ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 704－716
DOI:10. 16337/j. 1004-9037. 2023. 03. 018

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

基于深度域适应CNN决策树的跨语料库情感识别

孙林慧，赵　敏，王　舜

（南京邮电大学通信与信息工程学院，南京 210003）

摘　要：在跨语料库语音情感识别中，由于目标域和源域样本不匹配，导致情感识别性能很差。为了提高跨语料库语音情感识别性能，本文提出一种基于深度域适应和卷积神经网络（Convolutional neural network，CNN）决策树模型的跨语料库语音情感识别方法。首先构建基于联合约束深度域适应的局部特征迁移学习网络，通过最小化目标域和源域在特征空间和希尔伯特空间的联合差异，挖掘两个语料库之间的相关性，学习从目标域到源域的可迁移不变特征。然后，为了降低跨语料库背景下多种情感间的易混淆情感的分类误差，依据情感混淆度构建CNN决策树多级分类模型，对多种情感先粗分类再细分类。使用CASIA，EMO-DB和RAVDESS三个语料库进行验证。实验结果表明，本文的跨语料库语音情感识别方法比CNN基线方法平均识别率高19.32%～31.08%，系统性能得到很大提升。

关键词：跨语料库语音情感识别；深度域适应；迁移学习；决策树模型；卷积神经网络

中图分类号：TN912.3　　**文献标志码**：A

Cross-Corpus Emotion Recognition Based on Deep Domain Adaptation and CNN Decision Tree

SUN Linhui, ZHAO Min, WANG Shun

(College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In cross-corpus speech emotion recognition, the mismatch between target domain and source domain samples leads to poor performance of emotion recognition. In order to improve the cross-corpus speech emotion recognition performance, this paper proposes a cross-corpus speech emotion recognition method based on deep domain adaptation and convolutional neural network (CNN) decision tree model. Firstly, a local feature transfer learning network based on joint constrained deep domain adaptation is constructed. By minimizing the joint difference between the target and source domains in the feature space and Hilbert space, the correlation between the two corpora is mined and the transferable invariant features from the target domain to the source domain are learned. Then, in order to reduce the classification error of confusable emotions among multiple emotions in the cross-corpus context, a CNN decision tree multi-level classification model is constructed based on the emotional confusion degree, and multiple emotions are first coarsely classified and then finely classified. The experiments are validated using three corpora, CASIA, EMO-DB and RAVDESS. The results show that the average recognition rate of the proposed cross-corpus speech emotion recognition method are 19.32%—31.08% higher than that of CNN baseline method, and

基金项目：国家自然科学基金(61901227)；中国国家留学基金(202008320043)。
收稿日期：2022-07-15；**修订日期**：2023-02-23

the system performance is greatly improved.

Key words: cross-corpus speech emotion recognition; deep domain adaptation; transfer learning; decision tree model; convolutional neural network

引　言

语音情感识别是人机交互重要技术之一,人机交互产品若能独立于多样的文化和语言来解析人类情感,则可能达到真正的拟人化、智能化的人机交互。在现有的自动语音情感识别中,大多数研究使用相同的语料库利用分类器进行训练和测试,这种匹配条件下的情感识别技术已经取得了重大进展,但是要实现真正的跨语言拟人化交互面临一些问题。一方面,国内外并没有建立所有语言的语音情感数据库可用于研究,当然即使有相应的情感数据库,构建多个对应语言的情感分类模型成本也太大。另一方面,把匹配条件下的情感识别技术直接应用于跨语料库时,由于训练和测试集数据分布上的差异,会导致跨库情感识别性能很差,甚至无法正确识别。因此,建立一个泛化能力强的跨语料库语音情感识别系统的关键是解决训练语料库和测试语料库不匹配问题。

不同语言的同一情感语音特征存在语种自身的特性,也存在情感共性。提高跨语料库情感识别性能的一个研究思路是寻找不同语料库情感之间的共性特征,采用共性特征训练分类模型来减小训练和测试数据的域偏移。文献[1]提出了一种基于主成分分析特征排序的语料库相似性度量方法,通过计算相似性分数对不同数据集相似度进行排序,找到具有相似特征的最佳数据集组合来提升跨语料库的情感识别性能。最相似的数据集组合在情感识别性能上确实优于其他数据集组合,但是找寻出所有与情感高度相关特征相当耗时耗力,且所选特征的有效性在很大程度上依赖于识别模型,该方法通用性较低[2]。另一个研究思路是通过减小不同语料库语音数据的情感特性之间的差异来提高识别性能。Liu等[3]提出一种深域自适应卷积神经网络模型来减小不同领域之间的差异并提取鲁棒特征,以弥补情感鸿沟。Song等[4]利用基于非负矩阵分解和迁移学习技术的转移非负矩阵分解方法,获取标记源数据集和未标记目标数据集通用的鲁棒特征。薛艳飞等[5]提出一种基于对抗训练的跨语料库语音情感识别方法,通过语料库之间的对抗训练来缩小不同语料库之间的差异,并通过引入多头自注意力机制增强序列中情感显著特征的提取能力,该方法具有较好的情感识别能力。

本文通过寻找不同域情感特征之间的非线性映射关系,提出一种基于深度域适应神经网络的跨语料库语音情感识别方法,该方法通过关联源域和目标域的语音情感特征,使得目标域情感特征有更好的域适应能力。该深度域适应网络相当于一台翻译器,可以将一个域的情感特征翻译成类似于另一个域的特征分布,使得不匹配条件下的语音情感识别接近于匹配条件下的情感识别,从而提高分类性能。与使用域不变特征表示方法不同,本文使用的不是固定特征,而是专注于域自适应和深度特征学习来寻找域之间的映射关系,将域自适应嵌入到深度学习中,在源域和目标域的局部子空间中学习不同域之间的可迁移特征和不变性特征,使得源域语料库训练好的分类网络可以更好地泛化于特征对齐后的目标域数据,使得跨语料库识别受域偏移阻碍降低。

源域和目标域之间的相似性是迁移学习的前提,可以通过距离来度量两个域间的相似度。最大均值差异(Maximum mean discrepancy, MMD)将特征映射到高维希尔伯特空间来度量域间的相似度,在迁移学习域适应中被广泛应用[6-7]。另外,在深度学习的回归问题中常用均方误差(Mean square error, MSE)来约束预测数据和实际数据的距离。在本文的深度域适应神经网络中,联合使用这两种测度作为损失函数,在两个维度同时约束源域和目标域特征之间的距离。利用这种联合约束使得网络在训练

