·人工智能与模式识别·　　文章编号：1000-3428(2022)04-0113-06　　文献标志码：A　　中图分类号：TP183

基于异构并行神经网络的语音情感识别

张会云[1,2]，黄鹤鸣[1,2]

（1.青海师范大学 计算机学院，西宁 810008；2.藏语智能信息处理及应用国家重点实验室，西宁 810008）

摘　要：提取能表征语音情感的特征并构建具有较强鲁棒性和泛化性的声学模型是语音情感识别系统的核心。面向语音情感识别构建基于注意力机制的异构并行卷积神经网络模型AHPCL，采用长短时记忆网络提取语音情感的时间序列特征，使用卷积操作提取语音空间谱特征，通过将时间信息和空间信息相结合共同表征语音情感，提高预测结果的准确率。利用注意力机制，根据不同时间序列特征对语音情感的贡献程度分配权重，实现从大量特征信息中选择出更能表征语音情感的时间序列。在CASIA、EMODB、SAVEE等3个语音情感数据库上提取音高、过零率、梅尔频率倒谱系数等低级描述符特征，并计算这些低级描述符特征的高级统计函数共得到219维的特征作为输入进行实验验证。结果表明，AHPCL模型在3个语音情感数据库上分别取得了86.02%、84.03%、64.06%的未加权平均召回率，相比LeNet、DNN-ELM和TSFFCNN基线模型具有更强的鲁棒性和泛化性。

关键词：语音情感识别；谱特征；韵律特征；注意力机制；异构并行分支；循环神经网络

开放科学（资源服务）标志码（OSID）：

中文引用格式：张会云，黄鹤鸣.基于异构并行神经网络的语音情感识别[J].计算机工程，2022，48(4)：113-118.

英文引用格式：ZHANG H Y，HUANG H M. Speech emotion recognition based on heterogeneous parallel neural network[J].Computer Engineering，2022，48(4)：113-118.

Speech Emotion Recognition Based on Heterogeneous Parallel Neural Network

ZHANG Huiyun[1,2]，HUANG Heming[1,2]

（1.Computer College，Qinghai Normal University，Xining 810008，China；
2.State Key Laboratory of Tibetan Intelligent Information Processing and Application，Xining 810008，China）

【Abstract】The core of a Speech Emotion Recognition（SER）system is to extract features that can best represent speech emotion and construct an acoustic model with strong robustness and generalization.In this study，a heterogeneous parallel Recurrent Neural Network（RNN）model based on the attention mechanism AHPCL is constructed for SER.The Long Short-Term Memory（LSTM）network is used to extract the time-series features of speech emotion，and the convolution operation is used to extract the speech spatial spectral features. By combining temporal and spatial information to jointly represent speech emotion，the accuracy of the prediction results is improved. The attention mechanism is used to assign weights according to the contribution of different time-series features to speech emotion to select a time sequence that better represents speech emotion from a large amount of feature information. Low-level descriptor features such as pitch，Zero Crossing Rate（ZCR），and Mel-Frequency Cepstrum Coefficient（MFCC）are extracted from three speech emotion databases，namely CASIA，EMODB，and SAVEE，and the high-level statistical functions of these low-level descriptor features are calculated to obtain 219 dimensional features. The experimental results show that the proposed model achieves 86.02%，84.03%，and 64.06% Unweighted Average Recall（UAR）on the CASIA，EMODB，and SAVEE databases，respectively.Compared with the LeNet，DNN-ELM，and TSFFCNN baseline models，the AHPCL model exhibits greater robustness and generalization.

