# 融合说话者特征的个性化自然语音情感识别

贾 宁[1] 郑纯军[1,2] 孙 伟[1]

[1](大连东软信息学院 辽宁 大连 116023)

[2](大连海事大学 辽宁 大连 116023)

**摘 要** 情感特征的高级表示与说话者的个性化特征之间存在较强相关性,因此以提升个性化情感识别精度为目标,设计一组融合说话者特征和语音情感特征的识别模型,利用卷积神经网络模型获取说话者类别,在融合说话人特征高阶表达的基础上,利用卷积循环神经网络训练个性化情感识别模型,结合自建的成人自然情感语料库,在多项语音情感语料库上测试识别模型性能,从而验证该模型的有效性。

**关键词** 语音情感识别 说话者特征 卷积循环神经网络 语谱图 个性化模型 成人自然情感语料库

**中图分类号** TP3 TP183 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.12.030

# PERSONALIZED NATURAL SPEECH EMOTION RECOGNITION BASED ON SPEAKER CHARACTERISTICS

Jia Ning[1] Zheng Chunjun[1,2] Sun Wei[1]

[1](*Dalian Neusoft University of Information*, *Dalian 116023*, *Liaoning*, *China*)

[2](*Dalian Maritime University*, *Dalian 116023*, *Liaoning*, *China*)

**Abstract** There is a strong correlation between the high-level representation of emotional features and the speaker's personalized features. Therefore, in order to improve the accuracy of personalized emotion recognition, we design a group of recognition models which integrate speaker features and speech emotion features. The convolution neural network model was used to obtain the speaker category. Based on the high-order representation of the speaker's features, we adopted the convolutional recurrent neural network to train the personalized emotion recognition model. Combined with the self-built adult natural emotion corpus, we tested the performance of the recognition model on multiple speech emotion corpus to verify the effectiveness of the model.

**Keywords** Speech emotion recognition Speaker features Convolutional recurrent neural network Spectrogram Personalized model Adult natural emotion corpus

## 0 引 言

语音作为人类交流最方便、最自然的媒介,是相互传递信息时采取的最基本、最直接的途径。语音包含多种不同类型的信息,可以表达丰富的情感信息[1]。

语音情感识别旨在通过语音信号识别说话者的正确情绪状态,目前对于情感的研究仍然处于学科交叉的领域,至今也未有统一的定义与规范。由于语音并非情感生理信号的完整表达形式,在忽略其余感官结果的前提下,如何高效而精确地识别用户表达的情感,是近年来语音学研究的热点领域[2]。总体上,目前的语音情感的整体识别率较低,泛化能力不强,主要来源于以下情感特征提取方法和模型设计等方面的制约。

从富有情感的语音数据中学习有用的声学特征。主要方法有三种,分别为:

(1) 采用手工制作的特征[3]。从原始音频文件中

收稿日期:2020-06-01。辽宁省自然科学基金项目(20180551068)。**贾宁**,副教授,主研领域:语音情感识别,深度学习,大数据分析。**郑纯军**,副教授。**孙伟**,教授。

提取手工特征,捕获最原始的不同类型的声学特征,从而判定该特征所属的语音学任务类型。

(2) 将传统特征与深度学习模型融合[4],在交叉领域中突出特征的重点,由于不同任务的侧重点不同,其融合的方式体现多样化、个性化的特点。

(3) 通过对原始音频信号进行分析,获取其中的情感影响因子与规律。

由于第 3 种方式导致情感特征维数过多,过度增加了语音情感识别过程的计算量,也就无形中增加了语音情感识别系统的空间复杂度和时间复杂度[5],因此常用前两种方法进行特征提取。

