doi：10.12068/j.issn.1005－3026.2024.01.006

融合功能性副语言比例系数的语音情感识别

孙 颖，周雅茹，张雪英
（太原理工大学 信息与计算机学院，山西 太原 030024）

摘 要：语言中的非言语发声如笑声、叹息、抽泣等，称为功能性副语言，对情感表达起重要作用，但现有研究很少考虑多种功能性副语言在一种情感中的协同作用．针对该问题，提出了融合功能性副语言比例系数（functional paralanguage proportion coefficient，FPPC）的情感识别系统．首先，提取能体现多种功能性副语言在情感语句中出现的频率快慢和持续时间长短的FPPC特征；然后，搭建基于注意力机制的集成学习（attention stacking）为不同的基分类器赋予不同权重，并对FPPC特征进行训练；最后，通过自适应熵权重决策融合方法将传统语音情感识别与基于FPPC特征情感识别进行融合．实验结果显示，融合了FPPC特征后的情感识别结果提高了16.84%，证明融合FPPC特征能有效提高系统整体识别率．

关 键 词：语音情感识别；比例系数；功能性副语言；注意力机制；自适应熵权重决策融合
中图分类号：TN 912.3 **文献标志码**：A **文章编号**：1005-3026(2024)01-0040-09

Speech Emotion Recognition Fusing Functional Paralanguage Proportion Coefficient

SUN Ying, *ZHOU Ya-ru*, *ZHANG Xue-ying*
（College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China. Corresponding author: ZHANG Xue-ying, E-mail: tyzhangxy@163.com）

Abstract: Nonverbal vocalizations such as laughter, sighs, and sobs in speech are called functional paralanguage and play an important role in emotional expression. However, existing research has rarely considered the synergistic effect of multiple functional paralanguages in a single emotion. To address this issue, an emotion recognition system integrating functional paralanguage proportion coefficients (FPPC) is proposed. Firstly, FPPC features that reflect the frequency and duration of multiple functional paralanguages appearing in emotional statements are extracted. Then, an attention mechanism-based ensemble learning is constructed to assign different weights to different base classifiers and train the FPPC features. Finally, the adaptive entropy weight decision fusion method is used to fuse traditional speech emotion recognition with emotion recognition based on FPPC features. Experimental results show a 16.84% improvement in emotion recognition after integrating FPPC features, proving that integrating FPPC features can effectively improve the overall recognition rate of the system.

Key words: speech emotion recognition; proportion coefficient; functional paralanguage; attention mechanism; adaptive entropy weight decision fusion

随着人工智能的快速发展，语音情感识别被广泛应用于生活的各个方面．目前通过传统语音信号获得人类情感状态已取得许多成果[1-2]．但现实生活中的语音信号通常会包含携带大量情感信息的功能性副语言[3]，如一个人高兴时会发出笑声，悲伤时会伴随哭泣，一条语句中可能会包含不同种类的功能性副语言，同一种功能性副语言在不同情感状态下所包含的信息也会有所不同．但以往研究发现，包含功能性副语言的语句由于特征的突发性会导致整体情感识别率下

收稿日期：2022-07-22
基金项目：国家自然科学基金资助项目(62271342)；山西省自然科学基金资助项目(201901D111096)．
作者简介：孙 颖(1981-)，女，山西太原人，太原理工大学副教授，博士；张雪英(1964-)，女，河北行唐人，太原理工大学教授，博士生导师．

降[4],功能性副语言所携带的情感信息没有被有效利用. 因此,构建能有效融合功能性副语言信号与传统语音信号的系统并进行情感识别是语音情感识别领域亟待突破的关键技术之一.

目前针对功能性副语言的研究主要集中在功能性副语言的识别检测与实际应用两方面. 在功能性副语言识别检测方面,笑声[5]、呼吸声[6]和哭声[7]等已被成功检测,所使用的方法涵盖了机器学习和深度学习. Huang等[8]将提取的深度特征输入基于注意长短时记忆的序列–序列模型,检测准确率比传统方法提高了52. 0%. Knox等[9]将低级声学特征通过神经网络对笑声进行检测,达到了92. 1%的准确率. 赵小蕾等[10]通过基于定长分段的功能性副语言检测模型与基于静音帧的分割点确认算法将功能性副语言提取出来. 在功能性副语言的应用中,Laguarta等[11]通过语谱图实现从咳嗽记录中预筛选新冠患者. Schuller等[12]基于呼吸、干咳、湿咳、打喷嚏等声音研究不同声音类型下新冠患者的身体状况,为研究新冠疫情提供了更加丰富的依据. Kaya等[13]将功能性副语言中的沉默与呼吸声用于检测抑郁症. 以上方法都实现了对功能性副语言进行检测与应用,但功能性副语言中蕴含的情感信息没有被有效表征,针对功能性副语言的语音情感识别也仅限于单一种类的功能性副语言对应单一情感,如笑声对应高兴、哭声对应悲伤等,并未考虑到多种功能性副语言之间存在的协同作用对最终情感的影响.

