ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 325－330
DOI:10.16337/j.1004-9037.2016.02.012

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于稀疏特征迁移的语音情感识别

宋　鹏[1]　金　赟[2]　查　诚[3]　赵　力[2,3]

(1.烟台大学计算机与控制工程学院,烟台,264005;2.东南大学儿童发展与学习科学教育部重点实验室,南京,210096;3.东南大学信息科学与工程学院,南京,210096)

摘　要:为了解决语音情感识别系统中训练数据和测试数据来自不同数据库所引起的识别率降低的问题,提出了一种基于稀疏特征迁移的语音情感识别方法。通过引入稀疏编码获取情感特征在不同数据库条件下的共同稀疏表示;同时引入最大区分差异(Maximum mean discrepancy, MMD)来衡量不同数据库条件下稀疏表示分布之间的距离,并将其作为稀疏编码目标函数的约束条件,从而获得较为鲁棒的稀疏特征。实验结果表明,相比传统语音情感识别方法,基于稀疏特征迁移的语音情感识别方法显著提高了跨库条件下的情感识别率。

关键词:语音情感识别;特征迁移;稀疏编码

中图分类号:TN912.3　　**文献标志码**:A

Speech Emotion Recognition Using Sparse Feature Transfer

Song Peng[1],Jin Yun[2],Zha Cheng[3],Zhao Li[2,3]

(1. School of Computer and Control Engineering,Yantai University,Yantai,264005, China; 2. Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University,Nanjing,210096, China; 3. School of Information Science and Engineering,Southeast University,Nanjing,210096, China)

Abstract: In speech emotion recognition system, recognition rates will drop drastically when the training and the testing utterances are from different corpora. To solve this problem, a novel sparse feature transfer approach is proposed. By employing sparse coding algorithm, the common sparse feature representation of emotion features from different corpora is obtained. Meanwhile, the maximum mean discrepancy (MMD) algorithm is introduced to measure the distance between different distributions, and is used as the regularization term for the objective function of sparse coding. Finally, the robust sparse features are achieved for recognition. Experimental results show that, compared to traditional methods, the proposed approach can significantly improve the recognition rates for cross databases.

Key words: speech emotion recognition; feature transfer; sparse coding

引　　言

作为情感计算的一个重要分支,语音情感识别受到人们越来越多的关注。语音情感识别指从语音

基金项目:山东省自然科学基金(ZR2014FQ016,ZR2015PF010)资助项目;国家自然科学基金(61273266, 61403328, 61403329)资助项目;东南大学基本科研业务费(CDLS-2015-04)资助项目。
收稿日期:2014-11-16;**修订日期**:2015-01-03

中识别出喜、怒、哀、乐和愁等情感的一种技术[1]。它有着广泛的应用前景，如在医疗卫生领域，通过语音情感的识别来对病人的心理状态进行判断来辅助治疗；在刑侦领域，通过对犯罪嫌疑人的情感状态进行实时监控来推进案件审理的进程；在语音翻译系统中，通过对情感的自动分析从而在目标语句中合成出相应情感的语句等[2]。

为了有效地从语音中识别出情感信息，许多学者提出了各种方法，包括隐马尔科夫模型(Hidden Markov model, HMM)法、支持向量机(Support vector machine, SVM)法、高斯混合模型(Gaussian mixture model, GMM)法、人工神经网络(Artificial neural network, ANN)法及近来提出的深度神经网络(Deep neural network, DNN)法和回归方法等[1-4]，这些方法都在一定程度上实现了情感特征的有效分类。但是这些方法主要针对单一数据库的情况提出，在实际环境中，往往面对多个数据库情况。在这种情况下，从一个数据库中训练得到的分类器直接用于另一数据库下的情感识别将引起识别率的急剧下降。

目前对于跨库语音情感识别的研究还不是很多，文献[5]将无监督的学习方法用于包含 6 个数据库的跨库语音情感识别，从唤醒度和效价维两个维度对情感进行分类，在一定程度上提高了情感识别率；文献[6]对跨语言的语音情感识别进行了初步的讨论；文献[7]提出了一种基于无监督自适应编码的语音情感识别方法；文献[8]将迁移学习方法用于不同数据库条件下的语音情感识别，这些方法在跨库条件下取得了比传统识别方法更好的实验效果。但是这些研究主要是从寻找共同特征空间的角度出发来解决不同数据库的差异性，存在着运算量大及当数据库差异较大时识别率提升有限等问题。