深度学习方法可以从不同层次的输入中学习有效的语音信号的非线性表现形式,目前已经广泛应用于语音情感模型设计中,目前常见的深度学习模型可以分为有监督和无监督两种,针对语音情感识别任务,主要采用深度神经网络(Deep Neural Network,DNN)[6]、卷积神经网络(Convolutional Neural Network,CNN)[7]、循环神经网络(Recurrent Neural Network,RNN)[8]、卷积循环神经网络(Convolutional Recurrent Neural Network,CRNN)[9]等有监督模型,为了突出不同任务的信号特征,还会融合多通道识别技术和注意力机制来进行情感识别。

然而,大多数研究集中于通用语料库上的具有泛化性能的模型和识别方案设计,现有的开源语料库往往存在数据量不足、倾斜现象、包含背景噪声、多为外文语料、标注结果精度不够等缺陷,而且鲜有专家对不同语料库的特征之间、不同说话者之间、个性化特征与模型的相关性之间进行充分的挖掘,直接导致现有的模型进行语音情感识别任务的准确率不高。

围绕上述问题,本文针对语音情感特征提取、个性化的深度学习模型设计和学习方案等方面开展了相关的研究,提出一种基于自建成人情感语料库、具备说话者个性化特征的、准确率较高的语音情感识别模型。

## 1 成人自然情感语料库设计

目前,常见的情感采集方案主要针对自然语音、诱导语音和表演语音进行设计[10]。自然语音是在自然条件下的真实情感表达,它包含最佳的情感数据,但采集困难,而且涉及复杂的后期数据处理和背景噪声分离操作。诱导语音则是在固定的场景模型下激发个人的情感,一般在专业环境下采集,因此背景噪声较少,因其诱发的情感将说话人带入特定的场景,其具备一定的真实性,但是无法衡量说话人表达情绪的刻意程度。表演语音是基于指定台词的目标情绪表演,它的刻意性较强,而且情绪表达过于饱满,与自然语音的表达存在一定的差异。然而针对此类语音,它的采集方式是最便捷的。

为了确保情感语料库数据的覆盖面和规模,本文主要采集自然语音和诱导语音,并将其有效地融合在一起,其目标是设计一个规模大、年龄层覆盖面广、情感类别平衡、语音质量高、情感表达基本正确的情感语音数据库。目前,此数据库中收录的情感包括高兴、愤怒、平静和悲伤四种情绪。

为有效地实现诱导语音设计,准备了 30 条相关的中文语料信息,这些语料信息多为对话的形式,它的内容多数存在情感分歧,即情感的表达与语义无关,而且具备浓重的语音信息,要求受试者在融入特定环境后,以多种方式恰当地表达特定的几种感情。现有受试者为 16 人,年龄分布在 19 至 40 岁之间,男女比例平衡。

自然语音的采集使用特定的语音采集装置。采集装置存放于小范围内的室内场所,例如家庭、寝室、社区、小型诊所等,可使用语音唤醒的方式,与特定人群进行语音沟通,记录说话者的音频数据。由于采集装置的提示信息为日常的生活用语,说话者在回答时一般较为自然,可以判定为自然语音。此设备存在的问题是,录制的语音可能存在背景噪声,需要后期统一处理。

为保证数据集中处理的正确性,本数据库的录音文件以 WAV 格式保存,音频文件采样率为 16 000 Hz,精度为 16 bit,采用单声道进行录制。

在此基础上,对原始情感语料库数据进行标注,采用多级别刻度方式,每种情感分为 4 个等级刻度表,等级 1 的情感表达最弱,等级 4 的表达最强,每个音频均需标识四类情感的等级。数据标注过程分为预判阶段和正式阶段,预判阶段时需要在独立标注 10 至 20 个音频的基础上,进行专家组商讨并确定标注规范,当多数专家观点一致时,可进行正式标注。

标注完毕后,使用迭代的优化贪婪算法进行专家置信度的更新和标注结果的判断。针对所有标注专家,每个音频的标注准确率与上一次可信度的均值作为基准值,然后分别计算每个专家的标注结果与基准值的相关系数作为衡量其新的可信度的指标。随着标记数量的增多,可信度的指标即时进行调整,在得到新的可信度指标后,重新计算当前的标注结果,即将所有