针对以上问题,本文提取能体现多种功能性副语言协同作用的比例系数(functional paralanguage proportion coefficient, FPPC)特征;然后使用基于注意力机制的Stacking集成学习(attention stacking, ATT_Stacking)对FPPC特征进行训练;最后使用自适应熵权重决策融合的方法将传统语音情感识别与基于FPPC特征情感识别进行融合,由对比实验验证了FPPC特征的有效性. 证明了融合FPPC特征的语音情感识别可以提高系统整体情感识别率,能构建出更有效的情感识别系统.

1　系统框架

图1给出了系统结构框图. 系统主要由三部分构成:一是传统语音信号情感识别通道;二是功能性副语言情感识别通道;三是自适应熵权重决策融合. 在传统语音情感识别通道,首先对传统语音信号使用OpenSMILE[14]提取IS09情感特征,然后将特征输入到Xgboost(eXtreme gradient boosting)中进行情感识别. 在功能性副语言情感识别通道,对功能性副语言信号提取FPPC特征,使用ATT_Stacking对FPPC特征进行训练. 在得到功能性副语言与传统语音信号的情感识别结果后,采用自适应熵权重决策融合的方法进行决策融合得到最终识别结果.

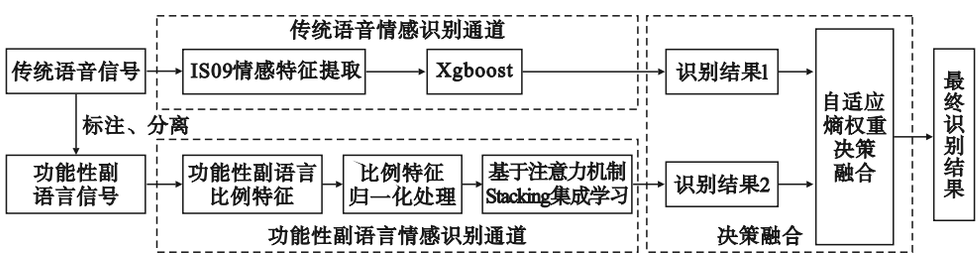

图1　系统结构框图

Fig. 1　System structure block diagram

2　功能性副语言比例系数情感识别

2.1　FPPC特征提取

不同种类的功能性副语言在同一信号中的比例信息蕴含着情感的变化,这种比例信息体现在功能性副语言持续时间长短和出现频次高低等方面,如抽泣在悲伤这一情感中持续时间较长、出现频次较高,笑声在高兴这一情感中持续时间较长、出现频次较高. 各种功能性副语言之间比例关系的改变会伴随情感的变化,通过分析FPPC特征的变化情况,可以实现对应语句特征提取.

2.2　持续时间比例特征

针对功能性副语言在传统语音信号中持续时间长短所蕴含的情感信息,并考虑到不同语料库对时间信息标注的标准不同,本文在提取持续时间比例特征时,首先统计不同种类的功能性副语言在一段对话中总的持续时间,然后分别计算

不同功能性副语言在该段对话中的时间占比. 设数据集中语音信号的集合为$X=\{x_1,x_2,\cdots,x_N\}$,其中N为样本数量,对于任一样本x_i,其包含的功能性副语言集合为$D_i=\{d_{i,1},d_{i,2},\cdots,d_{i,M}\}$,其中$M$为功能性副语言的种类数,每种功能性副语言的持续时间集合为$T_i=\{t_{i,1},t_{i,2},\cdots,t_{i,M}\}$,则$x_i$中功能性副语言$d_{i,j}$所占持续时间比例系数$b_{i,j}$定义为

$$b_{i,j}=\frac{t_{i,j}}{T_i}. \tag{1}$$

式中:$t_{i,j}$为功能性副语言$d_{i,j}$在信号x_i中持续时间;T_i为信号x_i声波振动的持续时间. 将信号x_i中不同种类的功能性副语言持续时间比例系数进行汇总可得

$$B_i=\{b_{i,1},b_{i,2},\cdots,b_{i,M}\}. \tag{2}$$

式中,B_i为信号x_i的持续时间比例特征.

2.3 频次比例特征

同种类的功能性副语言在同一信号中出现频次的高低同样影响情感的变化,针对这种影响,本文通过提取功能性副语言频次比例特征来进行表征,首先统计不同种类的功能性副语言在一段对话中总的出现频次,然后分别计算不同功能性副语言在该段对话中的频次占比. 设功能性副语言$d_{i,j}$在信号x_i中出现频次比例为$r_{i,j}$,具体定义为

$$r_{i,j}=\frac{n_{i,j}}{N_i}. \tag{3}$$

式中:$n_{i,j}$为功能性副语言$d_{i,j}$在信号x_i中出现频次;N_i为信号x_i中所有功能性副语言出现频次. 将信号x_i中不同种类的功能性副语言频次比例系数进行汇总可得频次比例特征R_i:

$$R_i=\{r_{i,1},r_{i,2},\cdots,r_{i,M}\}. \tag{4}$$

本文选取13个常用的统计函数应用于上述特征,得到总的FPPC特征,如表1所示.

