DOI:10.11918/j.issn.0367-6234.201806131

PAD 三维情感空间中的语音情感识别

陈逸灵, 程艳芬, 陈先桥, 王红霞, 李 超

(武汉理工大学 计算机科学与技术学院, 武汉 430063)

摘 要: 离散情感描述模型将人类情感标注为离散的形容词标签,该类模型只能表示有限种类的、单一明确的情感类型,而维度情感模型从情感的多个维度量化了复杂情感的隐含状态. 另外,常用的语音情感特征梅尔频率倒谱系数(MFCC)存在因分帧处理引起相邻帧谱特征之间相关性被忽略问题,容易丢失很多有用信息. 为此本文提出改进方法,从语谱图中提取时间点火序列特征、点火位置信息特征对 MFCC 进行补充,将这三种特征分别用于语音情感识别,根据识别结果从 PAD 维度情感模型的三个维度(Pleasure-displeasure 愉悦度、Arousal-nonarousal 激活度、Dominance-submissiveness 优势度)进行相关性分析得到特征的权重系数,加权融合后获得情感语音的最终 PAD 值,将其映射至 PAD 三维情感空间中. 实验表明,增加的时间点火序列、点火位置信息不但能探测说话人的情感状态,同时考虑了相邻频谱间的互相关信息,与 MFCC 特征形成互补,在提升基本情感类型离散识别效果的基础上,将识别结果表示为 PAD 三维情感空间中的坐标点,采用量化的方法揭示情感空间中各种情感的定位与联系,展示出情感语音中糅杂的情感内容,为后续复杂的语音情感分类识别奠定研究基础.

关键词: PAD 三维情感模型;语音情感识别;梅尔频率倒谱系数;时间点火序列;点火位置信息;相关性分析

中图分类号: TN912.34 **文献标志码**: A **文章编号**: 0367-6234(2018)11-0160-07

Speech emotionestimation in PAD 3D emotion space

CHEN Yiling, CHENG Yanfen, CHEN Xianqiao, WANG Hongxia, LI Chao

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: The discrete emotional description model labels human emotions as discrete adjectives. The model can only represent limited types of single and explicit emotion. The dimensional emotional model quantifies the implied state of complex emotions from the multiple dimensions. In addition, conventional speech emotion feature, Mel Frequency Cepstral Coefficient (MFCC), has the problem of neglecting the correlation between the adjacent frame spectral features due to frame division processing, making it susceptible to loss of much useful information. To solve this problem, this paper proposes an improved method, which extracts the time firing series feature and the firing position information feature from the spectrogram to supplement the MFCC, and applies them in speech emotion estimation respectively. Based on the predicted values, the proposed method calculates the correlation coefficients of each feature from three dimensions, P (Pleasure-displeasure), A (Arousal-nonarousal), and D (Dominance-submissiveness), as feature weights and obtains the final values of PAD in emotion speech after the weighted fusion, and finally maps it to PAD 3D emotion space. The experiments showed that the two added features could not only detect the emotional state of the speaker, but also consider the correlation between the adjacent frame spectral features, complementing to MFCC features. On the basis of improving the effect of discrete estimation of basic emotional types, this method represents the estimation results as coordinate points in PAD 3D emotion space, adopts the quantitative method to reveal the position and connection of various emotions in the emotion space, and indicates the emotion content mixed in the emotion speech. This study lays a foundation for subsequent research on classification estimation of complex speech emotions.

Keywords: PAD 3D emotion model; speech emotion estimation; Mel Frequency Cepstral Coefficient; time firing series; firing position information; correlation analysis

语音是人类最基本、最便捷的交流形式,承载复杂信息的语音信号不仅可以反映语义内容,还能够传递说话人内在的情感状态,语音情感识别对建立更加自然、和谐的人机交互环境至关重要[1-2],是人工智能领域重要发展目标. 近年来,情感特征提取的研究工作已经取得不错进展[3-4],例如将时长、基频、能量等韵律特征用于情感识别[5-6],将谱相关特征运用到语音情感的识别[7-13],提取声音质量特征进行语音情感识别[14-16]等. 常用的谱特征例如梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients,

收稿日期: 2018-06-21
基金项目: 国家自然科学基金(51179146)
作者简介: 陈逸灵(1995—),女,硕士研究生
通信作者: 王红霞,99575522@ qq.com

MFCC)是基于人耳听觉特性提出来的[17-18]，已经广泛地应用在语音相关研究领域. Sun Y 等[19]指出了 MFCC 存在因分帧处理引起相邻帧谱特征之间相关性被忽略的问题，导致很多有用信息丢失. 语谱图直观地呈现语音信号在各时段的频率分布情况，包含了重要的语言学信息[20-22]，梁泽[23]、阮柏尧[24]将语谱图送入脉冲耦合神经网络(Pulse Coupled Neural Network，PCNN)进行迭代处理得到描述语谱图信息的时间点火序列、点火位置信息特征，该特征不仅包含语谱图的灰度分布信息，更重要的是还包含其中相邻像素之间的相对位置信息，不但能探测说话人的情感状态，同时考虑了相邻频谱间的互相关信息，与 MFCC 特征形成互补. 因此本文提出提取时间点火序列、点火位置信息特征对 MFCC 特征进行补充，从而改善语音情感识别系统性能，实验结果表明，改进之后的方法提升了语音情感识别效果. 另一方面，本文尝试将情感识别结果表示为 PAD 三维情感空间中的坐标点，突破了离散形容词标签(高兴、生气、悲伤等)描述情感种类局限性，采用量化方法揭示情感空间中各种情感的定位和关系，在处理情感的单一维度问题上也更高效.

