A FEATURE SELECTION AND FEATURE FUSION COMBINATION METHOD FOR SPEAKER-INDEPENDENT SPEECH EMOTION RECOGNITION

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

[1] Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, P.R.China
[2]School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou, P.R.China
[3]Key Laboratory of Child Development and Learning Science, Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, P.R.China
{jinyun9999,songpengseu}@gmail.com, {wenming_zheng,zhaoli}@seu.edu.cn

ABSTRACT

To enhance the recognition rate of speaker independent speech emotion recognition, a feature selection and feature fusion combination method based on multiple kernel learning is presented. Firstly, multiple kernel learning is used to obtain sparse feature subsets. The features selected at least n times are recombined into another subset named n-subset. The optimal n is determined by 10 cross-validation experiments. Secondly, feature fusion is made at the kernel level. Not only each kind of feature is associated with a kernel, but also the full feature set is associated with a kernel which is not considered in the previous studies. All of the kernels are added together to obtain a combination kernel. The final recognition rate for 7 kinds of emotions on Berlin Database is 83.10%, which outperforms state-of-the-art results and shows the effectiveness of our method. It is also proved that MFCCs play a crucial role in speech emotion recognition.

Index Terms— speech emotion recognition, feature selection, feature fusion, multiple kernel learning

1. INTRODUCTION

Speech emotion recognition makes human-computer interaction(HCI) more closer to human-human interaction and makes its applications more usable and friendly. Feature selection and feature fusion are two popular research directions in speech emotion recognition.

Extracting a limited, meaningful, and informative set of features is an important step in automatic recognition of emotions [1]. So a lot of feature selection strategies have been put forward. Principle component analysis (PCA) [2] and linear discriminate analysis (LDA) [3] are two commonly used feature reduction techniques which project the input space onto a less dimensional one and hold as much information as possible. The filter methods utilize intrinsic properties of data as the criterion for feature subset evaluation, such as correlation-based solution [4]. The wrapper methods depend on the classifier's accuracy to select feature subsets, such as forward feature selection (FFS) [5] and sequential floating forward selection (SFFS) [6]. Feature selection methods based on kernel are also presented. Support vector machine was utilized for feature selection[7]. Using multiple kernels instead of one single kernel, the feature selection method based on multiple kernel learning (MKL) is then proposed [8].

Traditional feature fusion methods in speech emotion recognition simply concatenate different kinds of features into one large vector. For traditional kernel method, it is mapped into a high dimensional space with a single kernel function. However, different kinds of features are with different distribution in space. Such simple concatenation sometimes will lost some important classification information. Therefore, multiple kernels are adopted.

The outline of the paper is as follows. In section 2, MKL is simply reviewed and our feature selection and feature fusion combination method is proposed. In section 3, the database is introduced and the features extracted are listed. In order to prove the effectiveness of our method, experiments are conducted in section 4. Conclusion is given in section 5.

2. THE PROPOSED FEATURE SELECTION AND FEATURE FUSION COMBINATION METHOD

2.1. Review of MKL

Kernel methods such as support vector machine(SVM) have been proved to be effective for classification or regression problem during the past two decades. Let $\{x_i, y_i\}_{i=1}^{l}$ be the training samples, where x_i belongs to some input space $\mathcal{X}$ and y_i is the label of pattern x_i. The statement of the kernel learning problem can be written as follows:

$$f(x) = \sum_{i=1}^{l} \alpha_i^* K(x, x_i) + b^* \tag{1}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

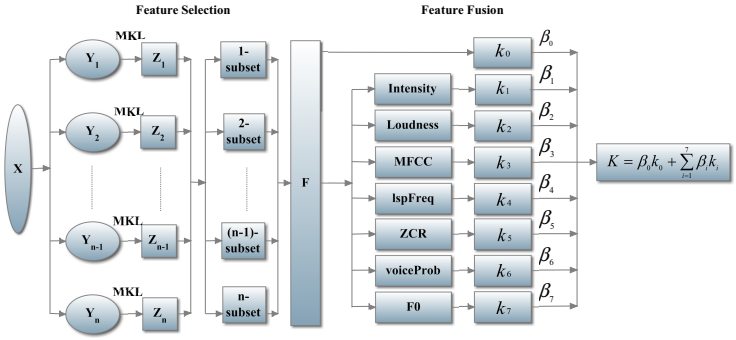

Fig. 1. Flowchart of the proposed feature selection and feature fusion combination method.