【Key words】Speech Emotion Recognition（SER）；spectral feature；prosodic feature；attention mechanism；heterogeneous parallel branch；Recurrent Neural Network（RNN）

DOI：10. 19678/j. issn. 1000-3428. 0061076

0　概述

语音情感识别（Speech Emotion Recognition，SER）是自动语音识别（Automatic Speech Recognition，ASR）领域的重要研究方向，在人机交互中具有重要作用。随着ASR技术的快速发展，以计算机、手机、

基金项目：国家自然科学基金（62066039）。

作者简介：张会云（1993—），女，博士研究生，主研方向为模式识别与智能系统、语音情感识别；黄鹤鸣，教授、博士。

收稿日期：2021-03-10　　**修回日期：**2021-04-27　　**E-mail：**1406043513@qq.com

平板等为载体的人工智能(Artificial Intelligence,AI)研究层出不穷。人机交互不再局限于识别特定说话人语音中的单一音素或语句,语音中的情感识别已成为ASR领域的新兴研究方向。例如:在远程教学中,实时检测学生情绪,能够提高教学质量[1];在移动通信中,增加情感分析功能,能够及时检测客户的情绪变化,并根据这种变化为客户提供更好的服务[2];在医学实践中,实时检测病人情绪能够提供更好的临床治疗[3];在侦察破案中,通过检测情感状态能识破嫌疑人是否撒谎,保证案件顺利进行[4];在电商领域中,通过识别用户情感可以调控流量[5]。总而言之,准确高效地识别语音情感有助于提高人们工作、学习和生活的效率与质量。

本文建立基于注意力机制的异构并行卷积循环神经网络(Recurrent Neural Network,RNN)模型AHPCL。该模型由2个异构并行分支和1个注意力机制构成:左分支由2个全连接层和1个长短时记忆(Long Short-Term Memory,LSTM)层构成,右分支由1个全连接层、1个卷积层和1个LSTM层构成,注意力机制由1个全连接层和1个注意力层构成。通过在EMODB、CASIA、SAVEE等3个语音情感数据库上提取音高(Pitch)、过零率(Zero Crossing Rate,ZCR)、梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient,MFCC)等低级描述符特征,同时计算这些特征的高级统计函数,得到共219维的特征作为输入来评估模型性能。

1 相关工作

SER是指利用计算机对语音信息进行预处理,提取情感特征,建立特征值与情感的映射关系,从而对情感进行分类[6]。SER主要包括语料库构建、情感信号预处理、情感声学特征提取以及声学建模4个环节。在预处理方面,语音情感信号的预处理与语音识别的预处理一样,均需要进行预加重、分帧、加窗、端点检测等操作[7]。情感声学特征提取是SER中一项极具挑战性的任务,对语音情感的识别严重依赖于语音情感特征的有效性。提取关联度更高的声学特征更有助于确定说话人的情感状态。通常以帧为单位提取语音信息的声学特征,并将全局统计结果作为模型的输入进行情感识别。一般而言,单一特征不能完全包含语音情感的所有有用信息,为了使SER系统性能达到最优,研究人员通常融合不同特征来提高系统性能。高帆等[8]利用深度受限玻尔兹曼机将韵律特征、谱特征进行融合,并在EMODB数据库上验证DBM-LSTM模型的性能。实验结果表明,与传统识别模型相比,DBM-LSTM模型更适用于多特征语音情感识别任务,最优识别准确率提升了11.00%。宋春晓[9]研究了语速、过零率、基频、能量、共振峰、MFCC等特征在EMODB数据库上的性能,采用SVM识别4类情感时获得了82.47%的准确率。GUO等[10]提取对数梅尔频谱特征,计算一阶差分和二阶差分,并融合这些统计值作为并行卷积循环神经网络模型的输入,在SAVEE数据库上取得了59.40%的未加权召回率。

声学模型是SER系统的核心。在识别过程中,情感特征输入到声学模型,计算机通过相应算法获取识别结果。MIRSAMADI等[11]利用LSTM网络提取深度学习特征,在IEMOCAP数据库上采用SVM识别情感,获得了63.50%的识别准确率。ZHANG等[12]提取了深度学习特征,在SEED和CK+数据库上采用循环神经网络识别情感,分别获得了89.50%和95.40%的识别准确率。传统LSTM网络假设当前时间步长的模型状态取决于前一个时间步长的模型状态,该假设限制了网络的时间依赖性建模能力,而TAO等[13]提出的Advanced-LSTM网络较好地克服了该限制,能更好地进行时间上下文建模,获得了55.30%的召回率,优于传统LSTM网络。