1　语音情感特征提取

针对 MFCC 存在因分帧处理引起相邻帧谱特征之间相关性被忽略的问题，本节在首先提取 MFCC 特征后，加入时间点火序列和点火位置信息特征提取，为后续根据获取的上述三种特征预测所得 P，A，D 值加权融合提供数据.

1.1　梅尔频率倒谱系数

本文提取 MFCC 特征参数的过程见图 1.

图 1　MFCC 提取流程图

Fig.1　Flow chart of MFCC extraction

预处理包括预加重和加窗分帧[25]，帧长取 256，帧移取 128，窗函数为汉明窗，将每帧语音信号的语音序列 $\boldsymbol{s}(n)$ 进行快速傅里叶变换(FFT)得到各帧的频谱 $\boldsymbol{X}(n)$，对其取模平方得到语音信号的能量谱 $|\boldsymbol{X}(n)|^2$，将 $|\boldsymbol{X}(n)|^2$ 通过梅尔滤波器组 $\boldsymbol{H}_{\mathrm{m}}(k)$，输出参数 $P_{\mathrm{m}}(m=0,1,2,\ldots,M-1)$，其计算方法为

$$\boldsymbol{P}_{\mathrm{m}}=\sum_{k=f_{\mathrm{m}-1}}^{f_{\mathrm{m}+1}}\boldsymbol{H}_{\mathrm{m}}(k)\ |\boldsymbol{X}(n)|^2, m=0,1,2,\ldots,M-1. \tag{1}$$

式中 M 为滤波器的个数，本文取 26.

最后对参数 $\boldsymbol{P}_{\mathrm{m}}$ 取对数做离散余弦变换(DCT)得到静态梅尔频率倒谱系数 $\boldsymbol{C}_{\mathrm{mel}}(k)$：

$$\boldsymbol{L}_{\mathrm{m}}=\ln(\boldsymbol{P}_{\mathrm{m}}), m=0,1,2,\ldots,M-1, \tag{2}$$

$$\boldsymbol{C}_{\mathrm{mel}}(k)=\sum_{m=0}^{M-1}\boldsymbol{L}_{\mathrm{m}}\cos\left(\frac{k(m-0.5)}{M}\right), k=1,2,3,\ldots,N. \tag{3}$$

对静态 MFCC 参数进行一阶差分，对得到的一阶差分系数进行二阶差分，得到动态的一阶 MFCC 参数和二阶 MFCC 参数. 通过实验对比发现，一阶 MFCC 参数对语音情感识别的准确率较静态 MFCC 参数有一定程度的提高，但二阶 MFCC 参数由于多了一倍的参数值，导致维数增加，在当前语音特征值维数与语料相比过大的情况下，维数的提升反而导致该参数识别性能下降，最终本文选取一阶差分 MFCC 参数.

1.2　神经元点火序列

将语音信号对应的语谱图作为 PCNN 的输入，得到神经元点火序列，过程见图 2，该序列描述语谱图各时刻释放脉冲(点火)的神经元总数，其中求语谱图流程见图 3. 帧长、帧移以及窗函数的选择与计算 MFCC 相同，用时间 n 作横坐标，频率 ω 作纵坐标，任一给定频率成分在给定时刻的强弱程度采用相应点的灰度表示，构成的二维图像即为语谱图[26].

图 2　特征时间序列提取流程图

Fig.2　Flow chart of characteristic time series extraction

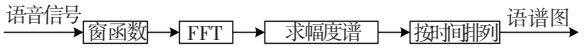

图 3　求语谱图过程流程图

Fig.3　Flow chart of speech spectrum seeking process

Eckhorn 依据猫的视觉皮层神经元释放同步脉冲现象提出了脉冲耦合神经网络(PCNN)模型[27]，PCNN 单个神经元由接收域、耦合连接域和脉冲发生器三部分组成，其简化结构模型见图 4.