人的标注结果和权重加权求和,得到最终确定的标注刻度结果。具体公式如下。

$$\sum_{exp=1}^{n} W_{exp} = 1 \tag{1}$$

$$Avg_i = \frac{1}{n}\sum_{exp=1}^{n} Result_{exp} \times W_{exp} \tag{2}$$

$$\tilde{W}_{exp} = \frac{1 - \left|\frac{Result_{exp} - Avg_i}{Avg_i}\right|}{\sum_{exp=1}^{n}\left(1 - \left|\frac{Result_{exp} - Avg_i}{Avg_i}\right|\right)} \tag{3}$$

$$Av\tilde{g}_i = \frac{1}{n}\sum_{exp=1}^{n} Result_{exp} \times \tilde{W}_{exp} \tag{4}$$

式中:$W_{exp}$和 $\tilde{W}_{exp}$分别是上一次和本次的可信度;$n$ 是标注专家的数量;$Result_{exp}$是本次标注结果;$Avg_i$和 $Av\tilde{g}_i$分别是利用上一次和本次可信度计算的计算标注结果。

由于每个音频表达的情绪不止一种,基于此种方案,可获得音频每种情绪的表达级别。同时动态调整专家对整体标注结果的贡献率,提升语料库的整体评价水平。

## 2　个性化语音情感识别模型设计

### 2.1　总体设计

随着情感语音数据量的增加,采用传统的机器学习方法无法有效地处理高维数据,分析高阶的内部关联。基于此,可将目前流行的深度学习技术引入其中,深入挖掘情感特征与模型之间的隐藏关系。

然而,由于说话者之间的差异,导致语音信息并非是情感表达的唯一关键要素,因此,基于语音建立一个通用的情感的判别模型是非常困难的。在没有其他模态数据辅助的前提下,可以通过将说话者的特征与情感识别模型相结合来提高识别的准确率,此时建立的模型具有很强的个性化信息,在指定的应用场景内,针对每类说话者定向建立情感识别模型,通过类内模型的微调,识别针对类内某人的情感表达。

模型整体分为两个阶段:说话者分类阶段和语音情感识别阶段。前一个阶段使用多组大尺寸的 1 维 CNN,在定位说话人员所属类别的同时,提取倒数第二个隐藏层的特征。第二个阶段将针对个体说话者进行情感语音识别,除第一个阶段提取的特征之外,还添加语谱图特征和 CRNN 模型,融合两者进行微调训练,以达到最佳的情感识别效果。图 1 是模型整体设计思路。