2 基于注意力机制的异构并行卷积循环神经网络模型

随着深度学习技术的不断发展,神经网络结构越来越复杂。与已有简单的前馈神经网络相比,RNN的隐含层之间既有前馈连接又有内部反馈连接[14]。RNN能较好地处理序列数据,但存在梯度问题,而LSTM中的门控循环单元能够较好地解决梯度问题,同时门控循环单元也能够对先前的信息进行选择性记忆[15],从而使得网络的预测结果更加准确。因此,本文选择LSTM提取语音情感的时间序列特征。但由于仅提取时间序列信息并不能很好地表征语音情感,因此同时采用卷积操作提取语音空间信息[16]。通过时间信息和空间信息共同表征语音情感,能使预测结果更理想。此外,注意力机制可以对来自不同时刻的帧特征给予不同关注[17]。

基于此,本文构建基于注意力机制的异构并行卷积循环神经网络模型AHPCL,如图1所示。该网络模型由2个异构并行分支和1个注意力机制构成,其中,左分支包含2个全连接层和1个LSTM层,右分支包含1个全连接层、1个卷积层和1个LSTM层,注意力机制包含1个全连接层和1个注意力层。拼接来自左右2个分支结构的数据,并在注意力层将拼接后的数据与原始输入数据中的对应元素相乘,将相乘后的结果输入到4个完全相同的全连接层,最终输入到Softmax层进行分类。

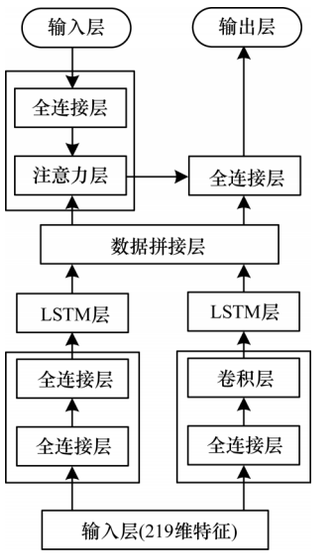

图1　基于注意力机制的异构并行卷积循环神经网络

Fig.1　Heterogeneous parallel convolutional recurrent neural network based on attention mechanism

AHPCL模型在卷积层的计算如下：

$$h=f\left(\frac{h^1 * \boldsymbol{F}}{S}\times N\right) \tag{1}$$

其中，*表示卷积运算；h^1表示第一个全连接层的输出；$\boldsymbol{F}=\left[k_1, k_2, \cdots, k_{512}\right]$表示卷积核；$N$表示滤波器个数；$S$表示步长。

AHPCL模型在数据拼接层y_L^B的计算如下：

$$F_c=\text{concatenate}\left(y_{\text{L}}^B, y_{\text{R}}^B\right) \tag{2}$$

其中，y_{L}^B、y_{R}^B分别是AHPCL网络的左右分支输出；concatenate(·)表示拼接来自左右2个分支的数据。

AHPCL模型在注意层的计算如下：

$$\alpha=\text{Multiply}\left(u\cdot F_c\right) \tag{3}$$

其中：u表示经过注意力机制后第一个全连接层的输出；F_c表示数据拼接层的输出；Multiply(·)表示对应元素的乘积。

3　数据库描述与特征提取

为评估AHPCL模型的性能，在EMODB、CASIA及SAVEE情感数据库上提取低级描述符特征，并计算相关的高级统计函数作为模型的输入。

3.1　数据库

CASIA[18]是由中科院自动化研究所在干净环境下录制的汉语语音情感数据库，包含4位专业发音人在高兴(Happiness，H)、恐惧(Fear，F)、悲伤(Sadness，Sa)、愤怒(Anger，A)、惊讶(Surprise，Su)、中性(Neural，N)等6类情感下演绎的9 600条情感语音，采样率为16 kHz。目前公开的CASIA库中包含1 200条情感语音，每类情感各200条情感语音。

