·人工智能与模式识别·　　文章编号：1000-3428(2022)07-0059-07　　文献标志码：A　　中图分类号：TP183

基于多操作网络的图式多域语音情感识别研究

张会云[1,2,3,4]，黄鹤鸣[1,2,3,4]

（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；
3.藏文信息处理教育部重点实验室，西宁 810008；4.青海省藏文信息处理与机器翻译重点实验室，西宁 810008）

摘　要：多域语音情感识别研究在语料标注方法、录制场景以及交互方式等方面存在差异性，使得构建多域语音情感识别系统变得较为复杂。设计一种基于多操作网络的多域语音情感识别模型，通过组合CASIA、EMODB、SAVEE 3个单域数据库，构建Hybrid-CE、Hybrid-ES、Hybrid-CS、Hybrid-CES 4种多域语音情感数据库及层级多操作网络(HMN)。HMN网络由2个异构并行分支组成，左分支由2个同构并行的一维卷积层构成，卷积层的神经元数量均为128，右分支由并行的Bi-GRU层和Bi-LSTM层构成，GRU和LSTM的记忆单元数量均为64。将原始数据投影到不同的变换空间进行计算，从而更准确地表征语音的情感信息。通过分层的Concate、Add和Multiply多操作运算，将左右分支提取的不同特征进行多重融合。在此基础上，计算梅尔频率倒谱系数、色谱图、谱对比度等低级描述符特征的高级统计函数，得到219维特征作为模型HMN的输入。实验结果表明，该模型在4种多域数据库上的F1-score分别达到82.22%、65.02%、70.59%、73.47%，具有较好的鲁棒性和泛化性。

关键词：语音情感识别；韵律特征；谱特征；多特征融合；多操作网络

开放科学(资源服务)标志码(OSID)：

中文引用格式：张会云，黄鹤鸣.基于多操作网络的图式多域语音情感识别研究[J].计算机工程，2022，48(7)：59-65.

英文引用格式：ZHANG H Y，HUANG H M.Research on schema multi-domain speech emotion recognition based on multi-operation network[J].Computer Engineering，2022，48(7)：59-65.

Research on Schema Multi-Domain Speech Emotion Recognition Based on Multi-Operation Network

ZHANG Huiyun[1,2,3,4]，HUANG Heming[1,2,3,4]

（1.School of Computer Science，Qinghai Normal University，Xining 810008，China；
2.The State Key Laboratory of Tibetan Intelligent Information Processing and Application，Xining 810008，China；
3.Key Laboratory of Tibetan Information Processing，Ministry of Education，Xining 810008，China；
4.Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province，Xining 810008，China）

【Abstract】Research on multi-domain Speech Emotion Recognition(SER) faces the problem that most available speech corpora differ from each other in crucial ways，such as annotation methods，recording scenarios，interaction mode，etc.，thereby making the construction of multi-domain SER system more complex.This paper proposes a multi-domain SER model based on a multi-operation network.First，databases such as CASIA，EMODB，and SAVEE，are combined for the first time to construct 4 multi-domain speech emotion databases.The HMN network is composed of two heterogeneous parallel branches.The left branch is composed of two isomorphic parallel one dimensional convolutional layers，both of which comprise 128 neurons.The right branch is composed of parallel Bi-GRU layer and Bi-LSTM layer，both of which have 64 memory units.The original data are projected to different transform Spaces for calculation so that the emotional information of speech can be more accurately represented.Multiple fusion of different features extracted from left and right branches is performed by hierarchical multi-operation operations Concate，Add，and Multiply.Accordingly，the advanced statistical functions of Mel Frequency Cepturm Comfficient(MFCC)，chroma，contrast，and other low level descriptor features were calculated，and 219 dimensional features were obtained as the input of model HMN.Experimental results reveal that the F1-score of the proposed model is 82.22%，65.02%，70.59%，and 73.47%，respectively，with good robustness and generalization.

【Key words】Speech Emotion Recognition(SER)；prosodic feature；spectral feature；multi-feature fusion；multi-operation network

DOI：10. 19678/j. issn. 1000-3428. 0061981

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、智能系统、语音情感识别；黄鹤鸣(通信作者)，教授、博士。

收稿日期：2021-07-05　　**修回日期：**2021-08-25　　**E-mail：**1406043513@qq.com

0 概述

情感被认为是对生存[1]或机体行为[2]有关情况的典型反应[3]。在几乎所有关于情感的理论解释中,感觉加工有着非常重要的作用[4-6],但是神经科学的观点认为,情感是由大脑的特定区域驱动的,例如,在边缘系统[7]和相关的皮层下回路[8]中,神经回路被认为是专门处理诸如恐惧和悲伤等情感类别的。根据上述观点,感觉皮层的活动被认为是情感的先决条件,而听觉作为一级感觉区,对情感信息的加工具有至关重要的作用[9]。