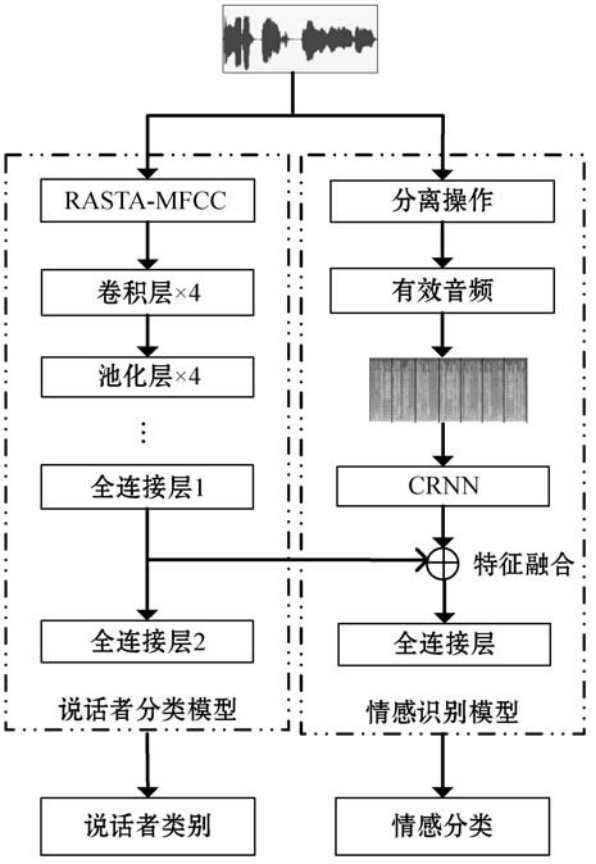


图 1　模型总体设计方案

### 2.2　说话者分类模型设计

目前,用于说话者识别的经典模型有高斯混合-通用背景模型(GMM-UBM)、联合因子分析(JFA)、i-vector[11]、x-vector[12]等,此类模型均是基于模板匹配的方法,从通用的模型中寻找最接近的说话者判别结果,这种形式适用于单任务的模型训练,且效果良好。

考虑到当前模型还需同时解决情感识别任务,如果仅针对个体识别创建模型,那么模型生成的中间结果将无法复用,此时将导致计算效率较低,浪费系统资源。基于此,本文的目标之一是寻找一种同时适用于说话者分类和情感识别的模型,将说话者的身份细化到某一类别,而并非某个人,同时配合各个阶段有效的特征表达,在保证识别准确率的同时,提升识别效率。

考虑到情感语音信号复杂度较高,而且含有未知的噪声,本文使用 RASTA(Relative Spectral)[13]滤波后的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficents,MFCC)[14]作为输入特征。MFCC 是目前语音情感识别中使用频率最高、最有效的谱特征,它是基于人耳的听觉机理而设计的。MFCC 一共有 13 个参数,可结合一阶和二阶差分共同使用,常用的 MFCC 为 1 - 4,其有效性较高。RASTA 滤波器通过对于声道的补偿,消除背景噪声对于短时频谱的负面影响,从而降低噪声的负面影响。

具体流程如下,在分帧和加窗的基础上,以帧为单位进行离散傅里叶变换,同时计算对数幅度频谱,在等带宽的梅尔滤波器组滤波和离散余弦变换的基础上,进行 RASTA 滤波,最终变换获得 RASTA-MFCC 特征。计算流程如图 2 所示。

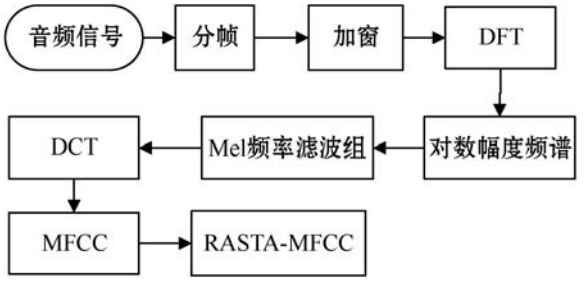


图 2 RASTA-MFCC 计算流程

在获得特征的同时,设计说话者分类模型,模型结构如图 3 所示。考虑到全部频带对于模型的影响,此处设计 4 个卷积层,均为大尺寸的一维卷积滤波器组,尺寸分别是 320 ×5,1 000 ×5,1 000 ×1,1 000 ×1,每类滤波器的步长均为 1,每个卷积层之间使用最大池化进行分隔,其后添加 2 个全连接层和 1 个 Softmax 层,从而获得说话者的分类信息。