EMODB[19]是由柏林工业大学在专业录音室录制的德语语音情感数据库，采样率为48 kHz。从40位说话人中选取10位(5男5女)对10句德语语句进行情感演绎并录音，包含中性、愤怒、恐惧、高兴、悲伤、厌恶(Disgust，D)、无聊(Boredom，B)等7类情感，共800条情感语音，考虑到每条语句的语音自然度，最终选取535个样本，对上述7类情感而言，每类情感包含的样本数量分别为79、127、69、71、62、46、81。

SAVEE[20]是由4名演员演绎愤怒、厌恶、恐惧、高兴、中性、悲伤、惊讶等7类情感得到的表演型数据库，共480条情感语音，语音情感数量分布相对平衡，除中性情感以外，其余6类均有60条情感语音。

由于上述3个数据库均未提供单独的训练数据和测试数据，因此本文采用说话人相关(Speaker-Dependent，SD)策略：每类情感的所有样本随机等分为5份，其中，4份作为训练数据，1份作为测试数据。实验重复10次取均值作为模型的整体性能评估数据。

3.2　特征提取

在提取音高、过零率、梅尔频率倒谱系数、幅度(Amplitude)、谱重心(Centroid)、频谱平坦度(Flatness)、色谱图(Chroma)、梅尔谱图(Mel)、谱对比度(Contrast)等低级描述符特征的基础上，计算这些特征的高级统计函数，得到共219维特征作为AHPCL模型的输入，所提取与计算的全部特征见表1。

表1　低级描述符与高级统计函数特征

Table 1　Low-level descriptors and high-level statistical function features

特征	低级描述符	高级统计函数
韵律特征	音高	均值、方差、最大值
	过零率	均值
谱特征	梅尔频率倒谱系数、幅度、谱重心	均值、方差、最大值
	频谱平坦度、色谱图、梅尔谱图、谱对比度	均值

4　实验与结果分析

在CASIA、EMODB以及SAVEE数据库上验证AHPCL模型性能。首先计算AHPCL模型在10次验证中的均值，用于评价模型的整体性能。其次选取AHPCL模型在10次验证中所获得的最佳混淆矩阵。最后将AHPCL模型与已有研究成果进行对比。

4.1　实验设置

实验运行在一台高性能服务器上，CPU为40核80线程，内存为64 GB。使用2块RTX 2080 Ti GPU进行加速训练。利用深度学习框架Keras和TensorFlow进行模型搭建。采用的优化器(Optimizer)为Adam，激活函数为Leaky ReLU，批处理(Batch_size)大小为32，丢弃率(Dropout)为0.5，迭代周期(Epoch)为100。基于混淆矩阵、准确率、精确率、未加权平均召回率(Unweighted Average Recall，UAR)、F1得分等

指标对模型性能进行评价。

4.2 结果分析

在CASIA、EMODB、SAVEE数据库上对AHPCL模型进行10次验证，模型在每个数据库上的均值和波动程度如图2所示，其中：箱体中间的一条虚线表示数据的中位数；箱体的上下限分别是数据的上四分位数和下四分位数，这意味着箱体包含了50%的数据；箱体的高度在一定程度上反映了数据的波动程度；在箱体的上方和下方各有一条线，分别表示最高准确率和最低准确率。