过程学习更好的可迁移特征和不变特征,训练得到一个拟合目标域特征到源域特征的深度域适应网络。

除了特征外,分类器也是语音情感识别中的重要模块,分类器的好坏直接影响最终的识别性能。常用的分类模型有:K最近邻模型[8]、高斯混合模型(Gaussian mixture model, GMM)[9]、决策树模型[10-12]和支持向量机模型(Support vector machine, SVM)[13]等,单一分类模型在分类任务中的性能有一定的局限性。由于神经网络强大的特征提取能力以及在庞大数据集上的高维度数据处理能力,研究人员开始将神经网络和传统分类方法结合应用于语音情感识别,并证明组合分类优于单分类效果。在训练和测试数据来自同一个语料库的情感识别任务中,Shahin等[14]将深度神经网络(Deep neural network, DNN)和GMM联合到一起构建组合分类器,相较于使用SVM分类器,在理想环境下识别率提升了4.6%,在噪声环境下识别率提升了5.9%。作者团队[13]提出一种新的多分类器联合决策算法,最终识别结果由支持向量机根据该算法进行联合决策得到。Yao等[15]通过对DNN、卷积神经网络(Convolutional neural network, CNN)与递归神经网络的分类结果做联合决策,其性能优于单个分类器分类。在跨语料库语音情感识别任务中,Albornoz等[16]用不同特征和分类器集为每一种语言建模,测试时使用已知语言训练好的分类器计算出的决策级融合结果作为未知语言的最终决策,这种模块化设计提高了跨语料库识别精度,但并没有从根本上解决不同语料库本身的差异对跨语料库识别性能的影响。

基于决策树的分类方法已成为监督学习中最有效的分类方法之一。Wang等[10]基于类间可分性度量,提出一种改进的基于决策树的支持向量机算法,实验验证了该方法的有效性。作者团队[11]提出了一种基于DNN决策树-SVM模型的语音情感识别方法,该方法的平均情感识别率比传统SVM和DNN-SVM分类方法分别高6.25%和2.91%。赵涓涓等[12]提出了一种基于决策树和改进SVM混合模型的语音情感识别方法,与传统SVM和人工神经网络方法相比,其具体更高的抗噪声能力和稳定性。在跨语料库背景下,多种情感的识别任务相对匹配条件下的识别任务情感间的混淆度更大,导致系统的整体识别性能不理想。本文构建一种基于CNN决策树多级分类模型的跨语料库语音情感方法,根据情感混淆度矩阵来划分树结构,不限制树节点可以拥有的分支数量,构建以CNN为树节点的多级CNN分类模型,对情感类别进行最终判决。该方法利用决策树对多种情绪进行不同层次分类来减少跨语料库情绪间混淆,并利用深度神经网络的强分类能力在跨语料库语音情感识别场景下获得较好的识别效果。

1 跨语料库情感识别

本文提出的基于深度域适应和CNN决策树模型的跨语料库语音情感识别框架如图1所示。主要由两大模块组成,第1个模块是特征对齐部分,包括预处理,特征提取和深度域适应,该模块主要是通过深度域适应网络在局部子空间进行跨库统计特征迁移学习,从而实现源域和目标域情感的特征对齐。第2个模块是情感分类部分,基于CNN决策树多级分类模型将多种情感类别进行分级处理,逐层实现分类。

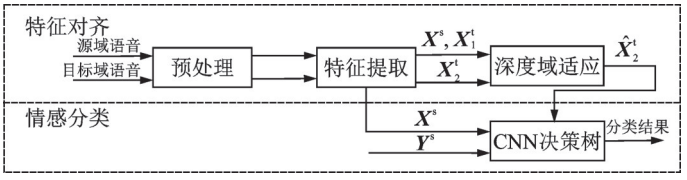

图1 基于深度域适应和CNN决策树模型的跨语料库语音情感识别框架图

Fig.1 Frame diagram of cross-corpus speech emotion recognition based on deep domain adaptation and CNN decision tree model

1.1 局部子空间跨库统计特征迁移学习

1.1.1 深度域适应神经网络

域适应是一种约束条件严格的迁移学习,可以通过减少数据集偏差能够使源域数据集训练的模型在目标域数据集上表现良好。本文提出一种基于深度域适应神经网络(Deep domain adaptive neural network, DDANN)的局部子空间跨语料库特征迁移方法,利用联合约束下的域自适应神经网络实现跨库统计特征对齐。该方法首先按照情感类别将源域和目标域划分为多个子空间,然后在每个特定类别的子空间进行局部特征迁移学习,实现源域和目标域特征的局部对齐,直到每种情感类别之间都建立不同领域特征之间的映射关系。最后,测试阶段的目标域数据采用DDANN对齐后,输入到源域训练好的分类器来实现分类。

传统的机器学习算法由于假设训练和测试数据是相同分布,直接应用在跨语料库情感识别任务效果很差。本文通过深度域适应神经网络来解决训练和测试数据分布不一致的问题,在训练阶段通过域适应挖掘不同语料库之间的深度可迁移特征在域特征和之间建立联系,从而训练得到具有把目标域特征拟合成类似源域的特征的DDANN,以提高分类模型的域适应能力,进而提高跨语料库语音情感分类性能。有大量标记样本的语料库称为源域,记为$D^{s}=\{\boldsymbol{X}^{s},\boldsymbol{Y}^{s}\}$;另一种用于跨库识别的语料库称为目标域,记为$D^{t}=\{\boldsymbol{X}^{t},\boldsymbol{Y}^{t}\}$。其中$\boldsymbol{X}^{s}=\{x_i^{s}\}_{i=1}^{n}$表示源域的特征集,$\boldsymbol{Y}^{s}=\{y_i^{s}\}_{i=1}^{n}$为源域情感特征对应类别标签集;$\boldsymbol{X}^{t}=\{\boldsymbol{X}_1^{t},\boldsymbol{X}_2^{t}\}$表示目标域的特征集,其中$\boldsymbol{X}_1^{t}=\{x_j^{t}\}_{j=1}^{m}$为用于训练的目标域数据集情感特征,对应类别标签集为$\boldsymbol{Y}_1^{t}=\{y_j^{t}\}_{j=1}^{m}$;$\boldsymbol{X}_2^{t}=\{x_j^{t}\}_{j=m+1}^{r}$为用于测试的目标域数据集情感特征,对应类别标签集为$\boldsymbol{Y}_2^{t}=\{y_j^{t}\}_{j=m+1}^{r}$(该部分标签只用来计算正确率),$n$和$m$分别为训练阶段源域和目标域的样本总数量。DDANN框架如图2所示,全连接(Full connection, FC)层FC_1和FC_2用来将目标域特征$\boldsymbol{X}_1^{t}$映射成$\hat{\boldsymbol{X}}_1^{t}$,最后一个全连接层作为域适应层(Domain adaptation, DA),在该层添加损失函数$Loss_{sum}$来计算$\hat{\boldsymbol{X}}_1^{t}$和$\boldsymbol{X}^{s}$间距离,在训练过程中通过最小化$Loss_{sum}$约束训练网络;测试阶段把目标域特征$\boldsymbol{X}_2^{t}$映射成类似源域特征$\hat{\boldsymbol{X}}_2^{t}$,用于跨语料库情感识别。