A single kernel can't accurately depict the data representation in space. So using multiple kernels instead of a single one can improve the performances. The kernel $K(x, x')$ can be considered as a linear combination of basis kernels:

$$K(x, x') = \sum_{m=1}^{M} \beta_m K_m(x, x') \tag{2}$$
$$s.t. \quad \beta_m \geq 0, \sum_{m=1}^{M} \beta_m = 1.$$

where M is the total number of kernels. Each basis kernel K_m may either use the full set of features or subsets of features from different data sources [9]. Such characteristics is used for feature fusion in our method. The kernels K_m can be gaussian kernels, polynomial kernels, exponential Kernels and so on. If the weights β_m are obtained, the data representation will be determined. Learning both the parameters α_i and the weights β_m in a single optimization problem is called MKL problem. Rakotomamonjy proposed the following constrained optimization problem for MKL and derived its dual problem[10].

$$\min_{\beta} J(\beta) \tag{3}$$
$$s.t. \sum_{m=1}^{M} \beta_m = 1, \beta_m \geq 0,$$

where

$$J(\beta) = \begin{cases} \min\limits_{\{f\},b,\xi,\beta} \frac{1}{2}(\sum\limits_{m} ||f_m||_{\mathcal{H}_m})^2 + C\sum\limits_{i} \xi_i \\ s.t. \quad y_i \sum\limits_{m} f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ \xi_i \geq 0 \quad \forall i. \end{cases} \tag{4}$$

And its dual problem is derived as follows:

$$\max_{\alpha} -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_{m} \beta_m K_m(x_i, x_j) + \sum_{i} \alpha_i$$
$$with \quad \sum_{i} \alpha_i y_i = 0 \tag{5}$$
$$C \geq \alpha_i \geq 0 \quad \forall i$$

An algorithm SimpleMKL in [10] is to solve the above optimization problem.

The constraint $\sum_{m=1}^{M} \beta_m = 1, \beta_m \geq 0$ in (3) called L_1-norm tends to result in a sparse solution of β_m. The majority of redundant kernels will be rejected and some important kernels will be kept. Therefore, MKL can be used for feature selection.

2.2. The proposed method

In this section, the proposed method will be introduced in details and the flowchart is shown in Fig.1.

L_1-norm constraint of MKL will lead to sparse solution on β_m. Specifically, an utterance is denoted by a n-dimensional vector $\mathbf{x} = [x_1, \cdots, x_n]^T$ and each feature x_j is associated with a kernel k_j. Then the combination kernel $\sum_{j=1}^{n} \beta_j k_j$ is obtained. Using L_1-norm of MKL, most of the kernel weights are forced to zero, and only the important ones are retained. The corresponding features are determined. This is the fundamental of our feature selection method. However, complementary information may be discarded if base kernels encode orthogonal information [11]. That means some useful features maybe abandoned during feature selection process. To compensate the lost information in this scenario, our feature selection method is proposed.

The total training samples are denoted by X, which are randomly split into N parts. $X = (X_1, \cdots, X_N)$. Each time, one part is left out and the remainder $(N-1)$ parts are recombined into a new group denoted by Y_i. $Y_i = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_N), i = (1, \cdots, N)$. Using SimpleMKL, feature selection is carried out on each

group Y_i to produce sparse feature subsets denoted by $Z_i, i = (1, \cdots, N)$. The reason why N groups are used for feature selection is to avoid some important features being randomly removed during one process[12]. Moreover, the frequency of all features selected in N subsets is computed. The features selected at least n times are regrouped as a subset named n-subset ($n = 1, \cdots, N$). Different n will lead to different recognition rates. So experiments will be conducted to determine the optimal n and the best n-subset F which is the final result of feature selection.