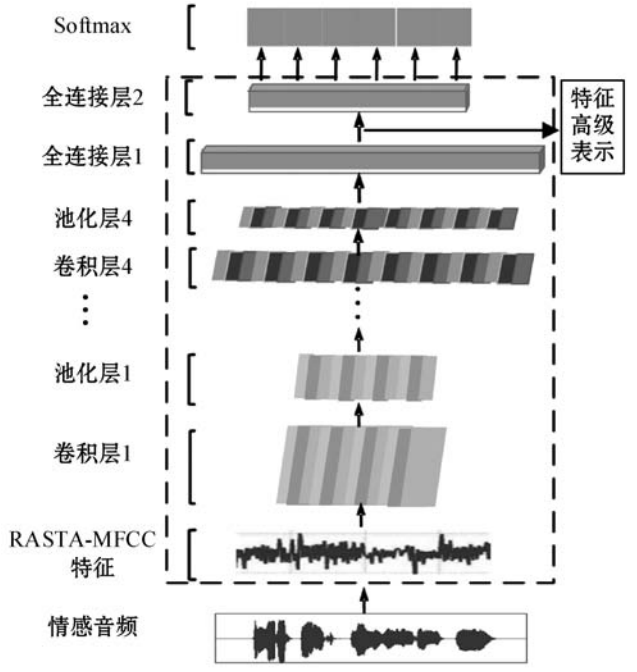


图 3 说话者分类模型

此模型在说话人分类时主要考虑 2 个要素:性别和基频。因此,模型的 Softmax 的初始类别数目是 5(2 个要素和 1 个其他类别),模型的输出为说话者所属类别,随着受试者人数的增多,模型的第 5 个类别(其他)将不断微调,当第 5 个类别数量与最多类别的数量相当时,将合并相似的声纹信息,分裂出新的类别。类别总数不超过 10 个。

由于不同的说话者类别在情感表达时的差异较大,为了进一步提升情感表达的精度,可以将说话者类别的特征作为附加语音情感特征,以缩小说话者类别对于情感表达识别产生的负面影响。

由于第 2 个全连接层的维度过少,本文考虑将说话者模型的第 1 个全连接层的输出用于情感特征的高级表达,与情感识别的特征组合进行第二阶段训练。

## 2.3 语音情感识别模型

由于不同说话者的发音习惯、发音方式、情感表达均不相同,其个性化的音频数据无法设计统一的识别模型参数,而且识别准确率会受到个体因素的影响。基于此,可针对上个阶段分类出的每位说话者,分别建立情感识别模型,该模型的特点是,采用通用的识别特征选择和识别模型的结构,但是通过深度学习获取各个模型的不同参数,从而突出个性化的特点。

在模型设计之前,首先需要完成语音信号与背景的信息分离,只保留与说话者声音有关的信息,可以将这个过程理解为简化版的去噪方案,此处选择软硬阈值折中的小波去噪方法。小波变换[15]在时频域都具有表征信号局部特征的能力,适合于环境噪声等背景信息的抽取。具体公式如下:

$$\lambda(j)=\delta\sqrt{2\log(N)}/\log(j+1) \tag{5}$$

$$\hat{w}_{j,k}=\mathrm{sgn}(w_{j,k})(|w_{j,k}|-a^{\lambda}) \quad |w_{j,k}|\geqslant\lambda \tag{6}$$

式中:$\lambda$ 为阈值;$\delta$ 是噪声强度;$j$ 是分解尺度;$N$ 为信号长度;$\hat{w}_{j,k}$、$w_{j,k}$分别表示估计前后的小波系数;$a$ 为参数,其范围为[0,1]。

通过小波去噪获得了表征能力较强的音频数据,然后针对此类数据进行特征提取,此时采用第一种手工制作的形式,将获得的音频信号进行时域和频域的切换,将其转化为频谱图的特征形式,此时原有的二维形式被转换成了三维的坐标形式,即语谱图。图 4 描述了语谱图的生成过程。此时将针对音频的处理转换为针对图像的处理过程,可采用深度学习中的图像处理技术辅助完成模型设计。