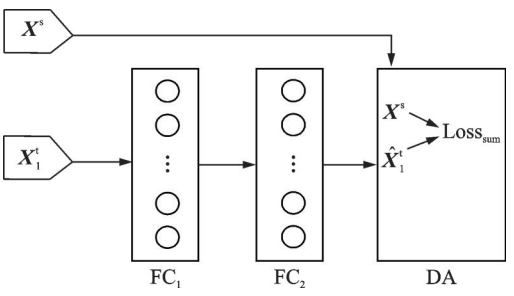

图2 DDANN框架

Fig.2 DDANN framework

具有强泛化能力的深度神经网络不仅可以学习到更强大、更有效的辨别性特征,而且能够有效地模拟非线性映射,更好地进行特征的迁移学习。本文所提出的DDANN模块就是利用DNN的逼近特性,使用DNN的多层网络结构实现不同语种语音情感特征之间的域迁移。DDANN的训练包括两个过程,在前向传播阶段,每一层通过激活函数学习一组非线性映射关系,得到最终的输出预测值。反向传播阶段是一个逐层域适应的过程,通过源域与目标域之间的距离来限制目标函数,使得神经网络能够学习到更多具有最小域偏移的“良好”特征表示。为了得到拟合度更高的神经网络模型,在自适应层对参数进行精细调整时,本文的损失函数联合考虑子空间跨库统计特征在特征空间和希尔伯特空间中的差异。在联合约束损失函数指导下,使用带动量随机梯度下降算法作为优化器,在不断更新每层网络的权重和偏置的过程中训练网络,使得迁移模型收敛速度更快、精度更高。

1.1.2 特征空间和希尔伯特空间联合约束

在局部特征迁移学习过程中,域适应层的损失函数直接影响DDANN训练的好坏。本文提出在低

维特征空间和高维希尔伯特空间联合约束目标域和源域特征距离，把该约束关系作为DDANN的损失函数来最小化域子空间特征的分布差异。该特征空间和希尔伯特空间联合约束损失函数$\mathrm{Loss}_{\mathrm{sum}}$如式(1)所示，均方误差$\mathrm{Loss}_{\mathrm{e}}$如式(2)所示，最大均值差异$\mathrm{Loss}_{\mathrm{d}}$如式(3)所示。

$$\mathrm{Loss}_{\mathrm{sum}}=\mathrm{Loss}_{\mathrm{e}}+\lambda\mathrm{Loss}_{\mathrm{d}} \tag{1}$$

$$\mathrm{Loss}_{\mathrm{e}}=\frac{1}{2N}\sum_{i=1}^{N}\left\| x_i^{\mathrm{s}}-\hat{x}_i^{\mathrm{t}}\right\|_2^2 \tag{2}$$

$$\mathrm{Loss}_{\mathrm{d}}=\mathrm{MMD}\left(\boldsymbol{X}^{\mathrm{s}},\hat{\boldsymbol{X}}_1^{\mathrm{t}}\right) \tag{3}$$

式中：N表示语音总条数，$\hat{x}_i^{\mathrm{t}}$表示输入为第i条目标域语音的情感特征时模型的预测输出；x_i^{s}表示该种情感类别对应的源域语音情感特征，两者之间的权重为超参数λ。

传统损失函数一般只在特征空间使用MSE来度量域间距离，为了实现更准确的域特征对齐，本文联合考虑了高维希尔伯特空间的域特征间距离。通过高维映射函数计算不同分布的两个语料库样本在再生希尔伯特空间(Reproducting kernel Hilbert space, RKHS)中期望之差的上界值，以此来衡量两个语料库分布在希尔伯特空间的距离

$$\mathrm{MMD}(F,P,Q)=\sup_{f\in\mathrm{H},\left\|f\right\|_{\mathrm{H}}\leqslant 1}\left(E\left[f\left(\boldsymbol{X}^{\mathrm{s}}\right)\right]-E\left[f\left(\hat{\boldsymbol{X}}_1^{\mathrm{t}}\right)\right]\right) \tag{4}$$

式中：源域语音特征集$\boldsymbol{X}^{\mathrm{s}}$分布为$P$；样本数量为$n$；目标域语音特征集的估值$\hat{\boldsymbol{X}}_1^{\mathrm{t}}$分布为$Q$；样本数量为$m$。$H$表示RKHS，$f(\cdot):X\rightarrow H$表示原始特征空间映射到RKHS的映射函数，Sup表示上界(最大值)。

由于一般无法直接计算源域和目标域的总体均值，利用核均值嵌入来近似计算，即

$$E\left[f(x)\right]=\left\langle\mu,f\right\rangle_{\mathrm{H}} \tag{5}$$

则式(4)改写为

$$\mathrm{MMD}(F,P,Q)=\sup_{f\in\mathrm{H},\left\|f\right\|_{\mathrm{H}}\leqslant 1}\left(\left\langle\mu_P-\mu_Q,f\right\rangle_{\mathrm{H}}\right) \tag{6}$$

再利用内积性质，即

$$\left\langle\mu_{\mathrm{P}}-\mu_Q,f\right\rangle_{\mathrm{H}}\leqslant\left\|\mu_P-\mu_Q\right\|_{\mathrm{H}}\left\|f\right\|_{\mathrm{H}} \tag{7}$$

在约束条件$\left\|f\right\|_{\mathrm{H}}\leqslant 1$的情况下，将式(6)进一步表示为

$$\mathrm{MMD}(F,P,Q)=\left\|\mu_{\mathrm{P}}-\mu_Q\right\|_{\mathrm{H}} \tag{8}$$