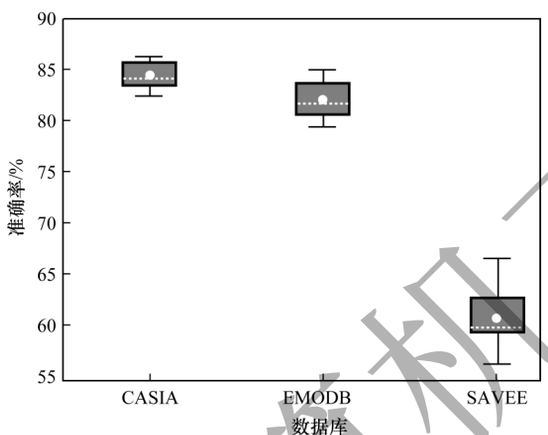

图2 AHPCL模型在CASIA、EMODB、SAVEE数据库上的箱线图

Fig.2 Box-plot of AHPCL model on CASIA, EMODB, and SAVEE databases

由图2可以看出：在10次验证中，AHPCL模型在CASIA、EMODB、SAVEE这3个数据库上的最高准确率依次为86.25%、85.05%、66.67%，最低准确率依次为82.50%、79.44%、56.25%，平均准确率依次为84.50%、82.06%、60.84%。由此可见：1）AHPCL模型在CASIA数据库上最高准确率和最低准确率相差最小，EMODB数据库次之，SAVEE数据库相差最大，即AHPCL模型在CASIA数据库上的波动程度最小，稳定性最好；2）AHPCL模型在CASIA数据库上的均值最高，表明取得了最佳性能。AHPCL模型在CASIA数据库上性能最佳的主要原因为：CASIA数据库仅包含6类情感，少于其他2个数据库中的7类情感，类别数少有利于识别；CASIA数据库中样本数据量是EMODB、SAVEE数据库的2倍多，模型得到了更好训练。

图3~图5选取了AHPCL模型在CASIA、EMODB、SAVEE数据库上的最佳混淆矩阵。如图3所示，AHPCL模型在CASIA数据库上6类情感的准确率、精确率、未加权平均召回率以及F1得分依次为86.25%、85.77%、86.02%、85.90%。从图3可以看出：愤怒、惊讶、中性这3类情感的召回率均达到了90.00%以上；恐惧和悲伤这2类情感的识别率较低且这2类情感容易混淆，即在恐惧类情感中，有15.79%的样本被预测为悲伤，同样地，在悲伤类情感的识别过程中，有23.08%的样本被预测为恐惧。如图4所示，AHPCL模型在EMODB数据库上7类情感的准确率、精确率、未加权平均召回率以及F1得分依次为85.05%、86.33%、84.03%、85.16%。从图4可以看出：高兴类情感的识别准确率较低，33.33%的样本被误判为愤怒类情感，13.33%的样本被误判为恐惧类情感，仅有46.67%的样本识别正确；其余情感均取得了较好的识别性能，愤怒情感的召回率达到了100.00%。如图5所示，AHPCL模型在SAVEE数据库上7类情感的准确率、精确率、未加权平均召回率以及F1得分依次为66.67%、64.35%、64.06%、64.20%。从图5可以看出，7类情感的平均召回率为64.06%，愤怒、厌恶、恐惧这3类情感的召回率均较低，高兴情感的召回率最高，达到81.82%。