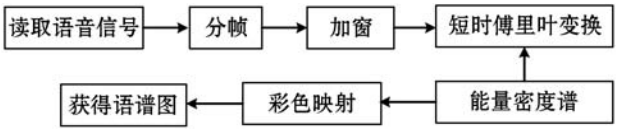


图 4 语谱图生成流程

针对语谱图,本文设计有效的 CRNN 模型。其中,CNN 模型与第一阶段相似,由 3 层卷积层、3 层池化层和 2 层全连接层,共 8 层构成,第一层卷积层的输入信息规模为 310 ×310 ×3,其中:310 为语谱图的长度和宽度;3 表示 RGB 三个通道。语谱图经过 64 个大小为 3 ×3 的卷积核,以步长为 1 的卷积操作后产生 64 个特征图,然后使用 ReLU 激活函数,经过最大池化操作

后得到 64 个特征图,第 2 层卷积层的输入源即第 1 层的输出特征图,计算过程与第 1 层一样,第 3 层同理,接下来是 2 层全连接层,每层为 1 024 个神经元,在此层上做 Dropout 操作,防止模型过拟合。

由于语音信号是基于时间序列的信息,其上下文之间存在着一定的关联,因此,除了设计适用于图像识别的 CNN 之外,同时考虑增加具有短期记忆能力的神经网络模型,引入 LSTM 来控制信息的累积速度,有选择地加入新的信息,并有选择地遗忘之前积累的信息。

此处采用了双向 3 层的 LSTM 模型,双向是指存在两个信息传递相反的循环层,第 1 层按时间顺序传递信息,第 2 层按时间逆序传递信息。它意味着过去和未来的信息均可以成功捕获,这是由于情感的时序因素,它可以由前后若干帧的信息共同决定,因此按照上述思路设计了 3 组双向 LSTM 模型,以利用上下文的个性化信息进行更准确的情感判断和参数学习。

语音情感识别模型如图 5 所示。除 CRNN 模型之外,在第 1 阶段获取的高级特征表示被添加至其中,与此时获取的特征共同完成训练过程,两组特征集合均为 1 024 维。其中,个性化特征体现在以下 3 处:

(1) 高级特征表示由每个语音独立生成,是上一个阶段模型的产物。

(2) 此处的 CRNN 模型为每一个说话人类别的定向模型,即针对每类说话人分别进行训练所得。

(3) 原始说话人分类依据:性别和基频,为每个类别提供了原始的通用信息,一定程度上抑制其他类别的混入噪声。

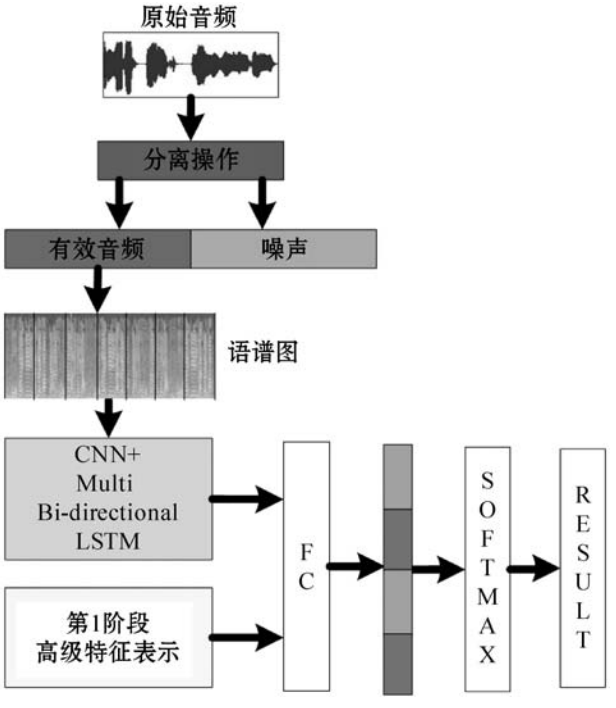


图 5　语音情感识别模型

## 3　实验设计与结果分析

### 3.1　实验准备

本文分别使用自建成人自然情感语料库和 Interactive Emotional Dyadic Motion Capture (IEMOCAP) 情感语料库进行实验。