式中

$$\mu=E\left[f(x)\right]=\frac{1}{n}\sum_{i=1}^{n}f(x_i) \tag{9}$$

因此，MMD可以表示为源域和目标域的总体均值之差，用来度量这两个域的分布差异，有

$$\mathrm{MMD}\left(F,\boldsymbol{X}^{\mathrm{s}},\hat{\boldsymbol{X}}_1^{\mathrm{t}}\right)=\left\|\frac{1}{n}\sum_{i=1}^{n}f\left(x_i^{\mathrm{s}}\right)-\frac{1}{m}\sum_{j=1}^{m}f\left(\hat{x}_j^{\mathrm{t}}\right)\right\|_{\mathrm{H}} \tag{10}$$

为了省去高维空间繁复计算，Pan等[17]提出可以将MMD问题转化为一个核学习问题，对MMD进行平方化简得到内积，用核函数可以表示为

$$\mathrm{MMD}^2\left[F,\boldsymbol{X}^{\mathrm{s}},\hat{\boldsymbol{X}}_1^{\mathrm{t}}\right]=\left\|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n}k\left(x_i^{\mathrm{s}},x_i^{\mathrm{s}'}\right)-\frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}k\left(x_i^{\mathrm{s}},\hat{x}_j^{\mathrm{t}}\right)+\frac{1}{m^2}\sum_{j=1}^{m}\sum_{j=1}^{m}k\left(\hat{x}_j^{\mathrm{t}},\hat{x}_j^{\mathrm{t}'}\right)\right\|_{\mathrm{H}}^2=\mathrm{tr}(\boldsymbol{KM}) \tag{11}$$

式中:$\boldsymbol{K}=\begin{bmatrix}k_{ss} & k_{st}\\ k_{ts} & k_{tt}\end{bmatrix}$,$\boldsymbol{K}$为维度为$(n+m)\times(n+m)$的核矩阵,$k_{ss}$,$k_{tt}$和$k_{st}$,$k_{ts}$分别为源域、目标域和混合域的核函数;$\boldsymbol{M}=\begin{bmatrix}\frac{1}{n^2} & -\frac{1}{nm}\\ -\frac{1}{nm} & \frac{1}{m^2}\end{bmatrix}$,$\boldsymbol{M}$为系数矩阵。当$x_i^s$、$\hat{x}_j^t\in D_s$时,$\boldsymbol{M}$的第$i$行第$j$列的元素$M_{ij}=\frac{1}{n}$;当$x_i^s$、$\hat{x}_j^t\in D_t$时,$M_{ij}=\frac{1}{m}$,其他情况为$M_{ij}=\frac{1}{nm}$。核函数主要有线性核、多项式核、高斯核、Sigmoid核以及Laplacian核等。高斯核函数也叫做径向基函数(Radial based function, RBF),可以拟合任意分布到无限维空间。本文选择的核函数是高斯核函数。当MMD中使用多个内核函数时称为多核最大均值差异(Multiple kernel maximum mean discrepancy, MK-MMD),在实际应用中比单内核MMD能获得更好性能。实验中通过联合多个不同RBF的高斯核函数,即每个核函数使用不同的超参数σ,来计算源域和目标域数据集特征之间的距离。

1.2 CNN决策树多级分类模型

1.2.1 CNN分类模型

卷积神经网络是一种分层的神经网络,是解决多分类问题的常用方法。在语音情感识别领域,大多数研究人员使用2D CNN模型进行特征提取和情感分类[18],为了匹配二维CNN的输入,需要将语音信号转换为语谱图。但是原始语音信号具有丰富的情感信息,把一维语音信号转换为二维语谱图,会丢失一些有用的语音情感特征信息。本文使用更适合处理语音时间序列数据的一维卷积来对原始语音进行特征提取和分类,所使用的一维CNN网络如图3所示,包括3个堆叠的卷积层和最大池化层,1个Flatten层,2个全连接层,2个Dropout层和用于判决的Softmax层。其中卷积层和池化层是卷积网络的主要组成部分,用于挖掘语音的情感细节。

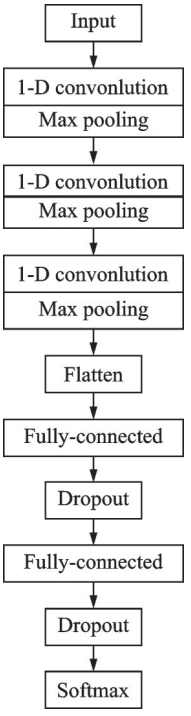

图3 CNN模型
Fig.3 CNN model

在卷积层利用卷积核对输入特征进行卷积运算,提取更深层次的情感信息,并增强对局部特征的学习。不同层间的连接通过卷积核操作,并且每个卷积层中卷积核的个数为多个,每个卷积核都采用局部连接方式连接到上层特征映射的局部区域。卷积层输出数据为上层所有特征映射并添加偏差激活后生成的映射特征,每个卷积核的特征映射可以表示为

$$z(n)=x(n)\otimes w(n)=\sum_{m=0}^{n}x(m)w(n-m) \tag{12}$$

$$z_i^{\,l}=\sigma\left(\sum_j z_j^{\,l-1}w_{ij}^{\,l}+b_i^{\,l}\right) \tag{13}$$

式中:1 D卷积层的输入信号为$x(n)$;卷积滤波器为$w(n)$;卷积结果为$z(n)$。$z_i^{\,l}$表示第l层第i个输入特征的输出映射特征,$z_j^{\,l-1}$表示第$(l-1)$层的第j个输入特征,$w_{ij}^{\,l}$表示第i个和第j个特征之间的卷积滤波器,$b_i^{\,l}$为偏置,$\sigma(\cdot)$为激活函数,通常为relu或tanh函数。

池化的目的是通过层间算法运算将输入数据抽象为高级特征表示,本文使用最大池化来计算局部特征映射的最大值,通过消除冗余和失真来降低输入特征的维度,以生成全局特征表示。这样做不仅可以从学习到的特征中提取最显著的情感线索,并且减少了大量的网络参数,在一定程度上缓解了过拟合。

通过一系列的卷积,最大池化和全连接层构成的特征提取器来挖掘深层次特征,捕获情感特征和标签的依赖关系,以抽象形式对输入数据进行建模。最后,在从多个卷积和池化层传递输入并提取高层次特征后。将二维数据扁平化到阵列并将其馈送到作为原始神经网络的前馈网络,将学习到的序列信息通过Softmax激活函数从最后一个全连接层传递,以产生不同情绪的概率。在语音情感分类任务中,使用多分类交叉熵损失函数来度量预测概率和真实概率分布间的距离,即有

$$\mathrm{Loss_c}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}P\left(x_i^c\right)\log\hat{P}\left(x_i^c\right) \tag{14}$$