表1 功能性副语言比例系数

Table 1 Functional paralanguage proportion coefficient

比例特征	统计特征
持续时间	最大值/最小值
	最大值位置/最小值位置
	第一/二/三/分位数
	均值
	平均绝对偏差
频次	标准偏差
	偏度、峰度、方差

3 基于注意力机制的Stacking集成学习

图2为本文搭建的基于注意力机制的Stacking集成学习总体框图,分为三部分:首先基模型对原始数据交叉验证训练得到新特征;然后使用注意力机制为不同基模型的特征赋予不同权重;最后输入到元模型中进行识别.

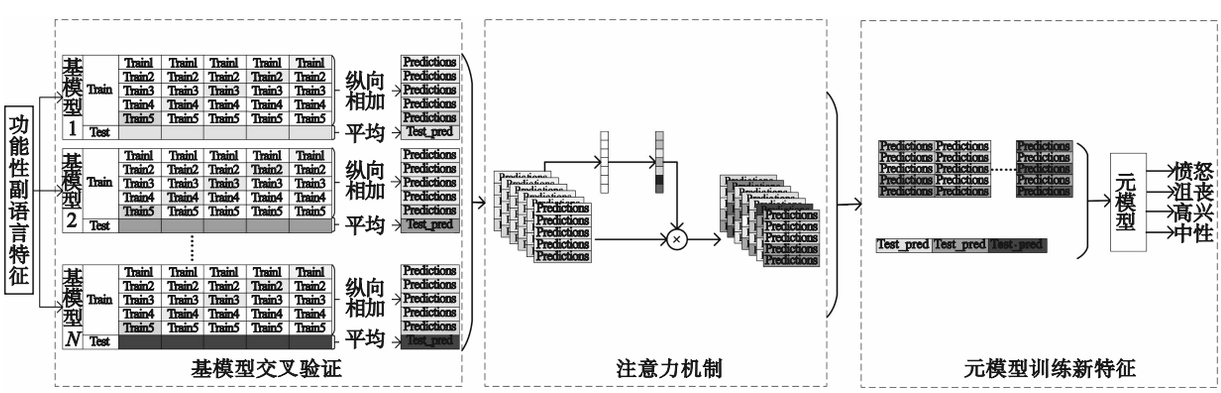

图2 基于注意力机制的Stacking集成学习总体框图

Fig. 2 Overall block diagram of Stacking ensemble learning based on attention mechanism

3.1 基、元模型的选择与构建

考虑FPPC特征数据结构,本文选择建立在统计学习理论基础上的Stacking集成学习. Stacking集成学习通过元模型将多个基模型对数据预测的结果进行再训练得到最终结果,在选择基、元模型时,模型之间相互独立,差异性越大越好,以此实现模型间信息有效融合与互补,本文选择的基、元模型如表2所示.

3.2 注意力机制

不同基模型对特征的处理能力不同,为实现多个基模型之间的优势互补,受通道注意力机制[15]的启发,本文对基模型引入注意力机制,通道注意力机制为不同通道赋予不同权重,而基模型注意力机制旨在为不同的基模型赋予相应的

表2　基、元模型的选择

Table 2　Selection of base model and meta model

模型选择		优点
基模型	KNN	时间复杂度低
	RF	抗过拟合能力强
	GBDT	适合低维数据
	Adaboost	精度高
	Extra Trees	泛化能力好
	LightGBM	训练速度快
元模型	SVM	非线性映射、泛化好

权重,具体操作分为 Squeeze 和 Excitation 两部分,如图3所示.

对基模型输出结果进行拼接得到总体输出维数为 $B\times C\times N$,其中 B,N 和 C 分别为基分类器数、样本数和样本类别数. 首先进行 Squeeze 操作,包括对输入特征进行最大池化 $S_{\max(N,C)}$ 和平均池化 $S_{\text{avg}(N,C)}$,并将两者拼接得到新的特征描述 $S_{\text{sq}}\in\mathbf{R}^{1\times1\times B}$;然后对 S_{sq} 执行 Excitation 操作,依次经过全连接层与激活层得到维数为 $1\times1\times B$ 的基模型注意力权重. 该过程具体可表示为

$$S_{\text{sq}}=F_{\text{sq}}(x)=S_{\max(N,C)}+S_{\text{avg}(N,C)}=\max_{N}\left(\max_{C}\left(x_B\right)\right)+\frac{1}{N\times C}\sum_{i=1}^{N}\sum_{j=1}^{C}x_B(i,j),\quad(5)$$

$$S=F_{\text{ex}}\left(S_{\text{sq}}\right)=\sigma\left(g\left(S_{\text{sq}},\boldsymbol{W}\right)\right)\sigma\left(\boldsymbol{W}_2\delta\left(\boldsymbol{W}_1,S_{\text{sq}}\right)\right).\quad(6)$$

式中:x_B 为输入信号;$\boldsymbol{W}_1$,$\boldsymbol{W}_2$ 为全连接层映射矩阵;$\boldsymbol{W}$ 为 $\boldsymbol{W}_1$,$\boldsymbol{W}_2$ 总写.