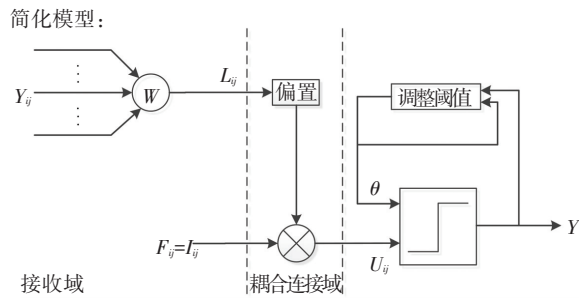

图 4　脉冲耦合神经网络神经元简化模型

Fig.4　Simplified model of pulse coupled neural network neuron

1)在接收域中，$\boldsymbol{Y}_{ij}$ 代表附近神经元上一时刻的输出，$\boldsymbol{Y}_{ij}$ 与对应系数相乘得到连接输入项 $\boldsymbol{L}_{ij}$. 反馈输入项 $\boldsymbol{F}_{ij}$ 即 $\boldsymbol{I}_{ij}$，代表来自外界刺激信号输入(本文

指语谱图像素构成矩阵中第 (i, j) 个像素灰度值),计算公式为

$$\boldsymbol{L}_{ij}[n] = \sum w_{ijpq}\boldsymbol{Y}[n-1], \tag{4}$$

$$\boldsymbol{F}_{ij}[n] = \boldsymbol{I}_{ij}. \tag{5}$$

式中 w_{ijpq} 为连接权值矩阵 $\boldsymbol{W}$ 的元素,代表 PCNN 中第 (i, j) 个神经元与第 (p, q) 个神经元的连接系数. 权值矩阵一般是 3 × 3 的矩阵,反映了相邻神经元传输给中心神经元信号强度的大小. 其元素定义为

$$w_{ijpq} = \frac{1}{(i-p)^2 - (j-q)^2}. \tag{6}$$

式中 (i,j) ,(p,q) 为像素点的坐标值.

2)在耦合连接域中,将连接输入项 $\boldsymbol{L}_{ij}$ 的偏置与反馈输入项 $\boldsymbol{F}_{ij}$ 相乘,得到内部活动项 $\boldsymbol{U}_{ij}$,表达式为

$$\boldsymbol{U}_{ij}[n] = \boldsymbol{F}_{ij}(1 + \beta\boldsymbol{L}_{ij}[n]). \tag{7}$$

式中 β 为连接强度系数.

3)在脉冲发生器中,比较内部活动项 $\boldsymbol{U}_{ij}$ 与动态阈值门限 $\boldsymbol{\theta}_{ij}$ 的大小,若内部活动项 $\boldsymbol{U}_{ij}$ 较大,则神经元释放脉冲(也称作点火)并且使输出的脉冲值重新对动态阈值门限进行反馈调节. 若 $\boldsymbol{U}_{ij}$ 不大于动态阈值,那么神经元不输出脉冲(也称作不点火). 输出脉冲和动态阈值门限的表达式为

$$\boldsymbol{Y}_{ij}[n] = \begin{cases} 1, & \boldsymbol{U}_{ij}[n] > \boldsymbol{\theta}_{ij}[n-1]; \\ 0, & \boldsymbol{U}_{ij}[n] \leqslant \boldsymbol{\theta}_{ij}[n-1]. \end{cases} \tag{8}$$

$$\boldsymbol{\theta}_{ij}[n] = \exp(-\alpha_\theta)\,\boldsymbol{\theta}_{ij}[n-1] + V_\theta\,\boldsymbol{Y}_{ij}[n-1]. \tag{9}$$

式中:α_θ 为 $\boldsymbol{\theta}_{ij}$ 的衰减时间常数,V_θ 为动态阈值门限固有常数,$\boldsymbol{Y}_{ij}[n]$ 为 PCNN 神经元的输出值,本文定义输出从 0 变为 1 为神经元的点火.

本文使用 50 次迭代的点火神经元的总数 $\boldsymbol{T}(n)$ 作为特征时间序列,定义为

$$\boldsymbol{T}(n) = \sum_{ij} \boldsymbol{Y}_{ij}(n). \tag{10}$$

式中 $\boldsymbol{Y}_{ij}(n)$ 为 n 时刻二值图像输出,$\boldsymbol{T}(n)$ 统计了 n 时刻整幅图像中发出脉冲的神经元总数.

1.3 神经元点火位置信息

将每次迭代所得点火神经元位置颁布图分别向时间轴和频率轴上投影,再把投影后的两个矢量合并成一个矢量. 最后,将每个时刻的点火位置分布图所得矢量按照时间排成多列,组成一个矩阵即为语音情感识别特征矩阵,过程见图 5.

图中,$\boldsymbol{R}_f$ 代表某一时刻对语谱图的点火位置图像的频率轴投影所得矢量,$\boldsymbol{R}_t$ 代表该时刻对时间轴投影所得矢量,$\boldsymbol{R}_{(i)}$ 表示两个矢量的组合矢量,$\boldsymbol{A}$ 是每个时刻得到的组合矢量按时间顺序排列合成的特征矩阵.

本文提取神经元点火位置信息的流程见图 6.

