feature selection methods have been used in speech emotion recognition including Principle Component Analysis (PCA) and Canonical Correlation Analysis (CCA). Furthermore, several search methods that evaluate a subset of features for the optimal subset have also been implemented. Best First Algorithm, Genetic Search Algorithms, Sequential Forward Search [2] and Sequential Forward Floating Search [3] belong to such search methods. However, these methods don't utilize the kernel method. If the feature space is nonlinear, the effect of the feature selection will not be satisfactory.

During the past decades, kernel methods such as support vector machine (SVM) have proved to be efficient tools for data representation, dimension reduction and classification. Using a single kernel, the data are mapped into a higher dimensional input space and an optimal separating hyperplane in this space is obtained. However, a single kernel cannot accurately depict the distribution of feature space. So in recent years, a more flexible learning model using multiple kernels instead of one, which is known as multiple kernel learning (MKL) [4], has been proposed. MKL has been shown to be more effective in many tasks, such as classification [5], regression [6] and feature selection [7]. In this paper, a two-layer MKL scheme is proposed. The first layer MKL is for feature selection and the second layer MKL is for classification.

The remainder of this paper is organized as follows. MKL method is briefly reviewed and our two-layer MKL scheme is then presented in Sect. 2. Section 3 introduces the Berlin dataset which is used in this paper and the features extracted from the dataset. In Sect. 4, the experiments are carried out to demonstrate the effectiveness of our proposed method in speech emotion recognition. Section 5 concludes the paper.

2. Two-Layer MKL Scheme

In this section, MKL is introduced and our proposed two-layer MKL scheme is presented.

2.1 Multiple Kernel Learning

Let $\{x_i, y_i\}_{i=1}^{l}$ be the learning set, where x_i belongs to some input space $\mathcal{X}$ and y_i is the corresponding label for pattern x_i. Multiple kernel learning can be formulated into the following optimization scheme [8]:

$$\min_{f\in\mathcal{H}_\gamma} \frac{1}{2}\|f\|^2_{\mathcal{H}_\gamma} + C\sum_{i=1}^{n} l(y_i f(x_i)) \tag{1}$$

where f is the decision function, and $\mathcal{H}_\gamma$ is a reproducing kernel Hilbert space associated with γ. It is denoted by the kernel function $k(\cdot,\cdot;\gamma) = \sum_{j=1}^{m} \gamma_j k_j(\cdot,\cdot)$. $l(\cdot)$ is the loss function. If the Hinge loss $l(t) = max(0, 1 - t)$ is used, the dual

Manuscript received May 22, 2013.

[†] The authors are with the Key Laboratory of Underwater Acoustic signal Processing of Ministry of Education, School of Information Science and Engineering, Southeast University, Nanjing 210096, P.R. China.

[††] The author is with the School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou, 221116, P.R. China.

[†††] The authors are with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, China.

a) E-mail: jinyun9999@gmail.com

DOI: 10.1587/transinf.E96.D.2286

problem of Eq. (1) is as following:

$$\min_{\gamma\in\Delta}\max_{\alpha\in Q} e^T\alpha - \frac{1}{2}(\alpha\circ y)^T\left(\sum_{j=1}^{m}\gamma_j K_j\right)(\alpha\circ y) \tag{2}$$

where α is a vector of Lagrange multipliers, and γ is a vector of the weights of multiple kernels. Q and Δ are the domains of α and γ respectively. e is a vector of all ones. $\{K_j\}_{j=1}^m$ is a set of base kernel matrices correlation with $\mathcal{H}_j'$, and $\circ$ is the elementwise product between two vectors. The domain Q is often denoted by $Q = \{\alpha \in \mathbb{R} : \alpha^T y = 0, 0 \le \alpha \le C\}$. If $\Delta = \{\gamma \in \mathbb{R}_+^m : \sum_{j=1}^m \gamma_j = 1, \gamma_j \ge 0\}$, it is called L_1-norm of kernel weights.

The above optimization problem can be regarded as a convex-concave problem, which alternate between the optimization of kernel weights and the optimization of the SVM classifier. So the coefficients α and the weighs γ are simultaneously learned in a single optimization problem. There are two advantages of MKL used in our paper. One is the sparsity of kernel weights using L_1-norm, and the other is the performance gain using MKL for classification.