不同于上述方法，受稀疏编码[9]和迁移学习[10]等方法的启发，本文从稀疏特征表示的角度出发，提出了一种有效的基于稀疏特征迁移的语音情感识别方法。首先引入稀疏编码方法对语音情感特征通过从训练数据中进行字典学习，训练获得特征的稀疏表示；其次引入迁移学习中广泛应用的最大区分差异(Maximum mean discrepancy, MMD)[11]来描述两个数据库特征分布之间的距离，并将其作为稀疏编码的约束条件，从而可以获得满足不同情感数据库条件的鲁棒稀疏特征表示，进而有效实现跨库条件下的语音情感识别。

1 稀疏特征迁移的语音情感识别

本文提出的基于稀疏特征迁移的语音情感识别方法如图 1 所示。在训练阶段，给定有类别标签的源情感语音库和无类别标签的目标情感语音库，分别提取它们的情感特征，通过稀疏特征迁移算法分别计算得到字典矩阵 $\boldsymbol{D}$ 和稀疏特征 $\boldsymbol{S}$，接着利用 $\boldsymbol{S}$ 训练得到情感分类器；在测试阶段，利用训练阶段得到的字典矩阵 $\boldsymbol{D}$ 估计测试语音情感特征的稀疏表示，并通过情感分类器对其进行分类得到情感类别标签。

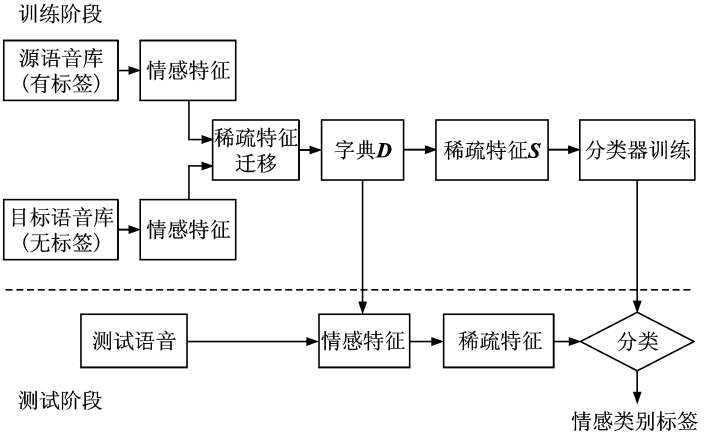

图 1 基于稀疏特征迁移的语音情感识别方法

Fig. 1 Speech emotion recognition using sparse feature transfer

1.1 情感特征的稀疏表示

与图像特征表示类似，在语音情感识别中提取到的情感特征往往具有非常高的维度(从几百维到上千维)，其具有很高的冗余度，在很大程度上会影响情感识别率，尤其跨库条件下的语音情感识别结果。受稀疏编码在人脸识别[12]及图像分类[9]等领域获得了成功

应用的启发,对于语音情感特征,同样可以从稀疏表示的角度出发,从情感语音特征 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in\mathbf{R}^{m\times n}$ 中学习到字典矩阵,并用其来获取相应的稀疏特征表示。假设 $\boldsymbol{D}=[d_1,d_2,\cdots,d_k]\in\mathbf{R}^{m\times k}$ 为字典矩阵,$\boldsymbol{S}=[s_1,s_2,\cdots,s_n]\in\mathbf{R}^{k\times n}$ 为稀疏特征矩阵,则可以通过如下的目标函数对 $\boldsymbol{D}$ 和 $\boldsymbol{S}$ 进行求解

$$\arg\min_{\boldsymbol{D},\boldsymbol{S}}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{S}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{S}\|_1\quad\text{s.t.}\quad\|\boldsymbol{d}_i\|^2\leqslant c,\quad i=1,2,\cdots,k\tag{1}$$

式中:$\|\cdot\|_{\mathrm{F}}$ 为 L_2 范数;$\|\cdot\|_1$ 为 L_1 范数;λ 为规整因子,用来调节规整项中的特征稀疏度及特征近似表示之间的平衡关系;c 为常数。