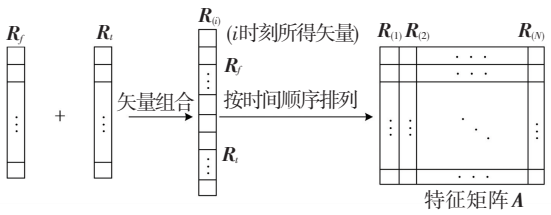

图 5 提取特征时间序列流程图

Fig.5 Flow chart of characteristic time series extraction

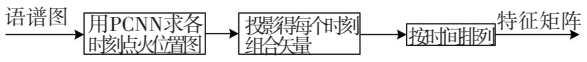

图 6 提取神经元点火位置信息流程图

Fig.6 Flow chart of neuron ignition position information extraction

2 PAD 三维情感空间中的语音情感识别

2.1 PAD 三维情感模型

学者们一般采用离散情感标签来分类少数几种基本情感,但悲喜交加、喜极而泣等已不再完全属于某一基本情感类别,这不利于创建新型的、更加人性化的人机交互环境. 基于维度论的 PAD 三维情感模型见图 7,该模型打破了传统的标签描述方法,是各种情感维度模型中较为成熟的一种.

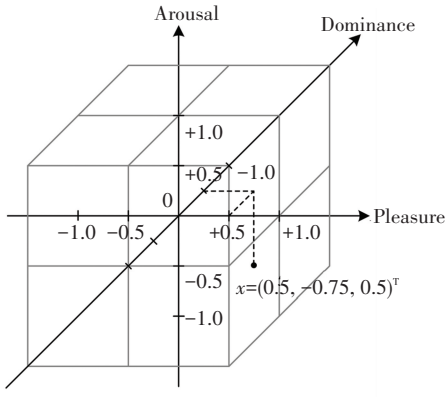

图 7 PAD 三维情感模型

Fig.7 PAD 3D emotion model

其中 P 代表愉悦度(Pleasure-Displeasure),表明了个体情感状态的积极或消极特性;A 代表激活度(Arousal-Nonarousal),表明了个体的神经生理激活程度;D 代表优势度(Dominance-Submissiveness),表明了个体对环境和他人的主观控制状态[28]. 该模型与离散情感描述模型相比,更关注情感内在成分的表达,通过计算待预测情感语音的 PAD 值与各基本情感类型的 PAD 量表值的距离,可以判别待测情感语音的情感状态并分析其内在情感组成,本文选择该模型进行语音情感分析.

2.2 特征权重计算

利用提取所得 MFCC、PCNN 神经元点火序列以及神经元点火位置信息三种语音情感特征,用支持

向量机回归 SVR(Support Vector Regression)建立语音情感识别模型,预测不同语音情感的 P、A、D 值,得到不同语音情感在三维情感空间中的分布. 最后,按照不同特征的相关性系数来确定特征权重,融合得到该语音信号在三维情感空间中的最终 PAD 值. 实现流程见图 8.

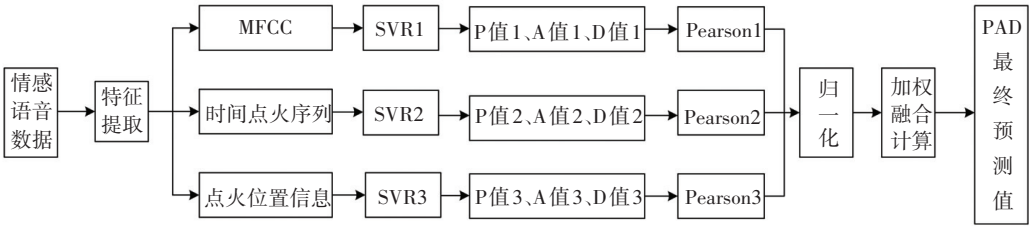

图 8　PAD 三维情感模型中的多特征融合

Fig.8　Multi-feature fusion in PAD 3D emotion model

计算 Pearson 相关系数确定特征权重,它在协方差的基础上除以两个随机变量的标准差,得到一个介于-1~1 之间的值,当相关系数等于 1 时,表示两个随机变量的线性关系正相关最强;反之当它等于-1 时,表示它们之间线性关系负相关最强;等于 0 说明它们之间不存在线性相关关系为

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \mu \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}. \tag{11}$$

式中: X 为预测值, Y 为 PAD 量表参考值, μ_X , μ_Y 分别为变量 X , Y 的平均值, σ_X , σ_Y 分别表示变量 X , Y 的标准差.

3　实验结果与分析

3.1　语音情感数据库

选择公开使用许可的中科院自动化所语音情感数据库(CASIA),柏林语音情感数据库(EmoDB)和 Surrey 视听表情情感数据库(SAVEE)评价所提方法的有效性,它们在语音情感识别中应用广泛[29],使用上述数据库可以方便与其他工作的实验结果对比.

CASIA 由 4 位录音人在 5 种不同情感(温和、惊吓、生气、难过、高兴、悲伤)下对 500 句文本发音得到,共9 600句语料. EmoDB 由 10 位演员对 10 条语句进行 7 种情感(温和、生气、害怕、高兴、悲伤、厌恶、难过)的演绎得到,共包含 800 句语料. SAVEE 由 4 个演员制作,设计了 7 种情感(生气、害怕、厌恶、高兴、温和、惊吓、悲伤),共 480 句语料.

上述 3 个数据库虽然是离散语音情感数据库,但是根据 Mehrabian 研制的原版 PAD 情绪量表以及中国科学院心理所修订的中文简化版 PAD 情绪量表与基本情感类型的对应关系,可以获得数据库中各情感类型的 PAD 量表值,所以以上 3 个数据库中语料能够作为本文实验所需的维度情感语音数据,涉及的基本情感类型对应的 PAD 量表值见表 1.