式中:N表示样本的数量;C表示情感类别数量;x_i^c表示第i个样本,真实类别为c;$P\left(x_i^c\right)$表示当前样本的真实概率分布;$\hat{P}\left(x_i^c\right)$表示当前样本通过Softmax输出的预测概率。此外,为了避免过拟合,在前两个全连接层之后还设置了系数为0.2的Dropout层。

1.2.2 决策树多级分类模型

在语音情感识别任务中,可以用情感混淆度来量化情感之间的差异[11],其数值越高,情感越相似越难分辨。实验发现,在跨语料库背景下进行情感分类比单一语料库下的情感混淆度更大,系统整体的识别率更低。在多分类问题中,决策树可以在相对较短的时间内对给定数据进行有效地划分,并且CNN被证明在处理分类任务时表现良好。本文将CNN和决策树相结合,构造一种以CNN为树节点的多级分类模型,该模型以情感混淆度作为情感分级分类的划分准则,将分类问题进行多层次分解,由粗到细对语音情感逐层分类。该模型在降低情感间混淆度的同时可以挖掘语音信号的易区分情感特征,有助于提高跨语料库语音情感识别系统性能。

本文决策树算法基本思想是混淆度小、易区分的情感在树的根节点分类;混淆度大、难区分的情感在树的叶节点分类。将情感混淆度作为类可分性测度,除叶节点外的其他每个节点都利用一个CNN模型进行分类。本文决策树构造算法首先确定决策树顶部节点的分类集群,将情感混淆度较大的类别视为同一个集群,较小的类别视为另一集群。以同样的方式生成下一级类别集群,一直迭代到叶节点,决策树构造完成。该算法的停止准则有两种形式:节点中情感只包含一个类;或者是节点中的情感间混淆度低于阈值。具体如下。

算法1:决策树构造算法

输入:情感类型的集合$E=\{e_1,e_2,\cdots,e_n\}$

输出:分割后的情感类型组合

步骤1 源域语料库作为训练数据,目标域语料库作为测试数据,计算情感混淆矩阵,根据混淆矩阵计算出不匹配条件下的情感混淆度。

步骤2 初次分类时设置阈值为$T\%$,将混淆度超过阈值T的情感分为一类,小于T的单独归为一类。在划分过程中,若不同组合情感类别没有重叠,则单独分组即可;若与其他组合情感有所重复,则将重复组并为一组。即若$C_{a,b}>T$,$C_{c,d}>T$,则将a和b分为一组,c和d分为一组;若$C_{a,b}>T$,$C_{b,c}>T$,则将a,b,c分为一组;若$C_{a,e_i}<T$,e_i为该节点处除a外的其他情感,则将a单独划分为一组。

步骤3 对于未分组的情感类别,计算与其他情感类别之间的混淆度转至步骤2,将其分入已有组或者单独成组。

步骤4 计算各组中情感类别的个数,如果个数大于2,则将阈值T增加至$2T$,并转至步骤1;否则,结束分组。所有情感都完成分组,结束。

用CASIA语料库训练模型,用EMO-DB语料库进行跨库测试来构建决策树,首先计算得到的5种

情感之间的混淆度矩阵如表1所示。根据决策树构造算法中训练阶段2的步骤2决策树构造流程,将初始阈值T设置为20%,由表1中5种情感混淆度矩阵可知害怕和伤心混淆度,害怕和平静的混淆度,快乐和生气的混淆度均大于初始阈值20%,根据决策树构造算法,将害怕、伤心、平静归为一大类,快乐和生气归为一大类,训练CNN1进行分类。由于害怕、伤心、平静的情感类别数大于2,后续细分这3类情感时阈值增加至40%,由表2的3类情感混淆度矩阵可知,害怕与伤心、平静之间的混淆度都小于40%,因此构建CNN2将3类情感单独分组即可。由于生气和快乐情感类别数等于2,后续可直接构建CNN3对这2类情感进行分类,情感分类之后不必再次计算这2类情感间的混淆度。至此,决策树模型构建完毕,如图4所示。

表1 5类情感的混淆度矩阵

Table 1 Confusion matrix of five types of emotions %

情感类别	伤心	生气	快乐	平静	害怕
伤心					
生气	0				
快乐	3.6	27.38			
平静	10	4.76	5		
害怕	22.6	12.7	5.5	20.8	

表2 3类情感的混淆度矩阵

Table 2 Confusion matrix of three types of emotions %

情感类别	害怕	伤心
伤心	27.2	
平静	8.8	7.1

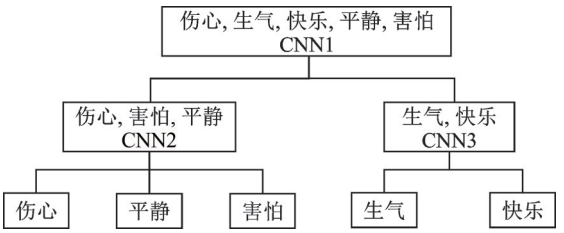

图4 CNN决策树模型

Fig.4 CNN decision tree model

根据划分好的情感类别组合训练针对不同分类目标的CNN模型,在测试阶段,利用训练好的决策树得到最终的分类结果。

1.3 跨语料库语音情感识别

跨语料库语音情感识别的本质是利用源域语音学习一个情感分类模型来预测目标域语音的情感类别标签如下。本文提出的跨语料库语音情感识别方法,在训练过程中的第1阶段训练深度域适应模型DDANN得到目标域和源域语料库情感特征间的映射关系;第2阶段是针对源域语料库训练对应的CNN决策树分类模型。测试阶段将目标域情感特征经过DDANN对齐后再利用CNN决策树进行分类,实现最终的跨语料库语音情感识别任务。该方法详细过程如下。

预处理:按照情感类别将源域和目标域分为多个子空间

训练阶段1:

输入:源域语音数据集$D^{s}=\{\boldsymbol{X}^{s},\boldsymbol{Y}^{s}\}$和目标域语音数据集$D^{t}=\{\boldsymbol{X}_1^{t},\boldsymbol{Y}_1^{t}\}$