语音情感识别是指计算机以帧为单位对情感信号进行特征提取,模拟人类感知并理解人类情感,进而推断语音情感类型的一种技术[10]。常用的语音情感识别(Speech Emotion Recognition,SER)方法是在标注的数据库上训练和测试分类器,或者将数据集划分为训练集、验证集和测试集进行交叉验证[11]。通过这种方式,识别模型在特定的说话群体、语言与情感类别等方面都取得了很好的性能。但这种识别模型能在多大程度上推广到不同交互场景和语言中还不能得出结论。

近年来,研究人员致力于多域语音情感识别研究。文献[12]对多域语音情感识别进行了初步探索,在不同语料库组合而成的训练集上验证了6种语音情感的识别性能,但由于不清楚哪些因素对识别结果产生影响,因此对识别结果的解释相对模糊;文献[13]对来自4个语系的8种语言进行研究,结果表明多域情感识别是可行的;文献[14]提出一种基于语言识别和模型选择的多域语音情感分类方法,在多域语音情感数据库上验证了模型的识别性能;文献[15]结合两种语言进行语音情感识别研究,利用直方图均衡化消除跨域语音情感表达之间的差异。

关于多域语音情感识别模型的分类性能,目前很难与其他多域语音情感识别模型在同一基准下进行比较,因为多域语音情感识别研究在诸如情感类别、训练集和测试集的划分、潜在的情感概念(离散情感或连续唤醒/效价维度)等方面没有统一标准[16],且目前各种多域语音情感识别研究至少在一个方面有所不同,因此,无法在同一基准下进行分类性能的比较。目前,对于多域和跨域语音情感识别[17]往往以单域语音情感识别为基线进行性能比较。

基于已有研究及上述问题,本文构建多域语音情感数据库Hybrid-CE、Hybrid-ES、Hybrid-CS及Hybrid-CES,通过多操作运算实现韵律特征和谱特征等低级描述符的高级统计函数特征的融合,提出一种新颖的图式层级多操作网络(Hierarchical Multi-operation Network,HMN)模型。最后通过实验验证HMN模型在多域语音情感数据库上的分类性能、鲁棒性和泛化性。

1 层级多操作网络

随着深度学习的不断发展,神经网络的结构越来越复杂。与前馈网络相比,循环神经网络(Recurrent Neural Network,RNN)[18]能较好地处理序列数据,但存在梯度消失或者梯度爆炸问题;而长短时记忆(Long Short-Term Memory,LSTM)网络和门控循环单元(Gated Recurrent Unit,GRU)能够较好地解决梯度问题,同时对信息实现选择性记忆[19]。为了更好地利用上下文语境信息,本文研究采用双向长短时记忆(Bi-LSTM)网络和双向门控循环单元(Bi-GRU)共同提取语音情感的时间序列信息[20],通过完整地表征语音情感特征,利用卷积操作提取语音空间信息[21-22]。同时,采用Concate、Add和Multiply多操作运算,更多地保留和突出原始语音的情感信息。基于此,本文构建了层级多操作网络HMN,如图1所示。HMN主要由两个异构并行分支和多操作层构成。

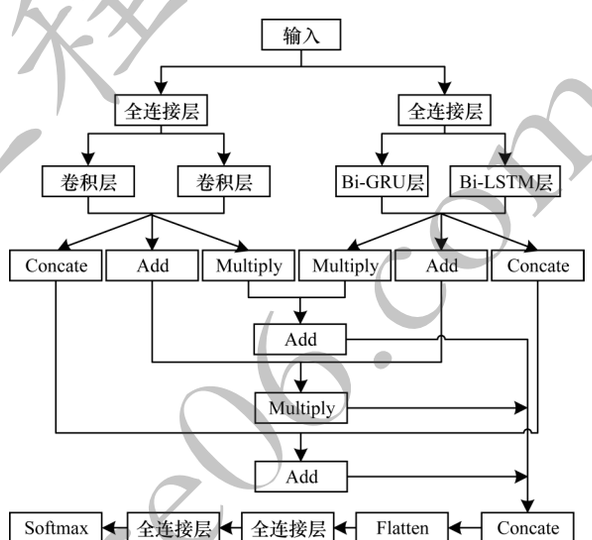

图1 层级多操作网络结构

Fig.1 Hierarchical multi-operation network structure

首先在两个异构并行分支中,左分支由两个同构并行的一维卷积层构成,卷积层的神经元数量均为128;右分支由并行的Bi-GRU层和Bi-LSTM层构成,GRU和LSTM的记忆单元数量均为64。设立左右分支的目的是将原始数据投影到不同的变换空间进行计算,从而更准确地表征语音的情感信息。