表 1　基本情感类型的 PAD 量表值

Tab.1　PAD scale for basic emotion types

情感类型	平均数		
	P	A	D
温和	0.27	−0.07	−0.14
高兴	0.66	0.74	0.32
悲伤	−0.28	−0.36	−0.78
生气	−0.86	0.66	0.91
惊吓	−0.33	0.65	−0.72
难过	−0.46	0.58	−0.13
厌恶	−0.44	0.22	0.36
害怕	−0.33	0.65	−0.72

3.2　基于三种特征的语音情感识别

实验的运行环境为 Matlab(R2014b),系统环境为 Win10,计算机配置为 Intel Core i3-3217U CPU (1.8GHz),8GB 内存.

MFCC 的提取方法依照提取流程图获得 MFCC. 为确保 PCNN 输出时间序列唯一,对语谱图的处理采用同一参数模型,取值见表 2.

表 2　PCNN 的参数及取值

Tab.2　Parameters and values of PCNN

参数	α_F	α_L	α_θ	V_F	V_L	V_θ	β
取值	0.1	1.0	1.0	0.5	0.2	20	0.1

表 2 中, α_F、α_L、α_θ 分别为反馈输入项 $\boldsymbol{F}_{ij}$, 连接输入项 $\boldsymbol{L}_{ij}$, 动态阈值门限 $\boldsymbol{\theta}_{ij}$ 的衰减时间常数. V_F, V_L, V_θ 分别表示 PCNN 的反馈放大系数、连接放大系数和阈值放大系数. 通过文献[30]提出的结合局部梯度确定时间衰减常数 α_θ 和放大系数 V_θ 方法指导参数 α_θ、V_θ 的设置,计算方法为

$$\alpha_\theta = \nabla f(x,y), \tag{12}$$

$$V_\theta = 1 - \nabla f(x,y). \tag{13}$$

式中 $\nabla f(x,y)$ 为语谱图 $f(x,y)$ 的梯度表示,计算方法为

$$\nabla f(x,y) = \left[\frac{\partial f}{\partial x}\ \frac{\partial f}{\partial y}\right]^{\mathrm{T}}. \tag{14}$$

表 2β 为连接强度系数,其值越大神经元兴奋的提升量越大,神经元也越容易点火. 相应地,语谱图像素的局部对比度可以反映神经元的兴奋状态,对比度高说明对应位置神经元比较兴奋,容易受到刺激而点火,因此将局部对比度的大小作为参数 β 设置的依据. $\boldsymbol{Y}$,$\boldsymbol{L}$,$\boldsymbol{U}$ 初始值设为 0,输入为归一化的灰度值,属于[0,1]之间. 连接域半径 r = 1.5,内部连接矩阵 $\boldsymbol{W}$ 是一个 3 × 3 的方阵,其中每一个元素数值为中心像素到周围每个像素的欧几里德距离平方的倒数(r^{-2}).

利用提取的三种语音情感特征,建立情感语音识别预测模型,本文 SVR 回归算法选用高斯径向基函数,实验样本的 80% 作为训练集,剩余样本作为测试集检验预测效果. 表 3 是以 CASIA 数据库中文本内容为“阳光使得你们温暖”,情感标签为高兴的测试数据为例,得到的三种特征的相关系数归一化结果. 首先将该语音的 MFCC 特征作为 SVR 的输入参数,SVR 的输出结果为该情感语音的 PAD 值,然后计算该预测值与 PAD 量表值的 Pearson 相关系数. 同理,继续利用该语音情感数据的点火时间序列和点火位置信息特征进行 P、A、D 值预测,计算 Pearson 相关系数. 根据表 3 中相关性分析结果我们可知道,不同的特征在 P(愉悦度)、A(激活度)和 D(优势度)三个维度上所对应的相关系数存在差别,其中 P、A 的相关系数由大到小的顺序是点火位置信息特征、时间点火序列、MFCC,而 D 的相关系数由大到小的顺序为时间点火序列、点火位置信息特征、MFCC,表明三种特征在语音情感识别的准确率中各有侧重. 因此,在对语音情感进行识别时,不宜选取单一特征,而应根据相关系数的大小给三种特征赋予不同的权值以提高识别精度.

表 3 三种语音情感特征的相关系数归一化结果

Tab.3 Correlation coefficient normalization of three speech emotion features

语音情感识别特征		相关系数归一化值					
		P 维度相关系数	归一化值	A 维度相关系数	归一化值	D 维度相关系数	归一化值
基于谱的相关特征	MFCC	0.404	0.262	0.588	0.304	0.204	0.234
基于语谱图的相关特征	时间点火序列	0.562	0.364	0.634	0.328	0.365	0.419
	点火位置信息特征	0.576	0.374	0.711	0.368	0.302	0.347

3.3 三种特征的加权融合

根据三种语音情感特征的相关系数归一化结果,获取待识别情感语音的 PAD 值加权融合计算结果. 计算方法为

$$P = P_1\lambda_1 + P_2\lambda_2 + P_3\lambda_3. \tag{15}$$

式中: P_1, P_2, P_3 依次代表采用 MFCC,点火时间序列,点火位置信息对该语音在 P 维度的预测值, λ_1, λ_2, λ_3 表示以上三种语音情感特征对该语音情感类型在 P 维度的相关系数归一化值,满足 $\lambda_1 + \lambda_2 + \lambda_3 = 1$. 由式(15)计算所得 P 即为该语音在三维情感空间中的愉悦度最终预测值. 同理可计算其优势度最终值 D 和激活度最终值 A .