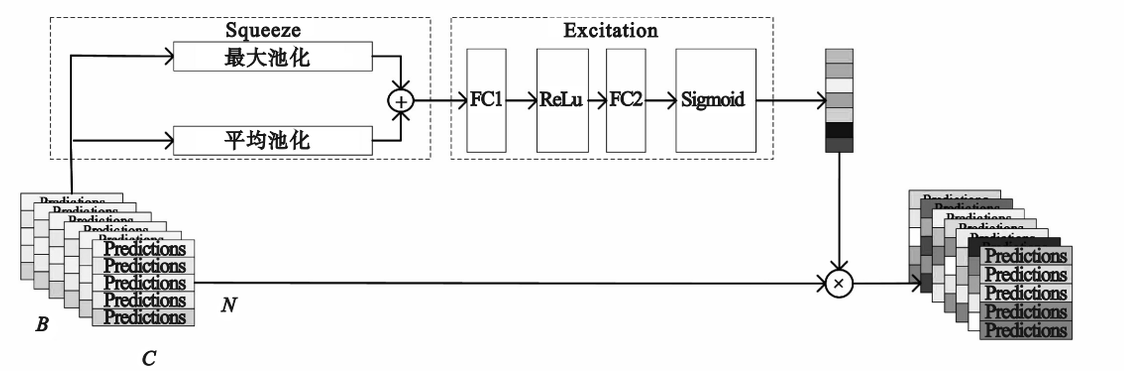

图3　基模型注意力机制

Fig. 3　Base model attention mechanism

4　自适应熵权重决策融合

为解决不同分类器输出概率最大值没有出现在同一类别的问题,引入一种自适应熵权重决策融合方法[16],目的是客观地衡量不同类别之间的差异性. 具体地,各类预测概率值之间的差别越小,则分类不确定性就越大,赋予的权重应越小,如果最大预测概率值与其他预测概率值的差别越大,说明分类的结果越可靠,赋予的权重也应越大. 因此一条语句的分类结果就由两个分类器的输出概率进行自适应熵权重决策融合得到. 其计算步骤如下.

1) 每个样本的输出概率矩阵为

$$\boldsymbol{P}(x)=\begin{bmatrix}p_{11}(x) & \cdots & p_{1n}(x)\\ p_{21}(x) & \cdots & p_{2n}(x)\end{bmatrix}.\quad(7)$$

式中:p_{1j} 为传统语音情感识别通道对输入样本 x 的概率输出值;p_{2j} 为功能性副语言语音情感识别通道对输入样本 x 的概率输出值;n>1 为情感种类数目;$\boldsymbol{P}(x)$ 为每个样本 x 的输出概率矩阵.

2) 引入信息熵[17]判定每个分类器对输入样本分类的不确定性,分类结果的信息熵值 $e_i(x)$ 为

$$e_i(x)=-\frac{1}{\ln n}\sum_{j=1}^{n}p_{ij}(x)\ln p_{ij}(x),i=1,2.\quad(8)$$

式中,$p_{ij}(x)$ 为第 i 个分类器将输出样本归为第 j 类情感的概率. 由式(8)可知,当同一分类器对样本 x 概率输出结果之间差距越大,分类不确定性越小,其信息熵就越小,该分类器对样本 x 的分类能力越强,得到的决策融合权重应越大,反之亦然.

3) 确定自适应熵权重的计算公式为

$$\omega_i(x)=\frac{1-e_i(x)}{\sum_{i=1}^{2}\left(1-e_i(x)\right)}.\quad(9)$$

式中,$\omega_i(x)$ 为样本 x 对应每个通道的自适应熵权重.

4) 得到融合权重后对原概率输出矩阵 $\boldsymbol{P}(x)$ 每一行都乘以对应权重并对变换后的矩阵按列求和,其中最大值对应的标签为自适应熵权重决策融合的结果,计算公式为

$$\text{label}(x)\underbrace{\arg\max}_{j=1,2,\cdots,n}\left(\sum_{i=1}^{3}\omega_i p_{ij}\right). \tag{10}$$

式中,label(x)为样本x的最终标签值.

自适应熵权重决策融合算法考虑了不同分类器对同一样本分类性能的差异性,并根据分类结果的可靠性为每种分类结果赋予不同的权重,因此本文选择自适应熵权重决策融合算法融合不同通道的识别结果.

5 实 验

5.1 实验数据库

为验证融合功能性副语言的语音情感识别系统的识别效果,本文选用NNIME中文交互式多模态情感语料库[18],该数据库中共有102组二元互动对话,包含的功能性副语言有笑声、抽泣声、叹气声、叫喊声以及观众笑声、背景噪声和沉默.本文选择更能表征声音突发特征的4种功能性副语言:笑声、叫喊声、叹息声和抽泣声.经整理,每种情感下所包含的上述4种功能性副语言具体情况如表3所示.