输出:训练好的DDANN模型

步骤1 源域语音的情感特征作为DDANN的输入,目标域语音情感特征为DDANN的训练目标。

步骤2 前向传播阶段,源域和目标域进行局部子空间特征的迁移学习,在每个局部子空间拟合一组$\boldsymbol{X}_1^{\mathrm{t}}$到$\boldsymbol{X}^{\mathrm{s}}$的非线性映射函数关系。

步骤3 反向传播过程中,在域适应层通过MSE和MK-MMD联合约束作为损失函数来度量子空间中的域间特征差异,通过优化器寻优得到最小损失值,迭代更新网络参数。

训练阶段2:

输入:源域语音数据集$D^{\mathrm{s}}=\left\{\boldsymbol{X}^{\mathrm{s}},\boldsymbol{Y}^{\mathrm{s}}\right\}$

输出:训练好的CNN决策树多级分类模型

步骤1 树节点1-D CNN模型的搭建。

步骤2 根据算法1决策树构造流程划分情感类型组合。

步骤3 源域语音特征$\boldsymbol{X}^{\mathrm{s}}$和对应情感标签$\boldsymbol{Y}^{\mathrm{s}}$作为训练数据,根据分割好的情感类别组合分别训练对应的1-D CNN分类模型。

测试阶段:

输入:目标域语音数据集$D^{\mathrm{t}}=\left\{\boldsymbol{X}_2^{\mathrm{t}}\right\}$

输出:语音情感识别率

步骤1 目标域语音特征$\boldsymbol{X}_2^{\mathrm{t}}$作为训练好的DDANN模型的输入,输出得到在局部子空间对齐后的特征数据$\hat{\boldsymbol{X}}_2^{\mathrm{t}}$。

步骤2 对齐后的目标域数据作为训练好的CNN决策树模型的输入,在源域语音训练好的分类网络中测试识别性能。

至此,完成跨语料库语音情感识别。

2 实验和结果

为了评估所提出的跨语料库语音情感识别方法的有效性,本文在CASIA,EMO-DB和RAVDESS 3个公共语料库上两两组合进行了6组实验,并对实验结果进行综合分析。

2.1 数据集描述

德国柏林语音情感数据库EMO-DB是由柏林工业大学录制[19]。10位演员以7种情感对10条语句进行语音录制,共计535条情感语音,其中7种情感分别为快乐、害怕、伤心、生气、平静、无聊和厌恶。中科院语音情感数据库CASIA是由中国科学院自动化研究所录制[20]。4位演员以6种情感倾向在纯净无噪的环境下进行录制,共计1 200条情感语音,所包含的6种情感分别为:快乐、生气、害怕、平静、伤心和惊讶。Ryerson情感言语和歌曲视听数据库RAVDESS[21]包括语音文件,歌曲文件和视频文件。其中,语音文件是由来自加拿大的24名专业演员以北美英语说出两个词汇相同的语句在专业录音室中单独录制,共计1 440条情感语音,包括中性、平静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶8种情感类别,除中性外,每种情感都是在两种情绪强度(正常、强烈)下产生的。

2.2 数据处理

(1) 情感类别选择。为了更好地进行对比实验,本文从3个语料库中选择重叠的情感类别(伤心,生气,快乐,平静,害怕)语音作为实验数据。(2) 在进行特征提取前对语音信号进行预处理。主要包含:预加重、分帧加窗和端点检测(Voice activity detection, VAD)。本文把每帧时间设置为25 ms,由于语音数据库的采样频率是16 kHz,所以在分帧时把帧长设置为512,帧移设置为帧长的1/4。(3) 特征提取、提取充分考虑人耳听觉特性的MFCC特征,在提取的帧级特征上做统计(平均值、最大值,最小值、均值、方差)得到包含更多跨语料库同种情感共性的句级特征来表征情感。(4) 由于统计后求得的各指

标数据数量级水平相差很大,为了保证结果的可靠性,基于原始数据的均值和标准差对数据进行Z-score标准化,标准化处理后的统计特征作为系统的输入特征。

2.3 结果和分析

实验是在Python 3.7的环境下进行的,实验所用到的深度学习框架主要是Keras。实验中训练样本和测试样本的数量比例为8:2。在本文提出的DDANN中,损失函数在特征空间和希尔伯特空间进行联合约束,其中超参数λ是这两个维度进行约束的权重,取值在0.1~0.9之间。超参数的选择直接影响模型的拟合效果进而影响最终分类性能,本文实验λ取0.8时,平均语音情感识别率最高。

图5中给出了多种跨语种组合下每类情感的识别率,其中,C、B和R分别代表CASIA、EMO-DB和RAVDESS语料库。C/B代表C是训练语料库,B是测试语料库;R/B表示R是训练语料库,B是测试语料库;B/C表示B是训练语料库,C是测试语料库。从图5可以看出,本文提出的DDANN+CNN方法与直接使用CNN分类方法相比,大部分情感的识别率都有明显提升。以图5(a)的C/B为例,相比直接使用CNN分类,本文提出的DDANN+CNN方法在伤心、生气、平静这3种情感的识别率分别提升了35.94%、66.06%和25.81%,提升非常大,这说明DDANN可以有效地将目标域B的情感特征映射成源