运用本文提出的方法对 CASIA 语音数据库中数据进行 PAD 值预测,将得到的 PAD 数据表示在三维空间中,绘制出 PAD 空间中不同情感的分布见图 9.

由分布图观察得,语音数据库中各标签类下语音 P、A、D 值分布情况见表 4.

图 9 显示 6 种情感样本类间分布比较分散,各个样本类内则较集中地分布在表 1 中 PAD 量表值对应坐标点附近,证明上述三种情感特征的预测结果通过相关系数加权融合后,在连续维度情感预测中对基本情感类型具有较好的区分效果. 分布图中绝大多数情感语音样本的位置并不与 PAD 量表表 1 中所示 6 种基本情感的 PAD 参考值重合,而是离散分布在基本情感标定点附近,因此可以通过计算其与基本情感 PAD 值的距离,进一步分析该情感语音状态的构成元素以及组成比例,从而能够识别出形如“温和偏悲伤”或者“温和偏高兴”等混合情感类型,更恰当地反映出情感表达的极性和程度.

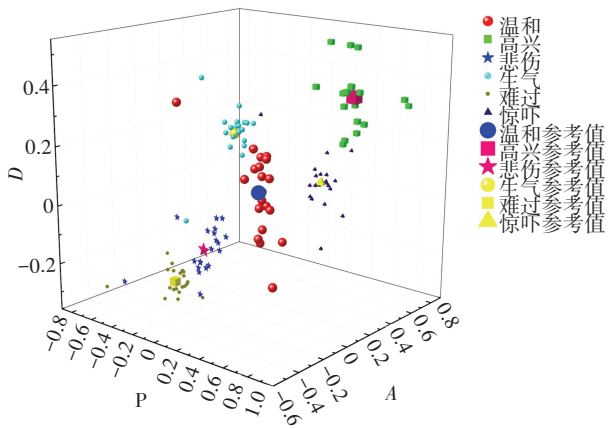

图 9 PAD 三维情感空间分布

Fig.9 PAD 3D emotional spatial distribution

表 4　6 种情感的 PAD 最终预测值分布范围

Tab.4　Distribution range of PAD final predictive value of the 6 emotion types

情感类型	各维度分布范围		
	P	A	D
温和	−0.1~0.1	−0.06~0.02	−0.2~0.01
高兴	0.4~0.53	0.67~0.73	0.17~0.3
悲伤	−0.4~−0.27	−0.41~−0.31	−0.4~−0.28
生气	−0.88~−0.8	0.62~0.7	0.14~0.2
惊吓	−0.32~−0.2	0.57~0.63	−0.12~−0.01
难过	−0.35~−0.2	−0.2~−0.15	−0.37~−0.24

3.4　与其他特征对比

采用均方根误差（root − mean − square error，RMSE）作为基本情感类型识别准确性的评价指标，计算方法为

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(X_{\mathrm{obs},i}-X_{\mathrm{model},i})^2}{n}}. \quad (16)$$

式中：$X_{\mathrm{obs},i}$ 为实验预测值，$X_{\mathrm{model},i}$ 为 PAD 量表参考值.

为证明时间点火序列、点火位置信息对 MFCC 有更好的互补效果，计算文献[31]中 MFCC、文献[32]中 MFCC+OTHER、文献[33]中 MFCC+PROS、本文 MFCC+FIRING 不同特征对 MFCC 补充之后预测结果的 RMSE 值见表 5. 其中 OTHER 由线性预测倒谱系数（Linear Predictor Cepstral Coefficient，LPCC）、过零幅度峰值（Zeros Crossing Peak Amplitude，ZCPA）组成，PROS 表示基频、短时能量、共振峰等韵律特征，FIRING 代表时间点火序列和点火位置信息. 表中每个维度的 RMSE 值都规范化至 0~1 之间，RMSE 越小说明对应的方法性能越好.