2.2 The Proposed Two-Layer MKL Scheme

The advantage of the L_1-norm is that it leads to a sparse solution, which means most of the kernel weights are forced to be zero and only a few base kernels carry significant weights. Such property is utilized in our proposed method for feature selection. Specifically, an utterance is denoted by a n-dimensional vector $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ and each feature x_j is associated with a kernel k_j, then the combination kernel $\sum_{j=1}^n \gamma_j k_j$ is obtained. Using L_1-norm of MKL, the kernel weights γ_j $(j = 1, \cdots, n)$ are sparsified, and only a few important kernels are kept with weights. One feature is associated with one kernel in our method, so only those important features are retained. This is the fundamental of our feature selection method, which is shown in Fig. 1.

Another advantage is that using multiple kernels instead of a single one can enhance the interpretability of the decision function and improve performances [9].

We utilize such two properties of MKL and propose our two-layer MKL scheme for speech emotion recognition which is shown in Fig. 2. The first layer is using MKL for feature selection and the second layer is using MKL for classification. The former layer is the main part of our scheme, so it will be introduced in detail in this section.

L_1-norm constraint will bring sparse solution on γ_j. However, it may also discard complementary information if base kernels encode orthogonal information. Some useful features maybe removed during the process of feature selection. To improve the performance in this scenario, we propose our feature selection method based on MKL.

Let X denote the total training samples which are randomly separated into n parts. That can be depicted as $X = (X_1, \cdots, X_n)$. Each time, one part is left out and the other $(n-1)$ parts are combined into one group which is denoted by Y_i, $Y_i = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$, $i = (1, \cdots, n)$.

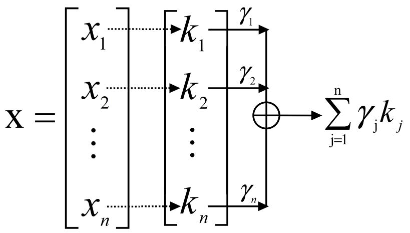

Fig. 1 The fundamental of feature selection based on MKL.

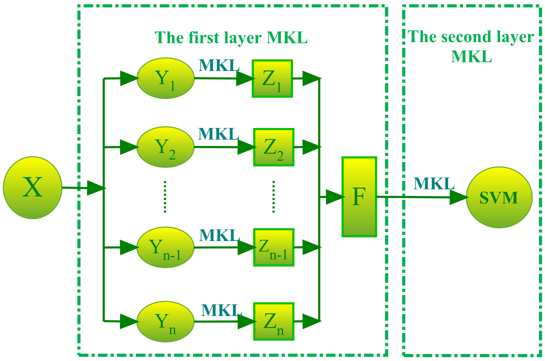

Fig. 2 The flowchart of our proposed two-layer MKL scheme for feature selection and classification.

For each group Y_i, feature selection is made using MKL to obtain a sparse feature set denoted by Z_i, $i = (1, \cdots, n)$. With intersection and union of Z_i, two feature subsets (F_{inter} and F_{union}) are obtained.

F_{inter} includes the features emerging in all n sparse subsets. So the features in F_{inter} are the most important features for classification. Though F_{inter} contains only a small set of features, it will lead to a relatively acceptable result. For one group, some important features may be abandoned in feature selection. If such process is repeated, some lost important features maybe appear in other group. The union of Z_i, F_{union}, is conducted to compensate the lost features and to include complementary information. F_{inter} and F_{union} are the results of our feature selection method.

3. Dataset and the Extracted Emotion Features

The Berlin Emotional Speech Dataset (EMO-DB) [10] is one of the most popular dataset used by researchers for speech emotion recognition, which covers the following seven speaker emotions: anger, boredom, fear, disgust, joy, sadness, neutral. The utterances were recorded by 10 German actors (5 male and 5 female) reading one of 10 preselected sentences typical of everyday communication. The whole dataset includes about 900 utterances, but only 494 utterances are marked after a listening experiment by 20 judgers. This selection set is used in the paper.

With the openEAR toolkit [11], 988 features are extracted as 19 functionals of 26 acoustic low-level descriptors (LLD) and corresponding first order delta. The 26 Low-level descriptors which are used in the paper are listed in Table 1. The statistical functionals and regression coefficients which

Table 1 26 Low-level descriptors (LLD).

Descriptor	Number
Intense	1
Loudness	1
MFCC 1-12	12
LSP 0-7	8
ZCR	1
Probability of voicing	1
F0	1
F0 Envelope	1
Total	26

Table 2 Statistical functionals, regression coefficients implemented in this paper.