表3 NNIME中功能性副语言分布情况

Table 3 Distribution of functional paralinguistics in NNIME

情感	功能性副语言			
	笑声	叫喊声	叹息声	抽泣声
愤怒	13	4	29	2
沮丧	18	25	23	9
高兴	108	19	6	0
中性	39	43	27	1
悲伤	10	5	44	102
惊喜	59	18	34	17

5.2 模型建立与评估

本文使用准确率(Acc)、精确率(Precision)、召回率(Recall)和$F1$以及混淆矩阵来评价语音情感识别效果.Acc表示总样本中被正确分类的样本比例;Precision表示预测与实际同为正例占预测总正例的比例,体现模型对负样本的区分能力;Recall表示预测为正样本占实际正样本的比例,体现模型对正样本的识别能力;$F1$为精确率与召回率的一个加权平均,表达模型的稳健程度.具体计算公式为

$$\text{Acc}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}, \tag{11}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{12}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{13}$$

$$F1=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{14}$$

式中:TP为被正确预测的正例;TN为被正确预测的反例;FP为被错误预测的正例;FN为被错误预测的反例.

混淆矩阵表达的是在多分类问题中出现的分类混淆情况,纵轴为样本实际类别,横轴为样本预测类别,对角线为每类情感被正确分类的概率.

本文采用Python 3.6进行模型搭建与训练.在功能性副语言情感识别通道,将数据集按照随机抽样的方法以7:3的比例划分为训练集和测试集,在此基础上对训练集使用五折交叉验证并输入到模型中进行训练,使用网格搜索确定基模型与元模型的最优参数.在传统语音情感识别通道,考虑到样本数量较少,不适合对复杂模型进行训练,因此选用Xgboost模型,具体模型参数如表4所示.

表4 各模型主要参数网格搜索结果

Table 4 Grid search results of main parameters of each model

模型	参数	数值或描述
KNN	K-近邻点	5
RF	决策树数量	400
	最大特征数	15
	分支所需最小样本数	5
GBDT	学习率	0.1
	最大深度	10
	最大特征数	10
Adaboost	决策树数量	200
	最大深度	13
ExtRa Tree	决策树数量	400
	分支所需最小样本数	10
LightGBM	最小损失减少值	0.37
	最小样本权重和	0.0001
SVM	核函数	Poly
	核函数系数	0.3
	核的阶数	3
Xgboost	最大树深度	5
	学习率	0.1
	决策树数量	200
	特征采样的比例	0.75

5.3 实验方案与实验结果

为验证融合功能性副语言语音情感识别系统的有效性,文中分别从验证FPPC特征有效性、验证基注意力Stacking模型有效性、验证Xgboost在识别IS09特征上的有效性和验证自适应熵权

重决策融合有效性4个方面设计实验.

方案一:验证FPPC特征的有效性. 本文使用IS09特征为对比,使用相同的模型对IS09和FPPC特征分别进行训练,包括基准模型SVM,Xgboost,Stacking集成模型和ATT_Stacking模型,并且将IS09和FPPC在进行简单的特征串联后,输入到上述模型中进行训练,并将每个模型的实验结果与ATT_Stacking模型的实验结果进行对比,得到改进增幅,实验结果如表5所示.

表5　不同特征与分类方法组合的识别结果

Table 5　Recognition results of different combinations of features and classification methods　%

特征	评价标准	SVM	Xgboost	Stacking	ATT_Stacking	改进增幅		
IS09	Precision	40.00	74.70	59.46	66.19	+26.19	−8.51	+6.73
	Recall	44.44	66.94	60.23	65.28	+20.84	−1.66	+5.05
	*F*1	40.36	66.44	58.84	64.90	+24.54	−1.54	+6.06
	Acc	48.07	70.00	61.50	63.33	+15.26	−6.67	+1.83
FPPC	Precision	43.67	53.67	50.20	52.00	+10.54	+0.54	+4.01
	Recall	45.40	54.40	49.61	53.64	+8.24	−0.76	+4.03
	*F*1	40.36	52.30	47.90	53.16	+12.8	+0.86	+5.26
	Acc	47.96	52.96	53.33	55.67	+7.71	+2.71	+2.34
IS09+FPPC	Precision	53.71	55.54	69.44	66.90	+13.19	+11.36	-2.54
	Recall	56.16	66.50	70.14	72.92	+16.76	+6.42	+2.78
	*F*1	54.58	57.68	68.36	73.15	+18.57	+15.47	+4.79
	Acc	57.69	64.44	73.33	76.67	+18.98	+12.23	+3.34

表5的结果显示,从单一特征的角度分析,IS09特征在各个模型上的性能均优于FPPC特征,特别是在Xgboost模型上,优势更加明显,准确率可达70.00%,而FPPC特征在ATT_Stacking模型上的识别性能高于其他模型,准确率可达55.67%,相比于其他模型最高可提升7.71%,这是由于FPPC特征数据结构简单,更适合集成学习;从特征融合的角度分析,融合后的特征在ATT_Stacking模型上训练效果最佳,准确率最高可达76.67%,说明FPPC特征的加入在一定程度上可以弥补IS09中对情感信息表征的不足.