接着通过分层的多操作运算将左右分支提取的不同特征进行多重融合。左分支中有两个子分支,将每一个子分支中的数据分别进行Concate、Add和Multiply操作。其中,Concate操作用于联合特征矩阵,这种操作增加了描述原始数据的特征维数,但每维特征对应的信息并未增加;Add操作叠加特征矩阵中对应位置的元素,这种操作虽未增加原始数据特征维数,但增加了每一维特征的信息量;Multiply操作将特征矩阵对应位置元素进行相乘,进一步突出显著性信息。对右分支中两个子分支中的数据同样进行Concate、Add和Multiply操作。

最后融合左右分支中的信息,即将左右分支中Multiply操作后得到的数据进行Add运算,Concate操作后得到的数据进行Add运算,Add操作后的数据进行Multiply运算,将得到的3个运算结果进行Concate操作拼接成219×512维的特征,并采用

Flatten操作将其平滑为一维数组，输入到神经元个数分别为128和64的两个全连接层中，最后采用Softmax函数进行分类。

HMN模型中数据的流动过程如下：1）将语音谱特征和韵律特征的高级统计函数值输入异构的两个并行分支；2）将左右两个分支的数据进行多重融合；3）拼接左右两个分支融合后的数据，进一步采取平滑操作后输入到2个全连接层；4）在输出层进行分类。

在模型HMN中，卷积层的计算为：

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

其中：h^1是第一个全连接层的输出；$F = \left[k_1, k_2, \cdots, k_{512}\right]$是卷积核；$N$是滤波器个数；$S$是步长。

操作Concate、Add和Multiply的计算公式如式（2）~式（4）所示：

$$F_C = \text{Concate}\left(y_L, y_R\right) \tag{2}$$

$$F_A = \text{Add}\left(y_L \oplus y_R\right) \tag{3}$$

$$F_M = \text{Multiply}\left(y_L \otimes y_R\right) \tag{4}$$

其中：Concate（·）拼接左右两个分支的数据y_L和y_R；Add（·）对y_L和y_R的对应元素求和；Multiply（·）将y_L和y_R的对应元素相乘。

2　数据集描述

为了评估HMN模型的性能，首先分别在自行构建的4个图式多域数据库Hybrid-CE、Hybrid-ES、Hybrid-CS以及Hybrid-CES上提取低级描述符（Low-Level Descriptor，LLD）特征[23]。其中，图式指存在于记忆中的认知结构或知识结构[3]，本文采用图式原理将单域数据集中的研究方法迁移到多域数据集中。其次计算LLD特征的高级统计函数（High-level Statistical Functions，HSF）值[24]作为HMN模型的输入。

2.1　单域数据集

CASIA是由中科院自动化研究所录制的中文语音情感数据库[22]。该库是由4位说话人分别演绎高兴（Happiness，H）、恐惧（Fear，F）、悲伤（Sadness，Sa）、生气（Anger，A）、惊讶（Surprise，Su）和中性（Neural，N）6类情感而录制的。在公开的CASIA库中包含6类情感，每类情感各200条，共1 200条情感语音。

EMO-DB是由柏林工业大学录制的德语语音情感数据库[25]。由10位说话人（5男5女）对10个德语语句进行中性（N）、生气（A）、恐惧（F）、高兴（H）、悲伤（Sa）、厌恶（Disgust，D）和无聊（Boredom，B）7类情感演绎得到。每类情感的样本数量依次为79、127、69、71、62、46、81，共535个样本。

SAVEE是由4名演员演绎生气（A）、厌恶（D）、恐惧（F）、高兴（H）、中性（N）、悲伤（Sa）以及惊讶（Su）7类情感得到的表演型数据库[26]。SAVEE语音情感数量分布相对平衡，共有480条情感样本，除中性外，其余6类情感均有60条语句。

2.2　多域数据集

通过合并CASIA、EMO-DB和SAVEE 3个单域数据集构建4种图式多域语音情感数据集Hybrid-CE、Hybrid-ES、Hybrid-CS以及Hybrid-CES。其中，Hybrid-CE由单域数据集CASIA[22]和EMODB[25]合并而成，Hybrid-ES由单域数据集EMODB和SAVEE[26]合并而成，Hybrid-CS由单域数据集CASIA和SAVEE合并而成，而Hybrid-CES由单域数据集CASIA、EMODB以及SAVEE合并而成。