Functionals	Number
Max./min, Range	3
Rel.position of Max./min	2
Arth.mean	1
Linear reg.coefficients and corresp.approx.err	4
Std.deviation, skewness, kurtosis	3
Quartiles and inter-quartile ranges	6
Total	19

are implemented are listed in Table 2.

4. Experiments

In this section, experiments will be conducted on the Berlin dataset to evaluate the performance of our proposed two-layer MKL scheme in speaker independent speech emotion recognition. The experiments are carried out with two stages. The first stage is to make feature selection using SimpleMKL [9]. The second stage is to make classification using SimpleMKL again with the selected features.

4.1 Results of the Proposed Method

In the first stage, feature selection is made based on SimpleMKL. Gaussian kernels are adopted with 10 different bandwidths σ on all features and with only 1 bandwidth σ on each single feature. The number of feature in the experiment is 988, therefore, there are totally 998 kernels (10 kernels for all features and 988 kernels for each single feature). Ten different bandwidths σ values are 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64. One σ value for each single feature is set 1 according to cross validation. There are two reasons for only using 1 bandwidth σ for each single feature. One reason is that too many kernels will bring tremendous time consuming. Another reason is that too many kernels will lead to overfitting which will deteriorate the recognition rate in the experiments.

The dataset is randomly split into 10 parts empirically. With our proposed feature selection method, two feature subsets F_{inter} and F_{union} are obtained. F_{inter} contains 12 features. The number of MFCC-related feature, LSP-related feature, loudness-related feature, zcr-related feature, voiceprob-related feature are 5, 2, 1, 1, 3 respectively. Because the above 12 features are included in all of the 10 fold groups, they contain the most important emotion information. The F_{union} contains 89 features. The number of MFCC-related feature, LSP-related feature, intensity-related feature, loudness-related feature, zcr-related feature, voiceprob-related feature and F0-related feature are 39, 31, 1, 2, 2, 9, 5 respectively. We can see that MFCC-related features and LSP-related features play important role in speech emotion recognition.

Table 3 Comparison with state-of-the-art results.

	F_{inter}	F_{union}	Tawari	Bitouk	Ruvolo
ACC	51.57%	82.76%	74.8%	78.2%	78.7%

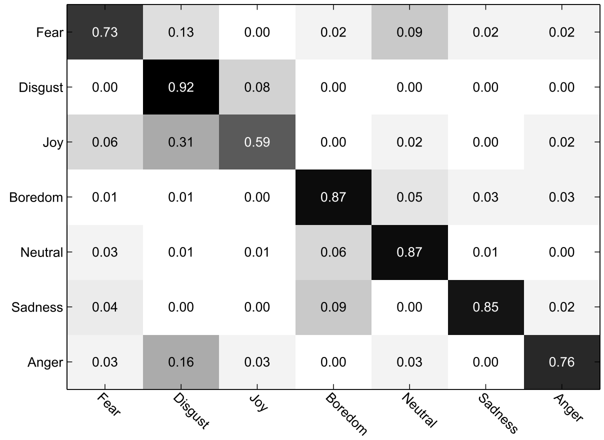

Fig. 3 The average confusion matrix of our two-layer MKL in speech emotion recognition with F_{union}.

With F_{inter} and F_{union}, experiments are carried out for classification based on the second layer of our scheme. Gaussian kernels are adopted with 10 different bandwidths σ on all features same as before and with 4 bandwidths σ on each single feature because the dimension of feature vector has been greatly reduced. To guarantee the speaker-independent, the whole dataset are separated into 10 parts according to 10 speakers. Each time, one speaker is left out for testing and the other 9 speakers are combined for training and make 10-fold cross validation. The average recognition rate using F_{inter} is 51.57% which shows that F_{inter} is highly effective. However, only with 12 features in F_{inter} we can't obtain relatively high recognition rate for 7 kinds of emotion. Using F_{union}, with more complementary features added, the overall recognition rate on this seven-way classification task is 82.76%. Its average confusion matrix is shown in Fig. 3. The results are listed in Table 3 comparing with state-of-the-art results. In the paper of Tawari [12], using the contextual information, the authors obtain 74.8% of weighted accuracy for speaker-independent analysis for seven emotions. In the paper of Bitouk [13], speaker-independent, multi-class emotion classification rates for six emotion task on Berlin datasets using prosodic and spectral features is 78.2%. In the paper of Ruvolo [14], the recognition rate using 10-fold leave one speaker out cross validation is 78.7%. The discrepancy in

Table 4 Comparison with other feature selection methods using F_{union}.