表 5　RMSE 值对比

Tab.5　RMSE value comparison

特征对比	RMSE 值								
	CASIA 数据库			EmoDB 数据库			SAVEE 数据库		
语音情感特征	P 维度	A 维度	D 维度	P 维度	A 维度	D 维度	P 维度	A 维度	D 维度
MFCC	0.77	0.59	0.81	0.81	0.57	0.79	0.80	0.59	.84
MFCC+OTHER	0.73	0.53	0.80	0.78	0.51	0.82	0.76	0.60	0.83
MFCC+PROS	0.34	0.31	0.37	0.41	0.33	0.32	0.31	0.35	0.36
MFCC+FIRING	0.22	0.17	0.21	0.26	0.18	0.19	0.16	0.20	0.23

从表 5 的对比结果看出，只用 MFCC 特征，识别结果不尽人意，虽然 MFCC 特征充分考虑了人耳的听觉特征，但在语音情感识别中表现却不及语音识别和说话人识别中优秀，其他谱特征对 MFCC 也几乎没有补充效果. 加入韵律特征后，识别效果提升很多，这说明韵律特征相对其他谱特征对 MFCC 有较大改进效果. 在 3 个数据库上 MFCC+FIRING 的 RMSE 值最小，其中在 SAVEE 数据库中 P 维度的 RMSE 值从 0.80 降低到 0.31，这一结果说明本方法引入时间点火序列和点火位置信息能提高情感识别效果，并且两者在对 MFCC 的互补效果上比其他谱特征和韵律特征性能更好.

4　结　论

本文在常用语音情感特征 MFCC 的基础上增加时间点火序列、点火位置信息特征分别单独进行语音情感预测，计算该语音预测结果与 PAD 量表值的相关系数，将相关系数归一化结果作为特征权重，加权融合获得语音的 PAD 最终预测值. 该方法在提升基本情感类型识别效果的基础上，将识别结果表示为 PAD 三维情感空间中的坐标点，采用量化方法揭示情感空间中各种情感定位与联系，更细腻地展现语音情感状态构成，为后续复杂语音情感的分类识别奠定研究基础.

参考文献

[1] 胡海龙.基于音节分布特征提取的语音情感识别[D]. 厦门：厦门大学，2011
HU Hailong.Speech emotion recognition based on syllable distribution feature extraction [D]. Xiamen：Xiamen University，2011

[2] 蒋海华，胡斌.基于 PCA 和 SVM 的普通话语音情感识别[J].计算机科学，2015，42(11)：270.DOI：10.11896/j.issn.1002−137X.2015.11.055
JIANG Haihua，HU Bing. Analysis ofmandarin speech emotion based on PCA and SVM [J]. Computer Science，2015，42(11)：270. DOI：10.11896 / j.issn.1002−137X.2015.11.055.

[3] SWAIN M，ROUTRAY A，KABISATPATHY P. Databases，features and classifiers for speech emotion recognition：a review [J]. International Journal of Speech Technology，2018，21(1)：93

[4] MUSTAFA M B，YUSOOF M A M，DON Z M，et al. Speech emotion recognition research：an analysis of research focus [J]. International Journal of Speech Technology，2018，21(1)：137

[5] ILIOU T，ANAGNOSTOPOULOS C N. Statistical Evaluation of Speech Features for Emotion Recognition[C]// International Conference on Digital Telecommunications. Colmar，France IEEE，2009：121

[6] WANG Y，DU S，ZHAN Y. Adaptive and Optimal Classification of Speech Emotion Recognition[C]// Fourth International Conference on Natural Computation. IEEE Computer Society，2008：407

[7] OOI C S，Seng K P，ANG L M，et al. A new approach of audio emotion recognition [J]. Expert Systems with Applications，2014，41

(13):5858

[8] LIKITHA M S, GUPTA S R R, HASITHA K, et al. Speech based human emotion recognition using MFCC[C]// International Conference on Wireless Communications, Signal Processing and NETWORKING. IEEE, 2018

[9] ROZGIC V, ANANTHAKRISHNAN S, SALEEM S, et al. Emotion Recognition using Acoustic and Lexical Features[C]// Conference of the International Speech Communication Association. 2012

[10] WANG K, AN N, LI B N, et al. Speech emotion recognition using fourier parameters [J]. IEEE Transactions on Affective Computing, 2017, 6(1):69

[11] BITOUK D, VERMA R, NENKOVA A. Class-Level spectral features for emotion recognition[J].Speech Communication, 2010, 52 (7/8):613. DOI: 10.1016/j.specom.2010.02.010

[12] CHAUHAN R, YADAV J, KOOLAGUDI SG, et al. Text independent emotion recognition using spectral features[C]//Proc. of the 2011 Int'l Conf. on Contemporary Computing. Berlin, Heidelberg: Springer-Verlag, 2011.359. DOI: 10.1007/978-3-642-22606-9_37

[13] WU SQ, FALK TH, CHAN WY. Automatic speech emotion recognition using modulation spectral features. Speech Communication, 2011, 53(5):768.DOI:10.1016/j.specom.2010.08.013

[14] LUGGER M, YANG B. The relevance of voice quality features in speaker independent emotion recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2007:IV-17. DOI:10.1109/ICASSP.2007.367152

[15] LUGGER M, YANG B. Psychological motivated multistage emotion classification exploiting voice quality features[C]// Proc. of the Speech Recognition. 2008

[16] LUGGER M, JANOIR ME, YANG B. Combining classifiers with diverse feature sets for robust speaker independent emotion recognition[C]//Proc. of the 2009 European Signal Processing Conf. Glagow: EURASIP, 2009. 1225