合并方式如下：将2个或者3个单域数据集合并为1个新的多域数据集；将拟合并单域数据集共有的情感类别对应的样本合并，得到多域数据集的一类；若某类情感在某个单域数据集上独有则单独作为一类。例如，通过合并单域数据库CASIA和EMODB构建多域数据库Hybrid-CE时，CASIA包含6类情感，EMODB包含7类情感，合并两个数据集共有的高兴、恐惧、悲伤、生气、中性5类情感，分别得到新构建的Hybrid-CE库中5类情感样本；惊讶类情感仅出现在CASIA库中，而EMODB库中无此类情感，此时将惊讶类情感作为Hybrid-CE库的一类新的情感；同理，EMODB库中包含无聊和厌恶类情感，而CASIA库中无此类情感，则将无聊和厌恶作为Hybrid-CE库中2个新的情感类别，最终Hybrid-CE库中包含8个情感类别：即愤怒、无聊、恐惧、厌恶、高兴、惊讶、中性、悲伤，如表1所示。多域数据库Hybrid-ES、Hybrid-CS以及Hybrid-CES的构建方式与Hybrid-CE类似。

表1　4种多域语音情感数据库的相关信息

Table 1　Relevant information of four multi-domain speech emotion databases

数据库	语言类型	说话人数	情感类别（样本数）	样本总数
Hybrid-CE	中文 德语	14位 （7男7女）	愤怒(327)，无聊(81) 恐惧(269)，厌恶(46) 高兴(271)，惊讶(200) 中性(279)，悲伤(262)	1 735
Hybrid-ES	德语 英语	14位 （9男5女）	愤怒(187)，无聊(81)，厌恶(106)，恐惧(129)，高兴(131)， 中性(199)，悲伤(122)，惊讶(60)	1 015
Hybrid-CS	中文 英语	8位 （6男2女）	愤怒(260)，厌恶(60) 恐惧(260)，高兴(260)，中性(320)，悲伤(260)，惊讶(260)	1 680
Hybrid-CES	中文 德语 英语	18位 （11男7女）	愤怒(387)，无聊(81)，厌恶(106)，恐惧(329)，高兴(331)， 中性(399)，悲伤(322)，惊讶(260)	2 215

表1展示了本文所构建的4种多域语音情感数据库的语言类型、说话人数、情感类别、每类情感中的样本数及总样本数等信息。

3 特征提取

韵律特征[27]和谱特征[28]是语音情感的主流特征，因此，本文提取了音高(Pitch)、调谐、过零率(Zero Crossing Rate，ZCR)等韵律特征以及梅尔频率倒谱系数(Mel Frequency Ceptrum Cofficient，MFCC)、幅度(Amplitude)、谱重心(Centroid)、频谱平坦度(Flatness)、色谱图(Chroma)、梅尔频谱(Mel)以及谱对比度(Contrast)等谱特征，并计算这些特征的高级统计函数值，将得到的219维特征作为HMN模型的输入。所提取的低级描述符、高级统计函数特征以及相应的维数如表2所示。

表2 低级描述符与高级统计函数特征

Table 2 Low level descriptors and high level statistical function feature

特征	低级描述符(维度)	高级统计函数
韵律特征	音高(1×63D)	均值，方差，最大值(1D)
	调谐(1×63D)	偏移(1D)
	过零率(1×1D)	均值(1D)
谱特征	幅度(1×63D)	均值，方差，最大值(1D)
	谱重心(1×63D)	均值，方差，最大值(1D)
	MFCC(20×63D)	均值，方差，最大值(20D)
	频谱平坦度(1×63D)	均值(1D)
	色谱图(12×63D)	均值(12D)
	梅尔谱图(128×63D)	均值(128D)
	谱对比度(7×63D)	均值(7D)

4 实验

单域数据库EMODB、CASIA、SAVEE以及由它们构建的4个多域数据库Hybrid-CE、Hybrid-ES、Hybrid-CS、Hybrid-CES均未提供单独的训练数据和测试数据。本文采用说话人无关(Speaker-Independent，SI)策略进行训练：每类情感的所有样本随机等分为5份，将其中的4份作为训练数据，剩余的1份作为测试数据[29]。实验重复10次，采用平均准确率(Average Accuracy，AA)、平均精确率(Average Precision，AP)、平均未加权召回率(Unweighted Average Recall，UAR)以及平均F1-得分(Average F1-score，AF)表征模型的整体性能。此外，采用混淆矩阵分析单个情感类别的识别精度。

4.1 实验设置

实验采用一台CPU为40核80线程、内存为64 GB的高性能服务器进行计算，使用RTX 2080 Ti GPU进行模型训练，根据深度学习框架Keras[30]搭建模型。采用的优化器(Optimiser)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。

4.2 实验分析

本文主要进行了以下3个方面的实验：1)以单域语音情感识别为基线来验证多域语音情感识别的可行性；2)验证HMN模型的鲁棒性和泛化性；3)分析HMN模型在多域语音情感数据库上的性能。

4.2.1 多域语音情感识别的可行性验证

HMN模型在单域与多域数据库上进行实验得到的平均性能如表3所示。

表3 HMN模型在单域(基线)与多域语音情感数据库上的性能对比

Table 3 Performance comparison of HMN model on mono-domain (baseline) and multi-domain speech emotion database %