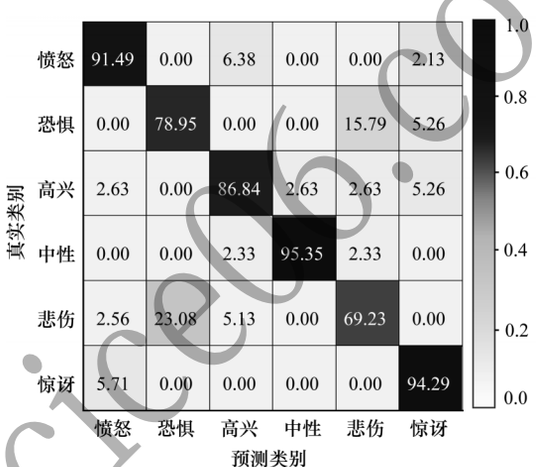

图3 AHPCL模型在CASIA数据库上的混淆矩阵

Fig.3 Confusion matrix of AHPCL model on CASIA database

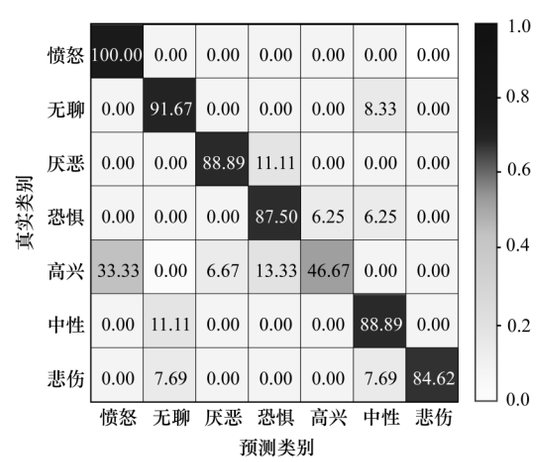

图4 AHPCL模型在EMODB数据库上的混淆矩阵

Fig.4 Confusion matrix of AHPCL model on EMODB database

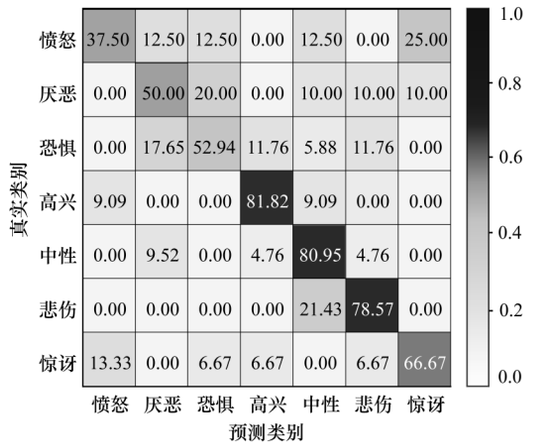

图5　AHPCL模型在SAVEE数据库上的混淆矩阵

Fig.5　Confusion matrix of AHPCL model on SAVEE database

AHPCL模型与DNN-ELM[16]、LeNet[18]、WADAN-CNN[19]、TSFFCNN[20]、GA-BEL[21]、HuWSF[22]、LNCMSF[23]、DCNN+LSTM[24]、FDNNSA[25]、RDBN[26]、ACRNN[27]、2D CNN[28]、RF[29]等同类模型的性能对比见表2，其中，WAR是指加权平均召回率（Weighted Average Recall，WAR），CASIA中的WAR与UAR相同的原因为CASIA中各类情感样本数量完全相等，均为200，即各类样本在总样本中所占的比重（权重）是一样的，因此这2个指标相等。

表2　在CASIA、EMODB、SAVEE数据库上AHPCL模型与现有模型的性能对比

Table 2　Performance comparison of AHPCL model with other models on CASIA，EMODB，and SAVEE databases　%

数据库	声学模型	WAR	UAR
CASIA	GA-BEL	38.55	38.55
	HuWSF	43.50	43.50
	RDBN	48.50	48.50
	DCNN+LSTM	72.80	72.80
	FDNNSA	83.00	83.00
	LeNet	85.80	85.80
	AHPCL	86.02	86.02
EMODB	HuWSF	81.74	
	RDBN	82.32	
	LNCMSF		74.46
	ACRNN		82.82
	WADAN-CNN	84.49	83.31
	2D CNN		83.40
	DNN-ELM		84.56
	AHPCL		84.03
SAVEE	GA-BEL	44.18	
	HuWSF	50.00	
	RDBN	53.60	
	RF		56.07
	TSFFCNN		62.54
	AHPCL		64.06