对于未知参数 $\boldsymbol{D}$ 和 $\boldsymbol{S}$ 的求解是一个非凸的最优化问题,可以通过迭代的方法来求解[9]。首先固定字典 $\boldsymbol{D}$,求解 $\boldsymbol{S}$,则问题转换为经典的 LASSO(Least absolute shrinkage and selection operator)问题进行求解;其次固定 $\boldsymbol{S}$,调整 $\boldsymbol{D}$,则问题转换为经典二次规划(Quadratic programming,QP)问题进行求解;反复迭代直到收敛,最后可以获取一组能很好地对原始特征进行稀疏表示的字典 $\boldsymbol{D}$。

1.2　稀疏特征迁移

通过稀疏特征编码在一定程度上可以获得同时满足不同数据库的共同稀疏特征表示,但是由于不同语音情感数据库的特征分布差别很大[8],导致获得的稀疏特征表示不够准确,会显著影响跨库条件下的语音情感识别率。同时,不同的语音情感数据库特征分布差别很大,且实际情况中大量有标签的情感数据很难直接获取。迁移学习是一种通过迁移相关知识来解决邻域中仅有少量标签甚至没有标签的一种机器学习方法[10],受这一思想的启发,从稀疏编码和迁移学习相结合的角度出发,本文提出了一种基于稀疏特征迁移的语音情感识别方法,考虑了不同情感数据库获取的稀疏特征分布之间的相似度,并将其作为稀疏编码目标函数的约束项,从而获得更为鲁棒的稀疏特征表示。

给定一个有类别标签的源语音情感数据库和一个无类别标签的目标语音情感数据库,分别提取它们的情感特征,表示为 $\boldsymbol{X}_{\mathrm{src}}=[x_1,x_2,\cdots,x_{n_l}]$ 和 $\boldsymbol{X}_{\mathrm{tar}}=[x_{n_l+1},x_{n_l+2},\cdots,x_{n_l+n_u}]$,其中 n_l 和 n_u 分别表示有标签和无标签的特征数。训练目标是通过稀疏编码找到一个能够满足 $\boldsymbol{X}_{\mathrm{src}}$ 和 $\boldsymbol{X}_{\mathrm{tar}}$ 稀疏表示的字典 $\boldsymbol{D}$,同时能够最小化有标签和无标签稀疏特征分布之间的距离。本文采用最大区分差异 MMD[11] 来描述两个分布之间的距离,表示为

$$\mathrm{Dist}=\left\|\frac{1}{n_l}\sum_{i=1}^{n_l}\boldsymbol{s}_i-\frac{1}{n_u}\sum_{j=n_l+1}^{n_l+n_u}\boldsymbol{s}_j\right\|^2=\sum_{i,j=1}^{n_l+n_u}\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{s}_j m_{ij}=\mathrm{tr}(\boldsymbol{S}\boldsymbol{M}\boldsymbol{S}^{\mathrm{T}})\tag{2}$$

式中:$\mathrm{tr}(\cdot)$ 表示矩阵的迹;$\boldsymbol{M}=\{m_{ij}\}_{i,j=1,2,\cdots,n_l+n_u}$ 为 MMD 矩阵,满足

$$m_{ij}=\begin{cases}\dfrac{1}{n_l^2} & \boldsymbol{s}_i,\boldsymbol{s}_j\in\boldsymbol{X}_{\mathrm{src}}\\ \dfrac{1}{n_u^2} & \boldsymbol{s}_i,\boldsymbol{s}_j\in\boldsymbol{X}_{\mathrm{tar}}\\ -\dfrac{1}{n_l n_u} & \text{其他}\end{cases}\tag{3}$$

将式(2)作为稀疏编码目标函数的约束,则式(1)变为

$$\arg\min_{\boldsymbol{D},\boldsymbol{S}}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{S}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{S}\|_1+\gamma\mathrm{tr}(\boldsymbol{S}\boldsymbol{M}\boldsymbol{S}^{\mathrm{T}})\quad\text{s.t.}\quad\|\boldsymbol{d}_i\|^2\leqslant c,\quad i=1,2,\cdots,k\tag{4}$$