域	数据库	AA	AP	UAR	AF
单域	CASIA	85.42	85.69	85.45	85.57
	EMODB	84.11	87.41	81.64	84.43
	SAVEE	59.38	60.15	55.61	57.80
多域	Hybrid-CE	84.15	83.38	81.09	82.22
	Hybrid-ES	65.52	63.73	66.37	65.02
	Hybrid-CS	75.60	71.07	70.12	70.59
	Hybrid-CES	76.30	73.67	73.26	73.47

从表1可以看出：

1)在单域数据库上，HMN模型在CASIA库上的性能最优，EMODB次之，SAVEE最差。数据库之间存在的差异是导致模型在这些数据库上识别性能存在差异的主要原因，例如：CASIA库仅有6类情感，识别难度相对较低，而SAVEE数据库包含7类情感且样本较少，因此识别难度相对较高。

2)HMN模型在本文构建的4类多域语音情感数据库上均取得了较为可观的识别结果，表明多域情感识别是可行的。具体而言，模型HMN在Hybrid-CE库上性能最优，在Hybrid-CS、Hybrid-ES、Hybrid-CES库上性能较低，主要原因是这3个库中都包含了SAVEE库，而SAVEE库是一个视听双模态数据库，仅使用音频信息不能精确地表征情感。

与Hybrid-ES相比，在Hybrid-CS库上的准确率提升了18.63个百分点，原因是Hybrid-CS库仅包含7类情感，识别难度降低，且该库的样本数量多于Hybrid-ES库，模型能得到充分训练。

3)HMN模型在多域数据库上的性能略低于在单域数据库上的性能，主要原因是受情感类别数量和语言类型等因素的影响。

4)多域数据库Hybrid-ES、Hybrid-CS以及Hybrid-CES上的性能均优于SAVEE库，这是因为混合后的数据库大幅增加了训练样本数量，能够更好地训练模型。

4.2.2 HMN模型的鲁棒性和泛化性验证

利用HMN模型分别在3个单域数据库和4个多域数据库上进行10次实验，得到HMN在每个数据库上对应的箱线，如图2所示。其中，横坐标是7类数据库，纵坐标是准确率；在箱体的上方和下方各有一条线，分别表示一组数据中的最大值和最小值；箱体的高度在一定程度上反映了数据的波动程度；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上四分位数和下四分位数，这意味着箱体包含了50%的数据；实心圆圈表示异常值。

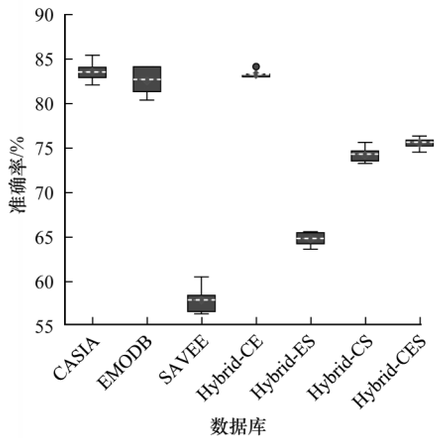

图2　HMN模型在多域数据库上的箱线图

Fig.2　Box-plot graph of HMN model on multi-domain database

从图2可以看出：1）对于3个单域数据库而言，模型在CASIA上的性能最高，而在SAVEE上的性能最差，平均性能最低，波动程度较大；2）在多域数据库Hybrid-CE、Hybrid-CES上，模型的波动程度较小，鲁棒性较好；3）无论是在单域数据库上还是在多域数据库上，模型HMN的性能均较好，表明该模型具有较好的泛化性。

图3利用AA、AP、UAR、AF 4个指标对HMN模型在4个多域数据库上的性能进行了较全面的对比。可以看出：1）在同一数据库上，无论在哪种评价指标下，HMN模型的性能相差均较小，表明模型鲁棒性较好；2）HMN模型在4种多域数据库上的性能均较好，尤其在Hybrid-CE数据库上的性能最好，表明HMN模型的泛化性较好。

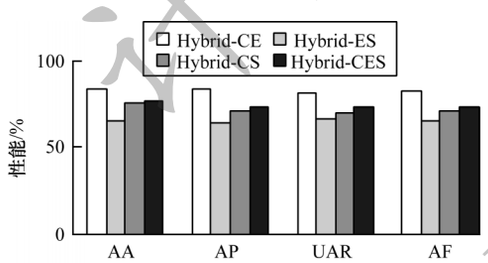

图3　HMN模型在多域数据库上识别性能对比

Fig.3　Identification performance comparison of HMN model on multi-domain database