The traditional feature fusion methods only adopt one kernel for mapping, which is not enough to depict the feature space distribution. In some studies, multiple kernels which are associated with each kind of features are combined to replace the single one. The local information of each kind of features is utilized in such combination, however, the global information of the full feature set is missing. Therefore in our method, a kernel associated with the full feature set is added into the combined kernel. And the experiment results show that such a kernel is critical in speech emotion recognition. Specifically, the full feature set F comprises of M subsets, $F = [F^{(1)}, \cdots, F^{(M)}]$. Each subset $F^{(i)}$ is associated with a kernel $k_i, (i = 1, \cdots, M)$. k_0 is a kernel associated with the full set F. The modified combination kernel K is written as follows:

$$K = \beta_0 k_0 + \sum_{m=1}^{M} \beta_m k_m \tag{6}$$
$$s.t. \;\; 0 \le \beta_m \le 1, \quad \sum_m \beta_m = 1, \quad m = 0, \cdots, M,$$

where β_m are the kernel weights and are determined by solving (5).

3. THE DATABASE AND FEATURE EXTRACTION

In this section, the Berlin Emotional Speech Dataset is introduced and the features extracted in our experiments are listed.

3.1. Berlin Emotional Speech Dataset

The Berlin Emotional Speech Dataset [13] is one of the most popular dataset used by researchers for emotion recognition. It contains the emotional utterances recorded by 10 German actors (5 female) reading one of 10 pre-selected sentences. The utterances cover the following seven kinds of emotions: *anger, boredom, fear, disgust, joy, sadness* and *neutral*. There are initially about 900 utterances in it. After a listening test by 20 judgers, only 494 sentences are kept.

3.2. Feature extraction

With the openEAR toolkit[14], the features are extracted as 19 functionals of 26 acoustic low-level descriptors(LLD) and corresponding first order delta. The 26 Low-level descriptors used in the experiments are listed in Table 1. The statistical functionals and regression coefficients are listed in Talbe 2. The feature vector per utterance contains $26 \cdot 2 \cdot 19 = 988$ attributes.

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 and regression coefficients

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

4. EXPERIMENT

In this section, experiments will be conducted to show the performance of our proposed method.

4.1. Feature Selection

Firstly, gaussian kernels are adopted with 10 different bandwidths σ (0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64) on full feature set and with 1 bandwidth σ on each single feature. There are totally 998 kernels (10 kernels for full feature set and 988 kernels for each single feature). The value of σ for each single feature is set 1 according to cross validation. Only 1 bandwidth σ is chosen because of the limitation of computer.

Table 3. The number of features with same frequency

Freq	1	2	3	4	5	6	7	8	9	10
Number	24	13	9	6	6	5	5	3	6	12

The dataset is split into 10 parts according to empirical experience. 10 feature subsets are selected from 10 groups. The frequency of each feature is computed and the number of feature with same frequency is listed in Table 3. For example, there are 12 features selected in all 10 groups and 24 features selected only in one group. The features selected at least n times are regrouped named n-subset ($n = 1, \cdots, 10$). Experiments are conducted to determine the optimal n using SVM. 10-fold cross validation is carried out. The average recognition rates are shown in Fig. 2. When n equals to 2, the best recognition rate 81.5% is reached. It means the features at least selected for twice in 10 groups are the optimal feature subset F. There are totally 65 features in it which includes 6 kinds of features. The number of loudness-related features, MFCC-related features, lspFreq-related features, zcr-related features, voiceProb-related features and F0-related features are respectively 2, 27, 22, 2, 8, 4. It is noted that the intense-related features are abandoned in feature selection.

4.2. Feature Fusion

F is utilized for feature fusion. Gaussian kernels are adopted with 10 different bandwidths σ on the full feature set F (10 kernels) and with 5 bandwidth σ (0.5, 1, 2, 4, 8) on each kind of features (30 kernels). There are totally 40 kernels. To guarantee the speaker-independent, the whole dataset is 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. 10-fold cross validation is carried out.

The average recognition rate is 83.10%. Because the weights of the kernels are various in each fold, the weights in one fold are listed in Table 4. The weight of the kernel for F is 0.8216, which means that the full feature set plays the most important role. The weight of the MFCC-realted features is much higher than those of the other features, which show its importance in speech emotion recognition.