式中:γ 为规整系数,用于调整稀疏度和特征分布相似度的平衡关系,当 $\gamma=0$ 时,则稀疏特征迁移的方法等价于传统的稀疏编码方法。

1.3　稀疏特征与字典求取

式(4)是一个非凸的最优化问题,与传统稀疏编码的求解方式类似,可以通过采用迭代的方法来估计未知参数 $\boldsymbol{D}$ 和 $\boldsymbol{S}$。

(1)固定字典 $\boldsymbol{D}$,则式(4)变为如下的形式

$$\arg\min_{\boldsymbol{S}} \| \boldsymbol{X}-\boldsymbol{D}\boldsymbol{S} \|_{\mathrm{F}}^{2}+\lambda\sum_{i=1}^{n_l+n_u}|\boldsymbol{s}_i|+\gamma\mathrm{tr}(\boldsymbol{S}\boldsymbol{M}\boldsymbol{S}^{\mathrm{T}}) \tag{5}$$

通过坐标下降法对式(5)进行求解,在每一步中固定所有的$\{\boldsymbol{s}_j\}_{j\neq i}$,对$\boldsymbol{s}_i$进行更新寻找目标函数最小值,则求解$\boldsymbol{s}_i$的目标函数表示为

$$\arg\min_{\boldsymbol{s}_i} \| \boldsymbol{x}_i-\boldsymbol{D}\boldsymbol{s}_i \|^{2}+\lambda\sum_{j=1}^{k}|\boldsymbol{s}_i^j|+\gamma\mathrm{tr}(m_{ii}\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{s}_i) \tag{6}$$

通过特征标记查询算法[13]对式(6)进行求解。

(2)固定$\boldsymbol{S}$,对字典$\boldsymbol{D}$进行求解,则式(4)变为如下的最小二乘求解的形式

$$\arg\min_{\boldsymbol{D}} \| \boldsymbol{X}-\boldsymbol{D}\boldsymbol{S} \|_{\mathrm{F}}^{2} \tag{7}$$

通过传统 QP 算法对式(7)进行求解。然后将新求得的$\boldsymbol{D}$重新代入式(6),对上述两个过程进行重复迭代,直到收敛。在测试阶段,利用训练得到的字典来获得输入特征的稀疏表示进行情感分类。

2 实验结果与分析

2.1 实验环境

本文采用两个经典语音情感数据库对提出的算法进行评价,分别是 Berlin 库[14]和 eNTERFACE 库[15]。其中 Berlin 库是一个由专业演员录制的德语情感语音库,包含 10 名说话人、494 条语句以及生气、烦躁、厌恶、害怕、高兴、伤心及中性等 7 类情感。而 eNTERFACE 库是一个由非专业演员录制的英文音视频情感库,它包含 42 位说话人、1 170 条情感语句及生气、厌恶、害怕、高兴、悲伤和惊讶等 6 种情感。在本次试验中采用这两个数据库共有的生气、厌恶、害怕、高兴、伤心这 5 类情感语句对相关算法的识别结果进行比较。

为了验证本文提出的稀疏特征迁移算法的有效性,设计了两种实验对识别结果进行评价:(1)采用 eNTERFACE 库作为有标签的训练数据,Berlin 库作为无标签的测试数据;(2)采用 Berlin 库作为有标签的训练数据,eNTERFACE 库作为无标签的测试数据。利用 openSMILE 工具箱[16]对情感语音进行特征提取。实验采用 Interspeech 2010 语音情感识别竞赛所采用的情感特征集[17],共 1 582 维特征,包含 38 个底层描述符(Low-level descriptors, LLDs)及其一阶差分,如表 1 所示。采用 21 个统计函数作用于上述 76 个 LLDs,同时舍弃 16 个零信息量的特征,并将基频数和时长信息加入特征集。

表 1 实验采用的 LLDs

Tab. 1 LLDs for the evaluations

LLDs	数量
响度	1
MFCC[0～14]	15
Log 域梅尔频带[0～7]	8
LSP [0～7]	8
基音频率	1
基音频率包络	1
浊音频率	1
局部抖动	1
连续抖动帧对	1
局部微扰	1