4.2.3　HMN模型在多域语音情感库上的性能

下文利用混淆矩阵详细分析HMN模型对多域数据库Hybrid-CE、Hybrid-ES、Hybrid-CS以及Hybrid-CES中每类情感的识别性能。

图4所示为HMN模型在多域数据库Hybrid-CE上所获得的最佳混淆矩阵，其中，AA为84.15%，AP为83.38%，UAR为81.09%，AF为82.22%。可以看出：1）模型的平均准确率为84.15%；2）模型在其他类情感的召回率均达到了79.00%以上，而厌恶与无聊两类情感的召回率较低，因为在多域数据库Hybrid-CE中，各类情感样本数量不均衡，其中，厌恶类情感仅有60个样本，模型未得到充分训练；3）无聊类情感与中性易混淆，有33.33%的无聊类样本被预测为中性，主要原因是无聊和中性两类情感在效价维和激活维上取值较为接近，且两类情感的激活程度均较低。

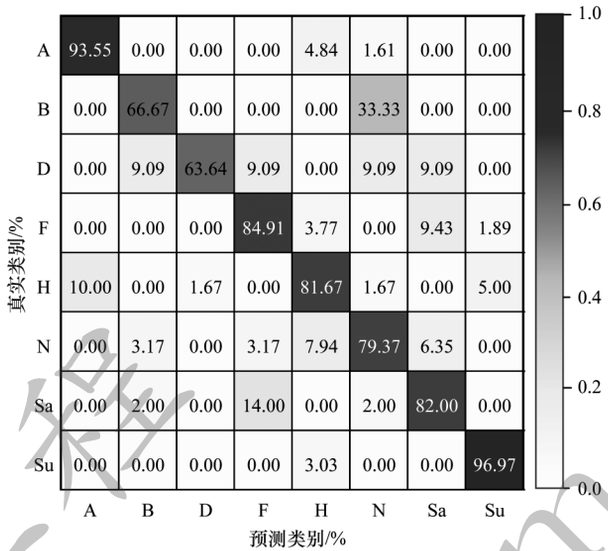

图4　HMN模型在Hybrid-CE数据库上的混淆矩阵

Fig.4　Confusion matrix of HMN model on Hybrid-CE database

图5所示为HMN模型在多域数据库Hybrid-ES上所获得的最佳混淆矩阵，其中，AA为65.52%，AP为63.73%，UAR为66.37%，AF为65.02%。可以看出：1）模型的平均准确率为65.52%；2）模型对恐惧类情感的识别率均较低；3）在多域数据库Hybrid-ES上，HMN模型的整体识别性能较低，主要是由SAVEE数据库引起的。

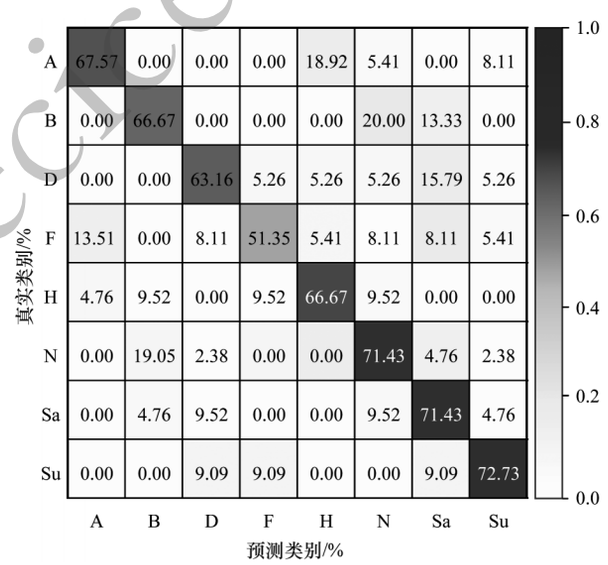

图5　HMN模型在Hybrid-ES数据库上的混淆矩阵

Fig.5　Confusion matrix of HMN model on Hybrid-ES database

图6所示为HMN模型在多域数据库Hybrid-CS上所获得的最佳混淆矩阵，其中，AA为75.60%，AP为71.07%，UAR为70.12%，AF为70.59%。可以看出：1）模型的平均准确率为75.60%；2）无聊类情感的识别率较低，仅为30.00%外，而其他类情感的识别

率均较为可观，主要原因是在多域数据库Hybrid-CS中，无聊类情感的样本较少，模型未能得到充分训练；3）在多域数据库Hybrid-ES中，HMN模型的整体识别性能较低，这仍然由SAVEE数据库引起的。