方案二:验证ATT_Stacking模型在FPPC特征上的有效性. 将ATT_Stacking模型与单一模型作对比,并将每种模型的识别准确率与ATT_Stacking模型进行对比得到改进增幅,实验结果如表6所示.

表6结果显示,ATT_Stacking集成模型的结果要比单一算法得到的结果更好,表现为与单一模型中准确率最高的Xgboost相比,ATT_Stacking的准确率提高2.71%,该结果说明ATT_Stacking模型可以准确地表达出信号的情感差异度,对FPPC特征识别效果更好. ATT_Stacking与未加基模型注意力机制的Stacking模型相比,ATT_Stacking的分类准确率、精确率以及体现模型稳定性的*F*1都比Stacking模型效果更好,说明基模型注意力机制的加入可以有效提高系统识别率.

方案三:验证Xgboost在识别IS09特征上的有效性. 本文在NNIME数据集上提取IS09特征,然后使用LSTM[19]和当前流行深度学习模型BLSTM,1D−CNN进行对比实验,并将每种模型得到的准确率与Xgboost模型得到的准确率进行对比,计算得到相应增幅,实验结果如表7所示,图4为Xgboost模型对IS09特征的分类结果混淆矩阵.

表7中结果显示,Xgboost的分类准确率、精确率以及体现模型稳定性的*F*1都比其他分类模型效果更好,平均识别率可达70%,这是因为LSTM和BLSTM更适合对时序信息进行分析,而1D−CNN则适用于存在空间关系的特征,IS09并不符合这些要求,而Xgboost能够充分学习到IS09所表示的情感信息,可以更准确地表达出信号的情感差异度. 从图4混淆矩阵可知,愤怒、沮丧、高兴和惊喜这4种情感的分类准确率相对较高,可达63%以上,而中性、悲伤的情感识别率相对较低,混淆主要发生在高兴和惊喜之间、中性和惊喜之间、悲伤和沮丧之间,借用情感维度空间模型进行分析,高兴和惊喜都属于高激活度的情感,空间分布区域相近,因此容易造成误判;悲伤和沮丧都属于低激活度的情感,也容易产生误判,而中性和惊喜之间的混淆则可能与选用的数据集有关,该数据集的录制人员在这两种情感上

表 6 不同分类方法在 FPPC 上的识别结果

Table 6 Recognition results of different classification methods on FPPC %

分类方法	Precision	Recall	*F*1	Acc	改进增幅
GBDT	42. 36	41. 33	41. 84	43. 33	+12. 34
KNN	43. 50	43. 46	42. 47	43. 33	+12. 34
Xgboost	53. 67	54. 40	52. 30	52. 96	+2. 71
RF	34. 45	33. 56	32. 09	36. 67	+19. 00
ET	41. 72	40. 77	40. 12	43. 33	+12. 34
LightGBM	42. 47	40. 21	41. 31	40. 00	+15. 67
SVM	43. 67	45. 40	40. 36	47. 96	+7. 71
LSTM	46. 17	45. 64	43. 16	44. 56	+11. 11
Stacking	50. 20	49. 61	47. 90	53. 33	+2. 34
ATT_Stacking	52. 00	53. 64	53. 16	55. 67	0. 00

表 7 不同方法在 IS09 上的识别结果

Table 7 Recognition results of different methods on IS09 %

分类方法	Precision	Recall	*F*1	Acc	改进增幅
LSTM[19]	—	—	38. 00	51. 12	+18. 88
BLSTM	50. 25	51. 39	49. 88	53. 33	+16. 67
1D_CNN	33. 17	57. 00	38. 26	43. 75	+26. 25
GBDT	60. 21	61. 39	59. 35	66. 67	+3. 33
RF	54. 95	59. 31	55. 65	63. 33	+6. 67
SVM	40. 00	44. 44	40. 36	48. 07	+21. 93
Xgboost	74. 70	66. 94	66. 44	70. 00	0. 00

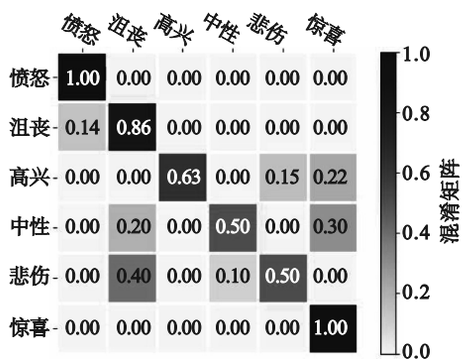

图 4 Xgboost 模型分类结果混淆矩阵

Fig. 4 Confusion matrix of Xgboost model classification results

的表达可能存在相似.

方案四:使用自适应熵权重决策融合将传统语音信号与基于 FPPC 特征识别结果进行融合,证明自适应熵权重决策融合有效性. 此外,增加同样使用 NNIME 数据集考虑功能性副语言情感识别的文献[19]和文献[8]进行对比,文献[19]使用 Resnet18 网络分别对功能性副语言和传统语音信号提取深度特征,并将得到的两种特征进行直接串联输入到 LSTM 模型中进行情感识别.