[17] 刘振焘, 徐建平, 吴敏等. 语音情感特征提取及其降维方法综述[J/OL].计算机学报,2017,1-22(2017-08-13).http://kns.cnki.net/kcms/detail/11.1826.TP.20170813.1200.006.html.
LIU Zhentao, XU Jianping, WU Min, et al. Aeeview of speech emotion extraction and its dimension reduction [J]. Chinese Journal of Computers, 2017, 1-2 (2017-08-13).http://kns.cnki.net/kcms/detail/11.1826.TP.20170813.1200.006.html

[18] 袁敏.汉语孤立词识别理论及关键技术研究[D].兰州大学, 2005
YUAN Min. Chinese isolated word recognition theory and key technologies [D]. Lanzhou University, 2005

[19] SUN Y, WEN G, WANG J. Weighted spectral features based on local Hu moments for speech emotion recognition[J]. Biomedical Signal Processing and Control, 2015, 18:80

[20] AJMERA P K, JADHAV D V, HOLAMBE R S. Text-independent speaker identification using radon and discrete cosine transforms based features from speech spectrogram [J]. Pattern Recognition, 2011, 44(10/11):2749

[21] KALINLI O, NARAYANNAN S. Prominence detection using auditory attention cues and task-dependent high level information. [J]. IEEE Transactions on Audio Speech & Language Processing, 2009, 17(5):1009

[22] 马义德, 袁敏, 齐春亮,等. 基于 PCNN 的语谱图特征提取在说话人识别中的应用[J]. 计算机工程与应用, 2005, 41(20):81
MA Yide, YUAN Mi, QI Chunliang, et al. Application of PCNN-based feature extraction in speaker recognition [J]. Computer Engineering and Applications, 2005, 41 (20): 81

[23] 梁泽. PCNN 在语音情感识别中的应用研究[D]. 兰州: 兰州大学, 2008
LIANG Ze. Application of PCNN in speech emotion recognition [D]. Lanzhou: Lanzhou University, 2008

[24] 阮柏尧. 脉冲耦合神经网络(PCNN)在基于语谱图的说话人识别中的应用[D]. 五邑大学, 2008
RUAN Baiyao. Application of pulse coupled neural network (PCNN) in speaker recognition based on speech spectrogram [D]. Wuyi university, 2008

[25] 张文克. 融合 LPCC 和 MFCC 特征参数的语音识别技术的研究[D].湘潭大学,2016
ZHANG Wenke. Research onspeech recognition technology combining LPCC and MFCC feature parameters [D] .Xiangtan University, 2016

[26] 李姗,徐珑婷.基于语谱图提取瓶颈特征的情感别算法研究[J].计算机技术与发展,2017,27(5):82.DOI:10.3969/j.issn.1673-629X.2017.05.018
LI Shan, XU LT. Research onemotion recognition algorithm of bottleneck characteristics extraction based on spectral graph [J]. Computer Technology and Development, 2017, 27 (5): 82. DOI: 10.3969 / j.issn.1673-629X .2017.05.018

[27] 宋静,张雪英,孙颖等.基于 PAD 情绪模型的情感语音识别[J].微电子学与计算机,2016,33(9):128
SONG Jing, ZHANG Xueying, SUN Ying et al. Emotion speech recognition based on PAD emotion model [J]. Microelectronics and Computers, 2016, 33 (9): 128

[28] 张雪英,孙颖,张卫,等. 语音情感识别的关键技术[J]. 太原理工大学学报,2015,46(06):629.
ZHANG Xueying, SUN Ying, ZHANG Wei, et al. Key Techniques of Speech Emotion Recognition [J]. Taiyuan University of Technology, 2015, 46 (06): 629

[29] 韩文静,李海峰. 情感语音数据库综述[J]. 智能计算机与应用, 2013,3(01):5
HAN Wenjing, LI Haifeng. Summary of Emotional Speech Database [J]. Journal of Microcomputer & Application 2013, 3 (01): 5

[30] 李建锋. 脉冲耦合神经网络在图像处理中的应用研究[D]. 中南大学, 2013
LI Jianfeng. Application of pulse coupled neural network in image processing [D]. Central South University, 2013

[31] 韩一, 王国胤, 杨勇. 基于 MFCC 的语音情感识别[J]. 重庆邮电大学学报(自然科学版), 2008, 20(5):597
HAN Yi, WANG Guoyin, YANG Yong. Speech emotion recognition based on MFCC [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2008, 20(5):597

[32] 孙亚新. 语音情感识别中的特征提取与识别算法研究[D]. 广州:华南理工大学, 2015
SUN Yaxin. Research on feature extraction and recognition algorithm in voice emotion recognition [D]. Guangzhou: South China university of technology, 2015

[33] 任浩, 叶亮, 李月,等. 基于多级 SVM 分类的语音情感识别算法[J]. 计算机应用研究, 2017,34(6):1682
REN Hao, YE Liang, LI Yue, et al. Speech emotion recognition algorithm based on multi-level SVM classification [J]. ComputerApplication Research, 2017, 34(6):1682

(编辑 苗秀芝)