自建成人自然情感语料库现有 13 500 余条有效语音,采用双重标注信息,第一层为情感标注,主要包括高兴、愤怒、平静和悲伤等 4 类情感。其中每类情感数据量较均衡。第二层为说话人分类标注,包括高基频(男和女)、低基频(男和女)、其他等 5 类。随着训练数据的增加,其他类别可再次分裂。受试者均为成年男女,一共为 16 人,其中男女各占 50%,以 18 至 30 岁为主,少数 30 至 40 岁。

IEMOCAP 数据集是使用动作、音频、视频录制的具有 10 个主题的 5 个二元会话中收集的,侧重于表达二元相互作用。每个会话由一个男性和一个女性演员执行脚本,并参与通过情感场景提示引发的自发的即兴对话。此数据集一共有 10 039 个标准语音,仅包含情感标注信息。需要将相关的同类情感进行合并操作,去除关联度较小的样本,最终使用 4 类情感数据:将 excited 类与 happiness 类别合并,除此之外,还有 sad 类别、angry 类别和 neutral 类别。其余类别的样本数据均被丢弃。基于此种分类方法,共保留 5 531 个样本,每类样本的数据量为 angry:1 103,happy:1 636,neutral:1 708,sad:1 084。

除了 angry 和 sad 类别的样本量偏少之外,其他类别的情绪样本数据量较均衡。

针对两个数据集,分别使用五折交叉验证方法进行实验。80% 数据用于训练深度神经网络,剩余的数据被用于验证和准确性测试。

在对语音数据进行预处理时,标准窗口大小为 25 ms,偏移量为 10 ms。特征被标准化为零均值。

在 CNN 和 CRNN 模型中,*Batch* 的大小为 100,最大轮次数为 100 000。同时设置学习速率为 0.001。Dropout 为 0.5。采用 ReLU 作为激活函数,Adam 作为优化器,使用均方误差作为损失函数。

### 3.2　说话者分类实验

针对说话者特征的分类,设计相关的实验,利用自建成人自然情感语料库进行训练,通过自建成人语料库和 IEMOCAP 数据集进行测试。使用 TensorFlow 框架进行网络模型结构的搭建,本文将当前说话者识别模型与 i-vector、x-vector 和基于 VGG 网络的方法进行

比较[16]。其中,基线:i-vector;模型 1:VGG;模型 2:x-vector(PLDA);模型 3:CNN(MFCC);模型 4:当前模型 CNN(RASTA-MFCC)。

表 1 和表 2 仅列出自建成人语料库的说话者分类模型的测试结果和不同说话者类别比例。

**表 1 说话者分类模型的测试结果**

| 模型 | EER | DCF |
|---|---|---|
| 基线:i-vector | 5.4 | 0.45 |
| 模型 1:VGG | 7.8 | 0.48 |
| 模型 2:x-vector(PLDA) | 7.1 | 0.57 |
| 模型 3:CNN(MFCC) | 6.6 | 0.54 |
| 模型 4:当前模型 | 5.4 | 0.45 |

**表 2 不同说话者类别比例**

| 说话者类别 | 类别占比/% |
|---|---|
| 男、基音高 | 15 |
| 男、基音低 | 28 |
| 女、基音高 | 27 |
| 女、基音低 | 14 |
| 其他 | 16 |

由表 1 中的测试结果可知,在相同数据源的条件下,本文提出的模型与 i-vector 效果持平,但明显优于 VGG 方法和 x-vector。与 i-vector 相比,除了可以获得相似声纹的数据之外,当前模型还可以获得语音情感的高维表达,进一步提升情感识别的准确率。表 2 中提供了自建成人语料库的说话者分类信息,可以看出,84% 的说话者可以隶属于前 4 个分类,其他类别的说话者比例较低,因此无须分裂出第 5 个类别。