The recognition rate of our method is listed in Table 5 comparing with state-of-the-art results which are all based on

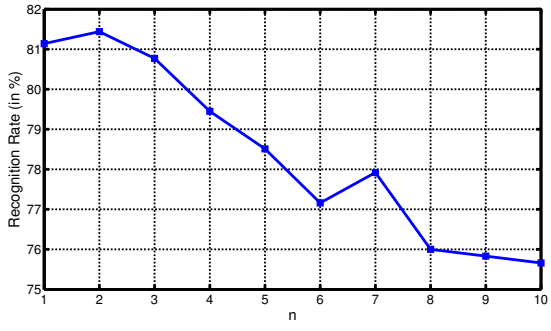

Fig. 2. The average recognition rates using n-subset with different n ($n = 1, \cdots, 10$).

Table 4. The weights of the kernels in one fold

Feature	loudness	MFCC	lspFreq	zcr
Weight	0.002	0.1042	0.0066	0.0168
Feature	voiceProb	F0	F	Total
Weight	0.0257	0.0231	0.8216	1

Table 5. Comparison with state-of-the-art results

Our method	Tawari	Bitouk
83.10%	74.8%	78.2%
Ruvolo	Zhang	Bhargava
78.7%	80.85%	80.60%

Berlin dataset and for speaker-independent. Using the contextual information, Tawari obtained 74.8% of weighted accuracy for seven emotions[15]. Multi-class emotion classification rates for six emotion task using prosodic and spectral features is 78.2% by Bitouk[16]. Ruvolo acquired the recognition rate of 78.7% using 10-fold cross validation[17]. Zhang obtained the recognition rate 80.85% for 7 kinds of emotions using an enhanced kernel isomap[18]. Bhargava acquired 80.60% using rhythm and temporal feature[19]. The discrepancy in recognition rate is the evidence that our proposed method can result in large gains in performance for speech emotion recognition.

5. CONCLUSIONS

In the paper, we proposed a feature selection and feature fusion combination method based on MKL. Firstly, 10 groups of feature subsets are obtained. The frequency of the features in all 10 subsets are calculated. Experiments are conducted to determine the optimal subsets. Secondly, each kind of feature is associated with a kernel. Added with another kernel associated with the full feature set, a combination kernel is obtained. The final recognition rate of speaker-independent speech emotion recognition on Berlin Database is 83.10%, higher than state-of-the-art results, which demonstrates the effectiveness of our method. Moreover, it is proved that the MFCC-related features play the most important role in speech emotion recognition.

6. ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Funding of China (No: 61231002, No: 61273266), the Natural Science Foundation of Jiangsu Province under Grant BK20130020 and the Ph.D. Program Foundation of Ministry Education of China under Grant 20120092110054.

7. REFERENCES

[1] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[2] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.

[3] Keinosuke Fukunaga, *Introduction to statistical pattern recognition*, Academic Pr, 1990.

[4] Lei Yu and Huan Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, 2003, vol. 20, p. 856.

[5] Björn Schuller, Stephan Reiter, R Muller, Marc Al-Hames, Manfred Lang, and Gerhard Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 864–867.

[6] Dimitrios Ververidis and Constantine 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.

[7] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik, "Feature selection for svms," *Advances in neural information processing systems*, pp. 668–674, 2001.

[8] Niranjan Subrahmanya and Yung C Shin, "Sparse multiple kernel learning for signal processing applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 788–798, 2010.

[9] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 6.

[10] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, Yves Grandvalet, et al., "Simplemkl," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[11] Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, and Michael R Lyu, "Simple and efficient multiple kernel learning by group lasso," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 1175–1182.

[12] JIN Yun, SONG Peng, ZHENG Wenming, ZHAO Li, and XIN Minghai, "Speaker-independent speech emotion recognition based on two-layer multiple kernel learning," *IEICE TRANSACTIONS on Information and Systems*, vol. 96, no. 10, pp. 2286–2289, 2013.

[13] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier, and Benjamin Weiss, "A database of german emotional speech," in *Proc. Interspeech*, 2005.

[14] Florian Eyben, Martin Wollmer, and Bjorn Schuller, "Openearłintroducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

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

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

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

[18] Shiqing Zhang, Xiaoming Zhao, and Bicheng Lei, "Speech emotion recognition using an enhanced kernel isomap for human-robot interaction," *INTERNATIONAL JOURNAL OF ADVANCED ROBOTIC SYSTEMS*, vol. 10, 2013.

[19] Mayank Bhargava and Tim Polzehl, "Improving automatic emotion recognition from speech using rhythm and temporal feature," *arXiv preprint arXiv:1303.1761*, 2013.