本文采用 SVM 作为情感分类器,采用基于主成分分析(Principal component analysis, PCA)做特征降维。经过特征降维后,分别比较了以下几种跨库条件下的语音情感识别算法:传统的 SVM 分类方法[2]、提出的基于稀疏编码(Sparse coding, SC)的 SVM 分类方法及提出的基于稀疏特征迁移(Sparse feature transfer, SFT)的 SVM 分类方法。同时选择基于 SVM 的单一情感库下的分类方法作为基线方法。最终通过情感识别率和混淆矩阵对实验结果进行评价。

在方案 1 中将 eNTERFACE 库的全部数据作为训练,将 Berlin 库的数据分成 5 份,每次选择 4/5 的无类别标签的数据用作训练,剩余的用于测试;同时,在方案 2 中将 Berlin 库的全部数据作为训练,将 eNTERFACE 库的无类别标签数据分成 5 份,每次选择 4/5 的数据用作训练,剩余的用于测试,共进行 5 重交叉验证来对λ和γ优化选择,其中λ在{0.001,0.005,0.010,0.050,0.100,0.500,1.000,5.000,10.000,50.000,100.000}中进行选择,最终被优化为 0.1;$\gamma\in[10^3,10^6]$,最终被优化设定为10^5,字典的大小k从{32,64,128,256,512}中进行选择[18]。不同测试环境取得最优识别率情况下 PCA 降维后的维度d的 5 次测试值分别为:129,134,135,131 和 132,取 5 次测试的平均取值,则d取值为 132。因此在本实验

中,k 被优化设定为 256。

2.2　结果与分析

表 3 给出了不同方案下得到的识别率,对于基线方法,分别选择测试库中 4/5 的有类别标签的数据作为训练,剩余的用作测试。从表 2 中可以发现,对方案 1,2 采用稀疏编码的方法都可以取得明显优于传统方法的结果,这说明采用稀疏编码方法可以提取到更能体现情感信息的特征表示;同时可以观察到,采用稀疏特征迁移的方法提升了稀疏编码的效果,这说明加入 MMD 约束条件的有效性,在提取稀疏特征的同时,考虑减少不同数据库之间的特征分布可以有效提高情感识别率。

表 2　不同方法得到的情感识别率比较

Tab. 2　Comparison of recognition rates using different methods　%

实验方案	不同方法的识别率			
	基线方法	传统方法	SC	SFT
方案 1	81.96	34.28	40.15	54.38
方案 2	62.05	23.15	31.68	45.23

图 2,3 分别给出了两种方案下采用稀疏特征迁移方法在取得最高识别率时得到的混淆矩阵。从图 2 可以发现,对于第 1 种方案,无论厌恶还是伤心,都取得了较高的识别率,分别为 79%和 77%,这表明本文提出的方法在很大程度上可以较好的区分这两类情感。从图 3 可以看到,对于第 2 种方案,生气、高兴及伤心这 3 类情感的识别率都超过了 50%;同时可以发现无论是哪一种情感,相对于方案 1,在方案 2 中的识别率整体比较低,这与表 1 中传统基线方法在单一数据条件下的识别结果相一致。