由表2可以看出：在CASIA数据库上，AHPCL模型的性能均优于6类基线模型，WAR和UAR比最好的基线模型LeNet[18]高出0.22个百分点；在EMODB数据库上，AHPCL模型的UAR仅比DNN-ELM模型[16]低0.53个百分点，除此之外，AHPCL模型的性能均优于其余6类基线模型；在SAVEE数据库上，AHPCL模型的性能均优于5类基线模型的性能，而且UAR比最优的TSFFCNN基线模型[20]高出1.52个百分点。

综上：AHPCL模型在CASIA、SAVEE这2个数据库上的性能均优于现有研究成果，在EMODB数据库上也与现有研究成果相当，证明了AHPCL模型的鲁棒性和泛化性均较好。

5　结束语

为提高语音情感识别性能，本文提出一种基于注意力机制的异构并行卷积循环神经网络模型AHPCL。在卷积层提取语音情感的空间谱特征，在LSTM层提取语音情感的时间序列特征，同时基于注意力机制，根据不同的时间序列特征对语音情感的贡献程度分配权重。实验结果表明，该模型能同时提取语音情感的空间谱特征和时间序列特征，具有较强的鲁棒性和泛化性。后续将使用向量胶囊网络替代AHPCL模型卷积层中的一维卷积，并将模型应用于混合语言的语音情感识别中，进一步提升鲁棒性和泛化性。

参考文献

[1] TRIANTO R，TAI T C，WANG J C. Fast-LSTM acoustic model for distant speech recognition[C]//Proceedings of 2018 IEEE International Conference on Consumer Electronics. Washington D. C.，USA：IEEE Press，2018：1-4.

[2] OCQUAYE E N N. Speech emotion recognition via domain adaptation[D]. Zhenjiang：Jiangsu University，2020.

[3] 张鹤鹏. 基于表情和语音信号的情感识别研究[D]. 济南：山东大学，2020.
ZHANG H P. Research on emotion recognition based on expression and speech signal[D]. Jinan：Shandong University，2020.（in Chinese）

[4] 张昕然. 跨库语音情感识别若干关键技术研究[D]. 南京：东南大学，2016.
ZHANG X R. Research on several key technologies in cross-corpus speech emotion recognition[D]. Nanjing：Southeast University，2016.（in Chinese）

[5] TZIRAKIS P，ZHANG J H，SCHULLER B W. End-to-end speech emotion recognition using deep neural networks[C]//Proceedings of 2018 IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2018：5089-5093.

[6] LIAN Z，LIU B，TAO J H. CTNet：conversational transformer network for emotion recognition[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021，29(1)：985-1000.