为便于表示,实验一为基于传统语音信号情感识别结果,即使用 Xgboost 对 IS09 进行情感识别;实验二为基于 FPPC 特征情感识别结果,即使用 ATT_Stacking 集成模型对 FPPC 特征进行情感识别;实验三为对 IS09 和 FPPC 进行简单的特征串联后使用 ATT_Stacking 集成模型进行情感识别;实验四为文献[19]的实验结果;实验五为使用自适应熵决策融合对实验一和实验二的识别结果进行决策融合得到的最终结果,实验结果如表 8 所示.

表 8 结果显示,无论是特征融合还是决策融合,在融合 FPPC 特征后情感识别结果都有了一定程度的提升. 与传统语音情感识别(实验一)和基于功能性副语言情感识别(实验二)相比,自适应熵权重决策融合(实验五)实验结果的精确率、召回率、*F*1 还是准确率都有了一定程度的提升,其中准确率分别提高 16. 84% 和 31. 17%;与简单拼接的特征融合(实验三)相比,自适应熵权重决策融合(实验五)从决策层的角度分析,更能挖掘数据中包含的情感信息,准确率提高 10. 17%;实验四在提取功能性副语言和传统语音信号深度特征后输入到 LSTM 和 BLSTM-LSTM 编解码结构中进行识别,对比本文使用的方法,LSTM 可以利用特征的时序信息,但由于数据集样本数量较少而网络结构复杂,因此识别效果有

所下降. 综上所述,自适应熵权重决策融合的方法可以对不同的识别方法进行有效融合,以提高系统整体识别率.

表 8　自适应熵权重决策融合实验结果

Table 8　Experimental results of adaptive entropy weight decision fusion　%

实验方案	评价标准	愤怒	沮丧	高兴	中性	悲伤	惊喜	均值	Acc	改进增幅
实验一	Precision	100.00	85.71	62.50	50.00	50.00	100.00	74.70	70.00	+16.84
	Recall	83.33	55.56	65.74	55.56	61.45	80.00	66.94		
	*F*1	80.00	71.43	71.43	51.39	54.50	68.65	66.44		
实验二	Precision	33.33	66.67	57.14	43.33	50.00	61.54	52.00	55.67	+31.17
	Recall	42.11	61.54	50.00	47.06	60.00	61.13	53.64		
	*F*1	37.50	63.33	54.55	44.44	54.55	64.59	53.16		
实验三	Precision	40.00	100.00	83.33	40.00	66.67	71.43	66.90	76.67	+10.17
	Recall	33.33	87.50	100.0	56.00	77.36	83.33	72.92		
	*F*1	40.00	93.33	100.0	50.00	88.89	66.68	73.15		
实验四	Precision	57.00	56.00	75.00	44.00	86.00	56.00	62.33	61.92	+24.92
	Recall	60.00	49.00	64.00	65.00	64.00	57.00	59.83		
	*F*1	58.00	52.00	69.00	53.00	73.00	57.00	60.33		
实验五	Precision	82.35	80.00	85.15	93.10	86.42	85.33	85.39	86.84	0.00
	Recall	84.00	82.00	90.33	84.61	96.15	89.65	87.72		
	*F*1	81.16	79.35	88.71	90.19	92.03	88.60	86.67		

6　结　　语

本文选择传统语音信号与其包含的功能性副语言作为研究对象,首先,提取FPPC特征;然后,使用基于注意力机制的集成学习对FPPC特征进行训练;最后,通过自适应熵权重决策融合的方法,从决策层融合的角度构建融合FPPC特征的语音情感识别系统. 通过实验验证了FPPC特征在语音情感识别系统中的有效性. 提取功能性副语言中更能体现情感差异性的特征以及研究更优的信息融合方法将是功能性副语言今后研究的重点.

参考文献:

[1] Akçay M B, Oğuz K. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. *Speech Communication*, 2020, 116: 56-76.

[2] 孙颖,胡艳香,张雪英,等. 面向情感语音识别的情感维度PAD预测[J]. 浙江大学学报(工学版), 2019, 53(10): 2041-2048.
(Sun Ying, Hu Yan-xiang, Zhang Xue-ying, et al. Prediction of emotional dimensions PAD for emotional speech recognition[J]. *Journal of Zhejiang University (Engineering Science)*, 2019, 53(10): 2041-2048.)

[3] Moore J D, Tian L, Lai C. Word-level emotion recognition using high-level features[J]. *Lecture Notes in Computer Science*, 2014, 8404: 17-31.

[4] 赵小蕾,毛启容,詹永照. 融合功能性副语言的语音情感识别新方法[J]. 计算机科学与探索, 2014, 8(2): 186-199.
(Zhao Xiao-lei, Mao Qi-rong, Zhan Yong-zhao. New method of speech emotion recognition fusing functional paralanguages[J]. *Journal of Frontiers of Computer Science & Technology*, 2014, 8(2): 186-199.)