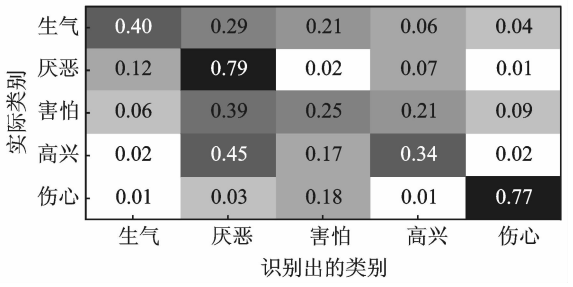

图 2　方案 1 得到的混淆矩阵

Fig. 2　Confusion matrix of scheme 1

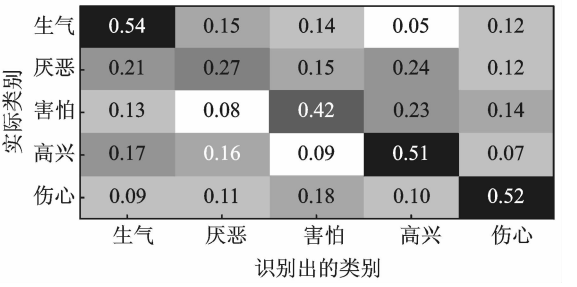

图 3　方案 2 得到的混淆矩阵

Fig. 3　Confusion matrix of scheme 2

3　结束语

针对跨库条件下的语音情感识别,本文提出了一种有效的稀疏特征迁移方法。首先,采用稀疏编码方法来获得不同数据库情感特征的稀疏表示;其次,考虑了不同数据库情感特征分布之间的差异,将其作为稀疏编码目标函数的约束条件,来获得满足不同情感数据库的鲁棒稀疏特征表示;最后,在经典的 Berlin 库和 eNTERFACE 库上对本文提出的方法进行语音情感识别实验评价。实验结果表明,相比传统方法,本文提出的稀疏特征迁移法可以有效提高跨库条件下的情感识别率。目前的实验仅在两个数据库上进行评价,下一步工作将引入更多的数据库对提出的算法的有效性进行验证。

参考文献:

[1]　Ayadi El M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases [J]. Pattern Recognition, 2011, 44(3): 572-587.

[2]　赵力, 黄程韦. 实用语音情感识别中的若干关键技术[J]. 数据采集与处理, 2014, 29(2): 157-170.
Zhao Li, Huang Chengwei. Key technologies in practical speech emotion recognition[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 157-170.

[3] Stuhlsatz A, Meyer C, Eyben F, et al. Deep neural networks for acoustic emotion recognition: Raising the benchmarks[C]// Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, USA: IEEE Signal Processing Society, 2011: 5688-5691.

[4] 余华, 黄程韦, 金赟, 等. 基于粒子群优化神经网络的语音情感识别[J]. 数据采集与处理, 2011, 26(1): 57-62.
Yu Hua, Huang Chengwei, Jin Yun, et al. Speech emotion recognition based on particle swarm optimizer neural network [J]. Journal of Data Acquisition and Processing, 2011, 26(1): 57-62.

[5] Zhang Z, Weninger F, Wollmer M, et al. Unsupervised learning in cross-corpus acoustic emotion recognition[C]// Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA: IEEE Signal Processing Society, 2011: 523-528.

[6] Jeon J H, Le D, Xia R, et al. A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception[C]// Proceedings of Interspeech. Lyon, France: ISCA, 2013: 2837-2840.

[7] Deng J, Zhang Z, Eyben F, et al. Autoencoder-based unsupervised domain adaptation for speech emotion recognition[J]. IEEE Signal Processing Letters, 2014, 21(9): 1068-1072.

[8] Song P, Jin Y, Zhao L, et al. Speech emotion recognition using transfer learning[J]. IEICE Transactions on Information and Systems, 2014, 97(9): 2530-2532.

[9] Huang K, Aviyente S. Sparse representation for signal classification[C]// Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2006: 609-616.

[10] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[11] Gretton A, Borgwardt K M, Rasch M, et al. A kernel method for the two-sample-problem[C]// Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2006: 513-520.

[12] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.

[13] Lee H, Battle A, Raina R, et al. Efficient sparse coding algorithms[C]// Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2006: 801-808.

[14] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]// Proceedings of Interspeech. Lisbon, Portugal: ISCA, 2005:1517-1520.

[15] Martin O, Kotsia I, Macq B, et al. The eNTERFACE'05 audio-visual emotion database[C]// Proceedings of International Conference on Data Engineering Workshops. Atlanta, USA: IEEE Computer Society, 2006: 8.

[16] Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor[C]// Proceedings of International Conference on Multimedia. Firenze, Italy: ACM, 2010: 1459-1462.

[17] Schuller B, Steidl S, Batliner A, et al. The interspeech 2010 paralinguistic challenge[C]// Proceedings of Interspeech. Makuhari, Japan: ISCA, 2010: 2794-2797.

[18] Zheng M, Bu J, Chen C, et al. Graph regularized sparse coding for image representation[J]. IEEE Transactions on Image Processing, 2011, 20(5): 1327-1336.

作者简介:

宋鹏(1983-),男,讲师,研究方向:语音信号处理,E-mail: pengsong@seu.edu.cn。

金赟(1979-),男,讲师,研究方向:语音情感识别。

查诚(1979-),男,博士生,研究方向:语音情感识别。

赵力(1958-),男,教授,研究方向:语音信号处理。