[7] DENG J, XU X Z, ZHANG Z X, et al. Semisupervised autoencoders for speech emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 31-43.
[8] 高帆, 张雪英, 黄丽霞, 等. 基于DBM-LSTM的多特征语音情感识别[J]. 计算机工程与设计, 2020, 41(2): 465-470.
GAO F, ZHANG X Y, HUANG L X, et al. Multi-feature speech emotion recognition based on DBM-LSTM[J]. Computer Engineering and Design, 2020, 41(2): 465-470. (in Chinese)
[9] 宋春晓. 情感语音的非线性特征提取及特征优化的研究[D]. 太原: 太原理工大学, 2018.
SONG C X. Research on nonlinear feature extraction and feature optimization of emotional speech[D]. Taiyuan: Taiyuan University of Technology, 2018. (in Chinese)
[10] GUO L L, WANG L B, DANG J W, et al. A feature fusion method based on extreme learning machine for speech emotion recognition [C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 2666-2670.
[11] MIRSAMADI S, BARSOUM E, ZHANG C. Automatic speech emotion recognition using recurrent neural networks with local attention [C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2017: 2227-2231.
[12] ZHANG T, ZHENG W M, CUI Z, et al. Spatial-temporal recurrent neural network for emotion recognition[J]. IEEE Transactions on Cybernetics, 2019, 49(3): 839-847.
[13] TAO F, LIU G. Advanced LSTM: a study about better time dependency modeling in emotion recognition [C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 2906-2910.
[14] PENG Z C, ZHU Z, UNOKI M, et al. Auditory-inspired end-to-end speech emotion recognition using 3D convolutional recurrent neural networks based on spectral-temporal representation[C]//Proceedings of 2018 IEEE International Conference on Multimedia and Expo. Washington D. C., USA: IEEE Press, 2018: 1-6.
[15] SHU X B, ZHANG L Y, SUN Y L, et al. Host-parasite: graph LSTM-in-LSTM for group activity recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 663-674.
[16] 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.
[17] XIE Y, LIANG R Y, LIANG Z L, et al. Speech emotion classification using attention-based LSTM[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1675-1685.
[18] 姜芃旭, 傅洪亮, 陶华伟, 等. 一种基于卷积神经网络特征表征的语音情感识别方法[J]. 电子器件, 2019, 42(4): 998-1001.
JIANG P X, FU H L, TAO H W, et al. Feature characterization based on convolution neural networks for speech emotion recognition[J]. Chinese Journal of Electron Devices, 2019, 42(4): 998-1001. (in Chinese)
[19] YI L, MAK M W. Improving speech emotion recognition with adversarial data augmentation network [EB/OL]. [2021-02-14]. https://www.researchgate.net/publication/346055097_Improving_Speech_Emotion_Recognition_With_Adversarial_Data_Augmentation_Network.
[20] WU M, SU W J, CHEN L, et al. Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition [EB/OL]. [2021-02-14]. https://www.researchgate.net/publication/338564019_Two-stage_Fuzzy_Fusion_based-Convolution_Neural_Network_for_Dynamic_Emotion_Recognition.
[21] LIU Z T, XIE Q, WU M, et al. Speech emotion recognition based on an improved brain emotion learning model[J]. Neurocomputing, 2018, 309: 145-156.
[22] SUN Y X, WEN G H, WANG J B. Weighted spectral features based on local Hu moments for speech emotion recognition[J]. Biomedical Signal Processing and Control, 2015, 18(1): 80-90.
[23] TAO H W, LIANG R Y, ZHA C, et al. Spectral features based on local Hu moments of Gabor spectrograms for speech emotion recognition[J]. IEICE Transactions on Information and Systems, 2016, E99-D(8): 2186-2189.
[24] 缪裕青, 邹巍, 刘同来, 等. 基于参数迁移和卷积循环神经网络的语音情感识别[J]. 计算机工程与应用, 2019, 55(10): 135-140, 198.
MIAO Y Q, ZOU W, LIU T L, et al. Speech emotion recognition model based on parameter transfer and convolutional recurrent neural network[J]. Computer Engineering and Applications, 2019, 55(10): 135-140, 198. (in Chinese)
[25] CHEN L, SU W J, WU M, et al. A fuzzy deep neural network with sparse autoencoder for emotional intention understanding in human-robot interaction [J]. IEEE Transactions on Fuzzy Systems, 2020, 28(7): 1252-1264.
[26] WEN G, LI H, HUANG J, et al. Random deep belief networks for recognizing emotions from speech signals[EB/OL]. [2021-02-14]. http://europepmc.org/article/PMC/5357547.
[27] CHEN M Y, HE X J, YANG J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters, 2018, 25(10): 1440-1444.
[28] 乔栋, 陈章进, 邓良, 等. 改进语音处理的卷积神经网络中文语音情感识别[J]. 计算机工程, 2022, 48(2): 281-290.
QIAO D, CHEN Z J, DENG L, et al. Speech emotion recognition based on improved speech processing and convolution neural network[J]. Computer Engineering, 2022, 48(2): 281-290. (in Chinese)
[29] NOROOZI F, MARJANOVIC M, NJEGUS A, et al. Audio-visual emotion recognition in video clips [J]. IEEE Transactions on Affective Computing, 2019, 10(1): 60-75.

编辑 陆燕菲