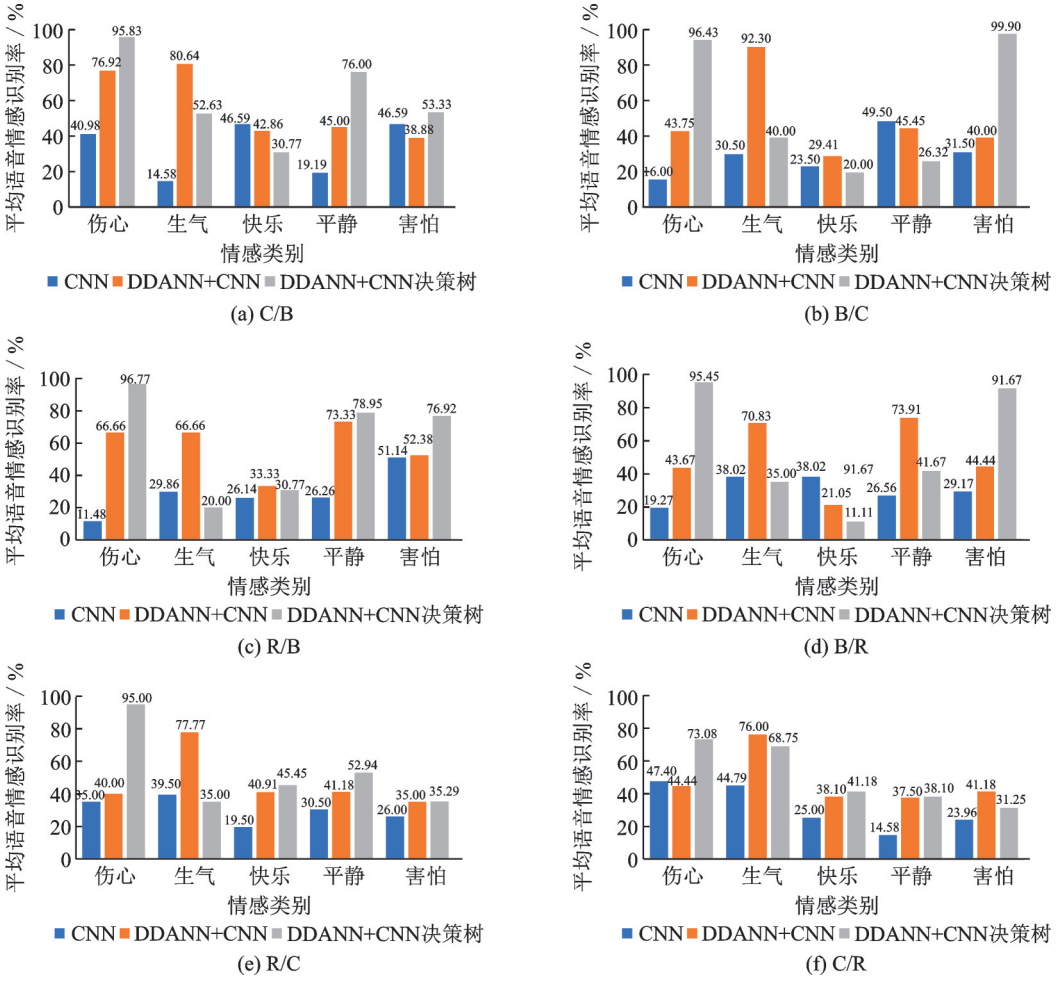

图5 跨语料库情感识别率

Fig.5 Cross-corpus emotion recognition rate

域C的情感特征，使得跨库条件下的情感识别接近匹配条件下的情感识别，从而提高了跨库情感识别性能。另外，比较6个跨语料库组合的情感识别率，生气情感的识别率是使用DDANN+CNN方法后所有情感中提升最多的，其中在C和B语料库的跨库组合情况下提升最多，C/B和B/C分别提升了66.06%和61.8%，说明本文提出的DDANN方法对这两种语言的生气情感特征迁移学习效果特别好，迁移后的目标域生气情感特征最接近源域的生气情感特征分布，使得该情感的识别率提升最多。

从图5也可以看出，DDANN+CNN决策树方法与DDANN+CNN方法相比，进一步提升了情感分类性能。以图5(a)的C/B为例，相比DDANN+CNN分类，使用DDANN+CNN决策树分类时伤心、平静、害怕的情感识别率都得到了提高，这说明CNN决策树多级分类方法可以进一步提升部分易混淆情感的分类能力。从6个跨语料库组合的情感识别率可以看出，使用DDANN+CNN决策树的整体分类性能优于DDANN+CNN。然而，快乐和生气有所下降，这可能是因为快乐和生气情感的语音之间存在一部分区分度小、相似的音频特征，并且在决策树模型中下一级分类的准确率会受到上一级分类准确率的影响，可能出现误差累积的情况。对这两种情感分类时，可以通过选择最佳特征组，构造条件更苛刻的情感分类器来进一步提高识别系统的性能。

表3总结了6个跨语料库组合下不同方法的平均情感识别率。结果表明，直接用CNN进行跨语料库语音情感识别时，识别率均为30%左右，分类性能很差。本文提出的DDANN+CNN方法与直接用CNN分类方法相比，在6个跨数据库组合下大部分情感的识别率都得到明显的提升，最终每个跨库组合下的平均识别率都提高了23.62%左右。其中，C/B情况下，平均识别率提升最多，为28.74%；C/R情况下，平均识别率提升最少，但也提升了17.83%。这主要是因为采用深度域适应神经网络将两个域的统计特征在局部子空间对齐，可以有效地提升情感特征的领域适应能力，使得大部分情感识别率都有明显的提高。另外，DDANN+CNN决策树方法的最终的识别率比DDANN+CNN方法提高了2.19%左右，表明CNN决策树模型将情感类别逐层划分后再分类的方式在跨语料库情感识别中可以进一步提升多语种情感分类性能。

表3 不同方法的跨语料库情感识别率

Table 3 Cross-corpus emotion recognition rates of different methods %

方法	源域/目标域					
	C/B	B/C	R/B	B/R	R/C	C/R
CNN	30.63	30.19	30.00	30.20	30.10	31.15
文献[22]	53.39	49.37				
文献[23]			44.60	34.50		
文献[24]			46.00			
DDANN+CNN	59.37	54.16	58.33	53.13	50.00	48.98
DDANN+CNN决策树	61.71	56.55	60.68	54.97	52.74	50.47

为了进一步验证本文提出的DDANN+CNN决策树方法在跨语料库语音情感识别中的有效性，本文还与现有的优秀的跨语料库语音情感识别方法进行了比较。文献[22]以B作为基础语料库，C作为迁移语料库，使用SVM分类器进行四分类得到的识别率为49.37%，而本文方法识别率为56.55%；以C为基础语料库，B为迁移语料库时，识别率为53.39%，而本文方法识别率为61.71%，本文方法在这两个跨语料库组合下的分类性能都明显优于该文献方法。文献[23]选择在R和B语料库训练得到的性能最好的网络进行跨语料库情感识别实验，本文方法明显优于该方法。文献[24]使用R创建了基于卷积神

经网络的基线语音情感识别模型,并在B进行测试,得到了46%的识别率,本文方法比其高14.68%。可以看出,本文提出的DDANN+CNN决策树方法在跨语料库情感识别任务中性能较好。

3 结束语

为了有效地提高跨语料库语音情感的识别性能,本文从特征对齐和分类模型的构建两个方面切入,提出了一种基于深度域适应和CNN决策树模型的跨语料库语音情感识别方法。在该方法中,本文首先在特征空间和希尔伯特空间联合约束深度域适应神经网络实现局部子空间的域特征对齐,然后将CNN和决策树结合,构建了一种以CNN作为树节点的分类模型。最后,将该模型用于语音情感识别。实验结果表明,相比直接跨语料库识别,该方法的系统性能得到很大提升。与其他优秀跨预料库情感识别方法相比,该系统不仅在识别率方面表现良好,而且操作简单,泛化能力强,可以使用现有语种数据集训练的模型对多语言情感识别进行推广,使得跨语料库语音情感识别适用于不同的应用场景与声学条件。

参考文献:

[1] SIEGERT I, BOCK R, WENDEMUTH A. Using a PCA-based dataset similarity measure to improve cross-corpus emotion recognition[J]. Computer Speech and Language, 2018, 51: 1-23.