[5] Reuderink B, Poel M, Truong K, et al. Decision-level fusion for audio-visual laughter detection[C]//Popescu-Belis A, Stiefelhagen R. International Workshop on Machine Learning for Multimodal Interaction. Berlin: Springer, 2008: 137-148.

[6] Schuller B, Weninger F. Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization[C]//2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, 2010: 5054-5057.

[7] Foo L S, Yap W S, Hum Y C, et al. Real-time baby crying detection in the noisy everyday environment[C]//11th IEEE Control and System Graduate Research Colloquium (ICSGRC). Shah Alam, 2020: 26-31.

[8] Huang K Y, Wu C H, Hong Q B, et al. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, 2019: 5866-5870.

[9] Knox M T, Mirghafori N. Automatic laughter detection using neural networks[C]//8th Annual Conference of the International Speech Communication Association Belgium, 2007: 2973-2976.

[10] 赵小蕾,赵慧青. 说话人功能性副语音自动检测算法[J]. 智能计算机与应用, 2015, 5(1): 73-76.
(Zhao Xiao-lei, Zhao Hui-qing. Automatic detection algorithm of functional paralanguage in speech[J]. *Intelligent Computer and Applications*, 2015, 5(1): 73-76.)

[11] Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings[J]. *IEEE Open Journal of Engineering in Medicine and Biology*, 2020, 1: 275-281.

[12] Schuller B W, Schuller D M, Qian K, et al.COVID-19 and computer audition: an overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis [J].*Frontiers in Digital Health*, 2021, 3: 564906.
[13] Kaya H, Fedotov D, Dresvyanskiy D, et al. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics [C]//Proceedings of the 9th International Conference on Audio/Visual Emotion Challenge and Workshop.Nice, 2019: 27-35.
[14] Jassim W A, Paramesran R, Harte N. Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features [J]. *IET Signal Processing*, 2017, 11(5): 587-595.
[15] Hu J, Shen L, Sun G.Squeeze-and-excitation networks [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Salt Lake City, 2018: 7132-7141.
[16] Chen W C, Liu W B, Li K Y, et al. Rail crack recognition based on adaptive weighting multi-classifier fusion decision [J].*Measurement*, 2018, 123: 102-114.
[17] Xiao F Y. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy [J]. *Information Fusion*, 2019, 46: 23-32.
[18] Chou H C, Lin W C, Chang L C, et al.NNIME: the NTHU-NTUA Chinese interactive multimodal emotion corpus [C]// Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). San Antonio, 2017: 292-298.
[19] Hsu J H, Su M H, Wu C H, et al.Speech emotion recognition considering nonverbal vocalization in affective conversations [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, 29: 1675-1686.

(上接第 39 页)

[7] Piccardi M. Background subtraction techniques: a review [C]//2004 IEEE International Conference on Systems, Man and Cybernetics. Hague, 2004: 3099-3104.
[8] 王振亚,曾黄麟. 一种基于帧间差分和光流技术结合的运动车辆检测和跟踪新算法[J]. 计算机应用与软件,2012,29(5):117-120,129.
(Wang Zhen-ya, Zeng Huang-lin. A new algorithm of moving vehicle detection and tracking based on combining frame difference method with optical flow technique [J]. *Computer Applications and Software*, 2012, 29(5): 117-120,129.)
[9] Zhang J Y, Barron J L. Optical flow at occlusion [C]//Ninth Conference on Computer and Robot Vision. Toronto, 2012: 198-205.
[10] 陈梅,王健. 图像并行细化算法改进研究[J]. 中国印刷与包装研究,2014,6(1):30-34.
(Chen Mei, Wang Jian. Research on improved image parallel thinning algorithm [J]. *China Printing and Packaging Study*, 2014, 6(1): 30-34.)
[11] Sivolobov S. Human gait feature extraction method [J]. *Procedia Computer Science*, 2021, 193: 220-227.
[12] Collins R T, Gross R, Shi J B. Silhouette-based human identification from body shape and gait [C]//Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition. Washington D C, 2002: 366-371.
[13] An W Z, Liao R J, Yu S Q, et al. Improving gait recognition with 3D pose estimation [C]//13th Chinese Conference on Biometric Recognition. Urumqi, 2018: 137-147.
[14] 胡少晖,王修晖,刘砚秋. 基于多支路残差深度网络的跨视角步态识别方法[J]. 模式识别与人工智能,2021,34(5):455-462.
(Hu Shao-hui, Wang Xiu-hui, Liu Yan-qiu. Cross-view gait recognition method based on multi-branch residual deep network [J]. *Pattern Recognition and Artificial Intelligence*, 2021, 34(5): 455-462.)
[15] 陈欣,杨天奇. 不受服饰携带物影响的步态识别方法[J]. 计算机工程与应用,2016,52(5):141-146,153.
(Chen Xin, Yang Tian-qi. Gait recognition method without influence of dress and carrying [J]. *Computer Engineering and Applications*, 2016, 52(5): 141-146,153.)