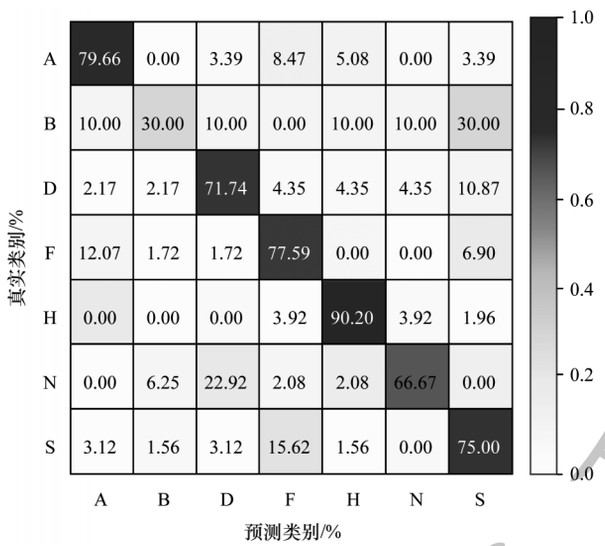

图6 HMN模型在Hybrid-CS数据库上的混淆矩阵

Fig.6 Confusion matrix of HMN model on Hybrid-CS database

图7所示为HMN模型在多域数据库Hybrid-CES上所获得的最佳混淆矩阵，其中，AA为76.30%，AP为73.67%，UAR为73.26%，AF为73.47%。可以看出：1）模型HMN的平均准确率为76.30%；2）厌恶类情感的识别率最低，仅有52.94%；3）与由两种语言混合的多域数据库Hybrid-CE、Hybrid-ES、Hybrid-CS相比，模型HMN在3种语言混合的多域数据库Hybrid-CES上的性能有所提升，这是因为该库包含的情感样本数增加，能够更好地训练模型。

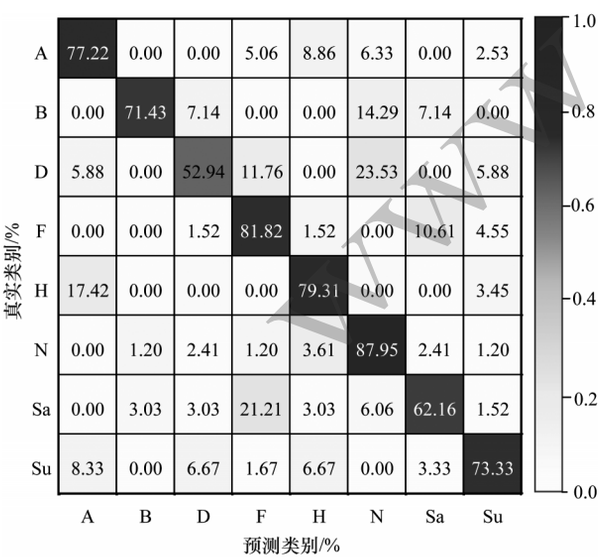

图7 HMN模型在Hybrid-CES数据库上的混淆矩阵

Fig.7 Confusion matrix of HMN model on Hybrid-CES database

总地来说，与作为基线的单域语音情感识别相比，多域语音情感识别因为情感类别数的增加导致区分难度加大，但本文提出的HMN模型在多域数据库上仍取得了较好的识别结果。

5 结束语

本文设计一种基于多操作网络的图式多域语音情感识别模型。通过3种单域数据库CASIA、EMODB、SAVEE构建多域语音情感数据库Hybrid-CE、Hybrid-ES、Hybrid-CS以及Hybrid-CES，在多域数据库上计算219维的高级统计特征作为层级多操作网络模型的输入，并在单域与多域数据库上对比HMN模型的识别性能、鲁棒性和泛化性。实验结果表明，该模型在4种多域数据库上均具有较高的识别性能。下一步将采用HMN模型在维度情感数据库上研究多域和跨域语音的情感识别。

参考文献

[1] TOOBY J, COSMIDES L. Evolutionary psychology and the emotions and their relationship to internal regulatory variables[M]. London, UK: The Guilford Press, 2008.

[2] LAZARUS R S. Emotion and adaptation: conceptual and empirical relations[C]//Proceedings of Nebraska Symposium on Motivation. Lanham, USA: University of Nebraska Press, 1968: 175-266.

[3] KRAGEL P A, REDDAN M C, LABAR K S, et al. Emotion schemas are embedded in the human visual system[J]. Science Advances, 2019, 5(7): 43-58.

[4] KOIRALA A, YU Z W, SCHILTZ H, et al. A preliminary exploration of virtual reality-based visual and touch sensory processing assessment for adolescents with autism spectrum disorder[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2021, 29(1): 619-628.

[5] DING K Q, DRAGOMIR A, BOSE R, et al. Sensory stimulation enhances functional connectivity towards the somatosensory cortex in upper limb amputation[C]// Proceedings of the 10th International IEEE/EMBS Conference on Neural Engineering. Washington D. C., USA: IEEE Press, 2021: 226-229.

[6] RUSSELL J A. Core affect and the psychological construction of emotion[J]. Psychological Review, 2003, 110(1): 145-172.