	Our method	FS	FFS	BFS	GS
ACC	80.93%	73.19%	71.63%	71.15%	74.03%

recognition rate is the evidence that the union of features selected by multiple kernel learning method can result in large gains in performance.

4.2 Comparisons with Other Feature Selection Methods

Our proposed feature selection method is also compared with best first search method (BFS), genetic search method (GS), forward selection method (FS) and floating forward selection method (FFS), which are popular feature selection methods. Feature selection techniques provided by WEKA [15] are utilized and SVM is adopted as classifier. For each method, 10 fold cross validation is used to obtain 10 sets of features. As same as before, union of such 10 sets generates a feature vector for classification. Features selected by different algorithms are fed into SVM for training. The parameters C and σ are determined according to the cross validation. Each time, one speaker is left out for test and the other nine speakers are combined for training and make 10 fold cross validation. The average recognition rates are obtained and listed in Table 4. The recognition rate of our feature selection method based on SVM is 80.93%, while the results of FS, FFS, BFS and GS are respectively 73.19%, 71.63%, 71.15% and 74.03%. It is shown that our method outperforms the other methods. Moreover, in order to see the robustness of our feature selection method, the intersection of above 10 sets is also obtained. If a feature selection method always selects the same important features in different conditions, it is thought to be robust. Six features appear in all 10 folds with BF method, one feature appear with FFS method, two features appear with FS method, and none feature appear with GS method. While in our method, 12 features appear in all 10 folds. That means our proposed method is more robust in feature selection.

5. Conclusion

In this paper, a two-layer MKL scheme is presented for speaker-independent speech emotion recognition. In the first layer, MKL is used for feature selection. The training samples are separated into n groups according to some rules. All groups are used for feature selection obtaining n feature subsets. The intersection and the union of all n subsets are the result of our feature selection method. In the second layer, MKL is used again for speech emotion classification with selected features. In order to evaluate our proposed two-layer MKL scheme, the result of our method is compared with state-of-the-art results of speaker-independent speech emotion recognition. The average recognition rate of our two-layer MKL is 82.76% which outperforms state-of-the-art results. Then another experiment is carried out and our feature selection method is compared with other popular feature selection methods. Using selected features with classifier of SVM, the recognition rate of 80.93% is obtained, which also results in large gain in performance.

Acknowledgement

This paper is supported by the National Natural Science Funding of China (No: 61231002, No: 61273266) and funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, vol.18, no.1, pp.32–80, 2001.

[2] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," IEEE International Conf. Multimedia and Expo, 2005, pp.864–867, 2005.

[3] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition," Signal Processing, vol.88, no.12, pp.2956–2970, 2008.

[4] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," Proc. Twenty-First International Conference on Machine Learning, p.6, ACM, 2004.

[5] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," Journal of Machine Learning Research, vol.7, pp.1531–1565, 2006.

[6] C. Yeh, C. Huang, and S. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," Expert Systems with Applications, vol.38, no.3, pp.2177–2186, 2011.

[7] M. Tan, L. Wang, and I. Tsang, "Learning sparse svm for feature selection on very high dimensional datasets," Proc. International Conference on Machine Learning, pp.1047–1054, 2010.

[8] Z. Xu, R. Jin, H. Yang, I. King, and M. Lyu, "Simple and efficient multiple kernel learning by group lasso," Proc. 27th International Conference on Machine Learning, pp.1175–1182, 2010.

[9] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," Journal of Machine Learning Research, vol.9, pp.2491–2521, 2008.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," Proc. Interspeech, 2005.

[11] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR—Introducing the munich open-source emotion and affect recognition toolkit," 3rd International Conf. Affective Computing and Intelligent Interaction and Workshops, 2009, pp.1–6, 2009.

[12] A. Tawari and M. Trivedi, "Speech emotion analysis: Exploring the role of context," IEEE Trans. Multimedia, vol.12, no.6, pp.502–509, 2010.

[13] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," Speech Communication, vol.52, no.7, pp.613–625, 2010.

[14] P. Ruvolo, I. Fasel, and J. Movellan, "A learning approach to hierarchical feature selection and aggregation for audio classification," Pattern Recognition Letters, vol.31, no.12, pp.1535–1542, 2010.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," ACM SIGKDD Explorations Newsletter, vol.11, no.1, pp.10–18, 2009.