[2] 吕惠炼,胡维平. 基于端到端深度神经网络的语音情感识别研究[J]. 广西师范大学学报(自然科学版), 2021, 39(3): 20-26.
LÜ Huilian, HU Weiping. Research on speech emotion recognition based on end-to-end deep neural network[J]. Journal of Guangxi Normal University (Natural Science Edition), 2021, 39(3): 20-26.

[3] LIU J T, ZHENG W M, ZONG Y, et al. Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network[J]. IEICE Transactions on Information and Systems, 2020, 103(2): 459-463.

[4] SONG P, ZHENG W M, QU S F, et al. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization[J]. Speech Communication, 2016, 83: 34-41.

[5] 薛艳飞,张建明. 基于对抗训练的跨语料库语音情感识别方法[J]. 微电子学与计算机, 2021, 38(3): 77-83.
XUE Yanfei, ZHANG Jianming. Cross-corpus speech emotion recognition based on adversarial training[J]. Microelectronics and Computer, 2021, 38(3): 77-83.

[6] SAITO K, WATANABE K, USHIKU Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, the United States: IEEE, 2018: 3723-3732.

[7] LIN W W, MAK M W, CHIEN J T. Multi-source I-vectors domain adaptation using maximum mean discrepancy based autoencoders[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2412-2422.

[8] CHEN P Z, ZHANG X, RU Y. Emotion recognition system based on enhancement of KNN algorithm[J]. Science Technology and Engineering, 2017, 17(19): 197-200.

[9] JIANG J B, WU Z Y, XU M X, et al. Comparing feature dimension reduction algorithms for GMM-SVM based speech emotion recognition[C]//Proceeding of Signal and Information Processing Association Summit and Conference. Kaohsiung, Taiwan: IEEE, 2013: 1-4.

[10] WANG X D, SHI Z H, WU C M, et al. An improved algorithm for decision-tree-based SVM[C]//Proceeding of World Congress on Intelligent Control and Automation. Dalian, China: IEEE, 2006: 4234-4238.

[11] SUN L H, ZOU B, FU S, et al. Speech emotion recognition based on DNN-decision tree SVM model[J]. Speech Communication, 2019, 115: 29-37.

[12] 赵涓涓,马瑞良,张小龙. 基于决策树和改进SVM混合模型的语音情感识别[J]. 北京理工大学学报, 2017, 37(4): 386-390.
ZHAO Juanjuan, MA Ruiliang, ZHANG Xiaolong. Speech emotion recognition based on decision tree and improved SVM mixed model[J]. Transactions of Beijing Institute of Technology, 2017, 37(4): 386-390.

[13] SUN L H, HUANG Y Q, LI Q, et al. Multi-classification speech emotion recognition based on two-stage bottleneck features

selection and MCJD algorithm[J]. Signal, Image and Video Processing, 2022, 16(5): 1253-1261.

[14] SHAHIN I, NASSIF A B, HAMSA S. Emotion recognition using hybrid Gaussian mixture model and deep neural network[J]. IEEE Access, 2019, 7: 26777-26787.

[15] YAO Z W, WANG Z H, LIU W H, et al. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN[J]. Speech Communication, 2020, 120: 11-19.

[16] ALBORNOZ E M, MILONE D H. Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles[J]. IEEE Transactions on Affective Computing, 2017, 8 (1): 43-53.

[17] PAN S J, KWOK J T, YANG Q. Transfer learning via dimensionality reduction[C]//Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. Chicago, United States: DPLP, 2008: 677-682.

[18] MAO Q R, DONG M, HUANG Z W, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. IEEE Transactions on Multimedia, 2014, 16 (8): 2203-2213.

[19] 金琴,陈师哲,李锡荣,等. 基于声学特征的语言情感识别[J]. 计算机科学, 2015, 42 (9): 24-28.
JIN Qin, CHEN Shizhe, LI Xirong, et al. Speech emotion recognition based on acoustic features[J]. Computer Science, 2015, 42 (9): 24-28.

[20] 李书玲,刘蓉,张鎏钦,等. 基于改进型SVM算法的语音情感识别[J]. 计算机应用, 2013, 33 (7): 1938-1941.
LI Shuling, LIU Rong, ZHANG Liuqin, et al. Speech emotion recognition algorithm based on modified SVM[J]. Journal of Computer Applications, 2013, 33 (7): 1938-1941.

[21] LIVINGSTONE S R, RUSSO F A. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English[J]. PloS one, 2018, 13 (5): e0196391.

[22] 钟琪,冯亚琴,王蔚. 跨语言语料库的语音情感识别对比研究[J]. 南京大学学报(自然科学版), 2019, 55 (9): 765-773.
ZHONG Qi, FENG Yaqin, WANG Wei. Comparison of speech emotion recognition in cross language corpus[J]. Journal of Nanjing University (Natural Science), 2019, 55 (9): 765-773.

[23] ALGHIFARI M F, GUNAWAN T S, HASHIM N N W N, et al. On the effect of feature compression on speech emotion recognition across multiple languages[C]//Proceedings of Recent Trends in Mechatronics Towards Industry 4.0: Selected Articles From iM3F 2020,Malaysia. Singapore: Springer, 2022: 703-713.

[24] BAKI P, ERDEN B O, NCU L S. A comparative study on different labelling schemes and cross-corpus experiments in speech emotion recognition[C]//Proceeding of Signal Processing and Communications Applications Conference (SIU). Istanbul, Turkey: IEEE, 2021:1-4.

作者简介:

孙林慧(1979-),女,**通信作者**,博士,副教授,研究方向:语音处理与现代语音通信、深度学习和稀疏表示, E-mail: sunlh@njupt.edu.cn。

赵敏(1998-),女,硕士研究生,研究方向:深度学习与语音情感识别, E-mail: 1220014002@njupt.edu.cn。

王舜(1998-),男,硕士研究生,研究方向:深度学习与语音情感识别, E-mail: 1220013714@njupt.edu.cn。

(编辑:刘彦东)