[7] ZAPARA T, ROMASHCHENKO A, PROSKURA A, et al. Mechanisms and functions of neurogenesis in the limbic system of adult animals[C]//Proceedings of Cognitive Sciences, Genomics and Bioinformatics. Washington D. C., USA: IEEE Press, IEEE Press, 2020: 174-179.

[8] PARADISO S. Affective neuroscience: the foundations of human and animal emotions[J]. American Journal of Psychiatry, 2002, 159(10): 1805.

[9] LIU T J, LI F L, JIANG Y, et al. Cortical dynamic causality network for auditory-motor tasks[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017, 25(8): 1092-1099.

[10] 余莉萍，梁镇麟，梁瑞宇. 基于改进LSTM的儿童语音情感识别模型[J]. 计算机工程，2020，46(6)：40-49.

YU L P，LIANG Z L，LIANG R Y. Emotion recognition model for children speech based on improved LSTM[J]. Computer Engineering，2020，46(6)：40-49. (in Chinese)

[11] NEUMANN M，THANG VU N G. CRoss-lingual and multilingual speech emotion recognition on English and French[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，IEEE Press，2018：5769-5773.

[12] SCHULLER B，VLASENKO B，EYBEN F，et al. Cross-corpus acoustic emotion recognition：variances and strategies[J]. IEEE Transactions on Affective Computing，2010，1(2)：119-131.

[13] FERARU S M，SCHULLER D，SCHULLER B. Cross-language acoustic emotion recognition：an overview and some tendencies[C]//Proceedings of International Conference on Affective Computing and Intelligent Interaction. Washington D. C.，USA：IEEE Press，2015：125-131.

[14] SAGHA H，MATĚJKA P，GAVRYUKOVA M，et al. Enhancing multilingual recognition of emotion in speech by language identification[C]//Proceedings of IEEE ISCA'16. Washington D. C.，USA：IEEE Press，2016：2949-2953.

[15] CHIOU B C，CHEN C P. Speech emotion recognition with cross-lingual databases[C]//Proceedings of IEEE ISCA'14. Washington D. C.，USA：IEEE Press，2014：558-561.

[16] ELGAAR M，PARK J，LEE S W. Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2020：7769-7773.

[17] ZHANG J C，JIANG L，ZONG Y，et al. Cross-corpus speech emotion recognition using joint distribution adaptive regression[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2021：3790-3794.

[18] 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.

[19] TAO F，LIU G. Advanced LSTM：a study about better time dependency modeling in emotion recognition[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2018：2906-2910.

[20] SEPAS-MOGHADDAM A，ETEMAD A，PEREIRA F，et al. Facial emotion recognition using light field images with deep attention-based bidirectional LSTM[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2020：3367-3371.

[21] PENG Z X，LU Y，PAN S F，et al. Efficient speech emotion recognition using multi-scale CNN and attention[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2021：3020-3024.

[22] 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.

[23] BLASZKE M，KOSZEWSKI D. Determination of low-level audio descriptors of a musical instrument sound using neural network[C]//Proceedings of Signal Processing：Algorithms，Architectures，Arrangements，and Applications. Washington D. C.，USA：IEEE Press，2020：138-141.

[24] WANG X，DU P J，CHEN D M，et al. Change detection based on low-level to high-level features integration with limited samples[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2020，13：6260-6276.

[25] WEN X C，LIU K H，ZHANG W M，et al. The application of capsule neural network based CNN for speech emotion recognition[C]//Proceedings of the 25th International Conference on Pattern Recognition. Washington D. C.，USA：IEEE Press，2021：9356-9362.

[26] FU C Z，LIU C R，ISHI C T，et al. An end-to-end multitask learning model to improve speech emotion recognition[C]//Proceedings of the 28th European Signal Processing Conference. Berlin，Germany：Springer，2021：1-5.

[27] GKIOKAS A，KATSOUROS V，CARAYANNIS G. Towards multi-purpose spectral rhythm features；an application to dance style，meter and tempo estimation[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2016，24(11)：1885-1896.

[28] SHAH A F，ANTO P B. Hybrid spectral features for speech emotion recognition[C]//Proceedings of International Conference on Innovations in Information，Embedded and Communication Systems. Washington D. C.，USA：IEEE Press，2017：1-4.

[29] CAO W H，XU J P，LIU Z T. Speaker-independent speech emotion recognition based on random forest feature selection algorithm[C]//Proceedings of the 36th Chinese Control Conference. Dalian，China：[s. n.]，2017：10995-10998.

[30] YANG X，ROOP P，PEARCE H，et al. A compositional approach using Keras for neural networks in real-time systems[C]//Proceedings of Design，Automation & Test in Europe Conference & Exhibition. Washington D. C.，USA：IEEE Press，2020：1109-1114.

编辑　索书志