## 3.3 语音情感识别实验

针对语音中情感表达的识别,利用自建成人自然情感语料库和 IEMOCAP 数据集进行训练和测试,使用 TensorFlow 框架进行网络模型结构的搭建,为了避免不同情感数量不均衡产生的影响,本文采用加权精度(Weighted accuracy,WA)和未加权精度(Unweighted accuracy,UA)作为指标,针对不同的情感分类模型进行测试。

实验以未使用说话者分类特征的 CRNN 模型作为基线,其输入语音为原始音频,未经任何处理。同时对比以下几个模型,模型 1:处理后音频 + 单向 3 层 LSTM;模型 2:处理后音频 + 双向 3 层 LSTM;模型 3:处理后音频 + CRNN;模型 4:当前模型(处理后音频 + CRNN + 第一阶段高级表达)。这里的 UA 和 WA 分别代表所有类别模型准确率的平均值,分别计算每个模型的情感识别的准确性。表 3 为经过实验验证后,不同语音情感识别模型的准确度。

**表 3 语音情感识别模型的测试结果(%)**

| 模型 | 自建语料库 UA | 自建语料库 WA | IEMOCAP UA | IEMOCAP WA |
|---|---|---|---|---|
| 基线:未使用说话者分类特征 | 58 | 57 | 51 | 50 |
| 模型 1:处理后音频 + 单向 3 层 LSTM | 64 | 63 | 58 | 57 |
| 模型 2:处理后音频 + 双向 3 层 LSTM | 67 | 67 | 60 | 61 |
| 模型 3:处理后音频 + CRNN | 74 | 73 | 65 | 64 |
| 模型 4:当前模型 | 78 | 78 | 71 | 70 |

由表 3 可知,针对两个数据集合,当前模型的表现最佳,拥有最优的平均 WA 和 UA,超过未使用说话者分类特征的模型和未处理音频数据的模型。由此可以确定,融合了说话者分类特征的模型可以提升情感识别的精度,确定了语谱图对于情感识别任务的积极作用。

图 6 描述了针对自建语料库,当前情感识别模型的误差变化趋势,以 *Batch* 的大小作为衡量周期,可以看出,平均在 *Batch* 为 1 900 时,模型趋于稳定状态。

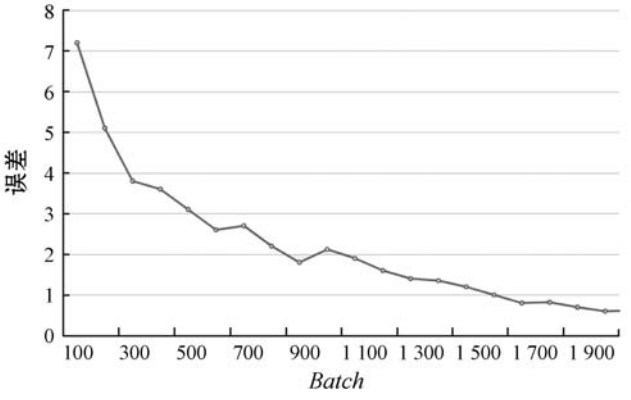


图 6 语音情感识别模型误差

表 4 是针对自建语料库中的音频,使用当前模型进行情感识别的混淆矩阵。可以看出,对于唤醒度较高的情绪,识别准确度较高,例如高兴、愤怒等类别。反之,针对平静、悲伤等唤醒度较低的类别,识别准确率较低。

**表 4 语音情感类别混淆矩阵(%)**

| 准确率 | 高兴 | 愤怒 | 平静 | 悲伤 |
|---|---|---|---|---|
| 高兴 | 82 | 9 | 5 | 4 |
| 愤怒 | 9 | 83 | 3 | 5 |
| 平静 | 5 | 3 | 74 | 18 |
| 悲伤 | 4 | 5 | 18 | 73 |

## 4 结　语

从语音中识别特定的情感是一项具有挑战性的任务,其结果常常依赖于语音信号特征的准确性和模型的有效性。本文设计一种针对个性化特征的、结合说话者分类任务、多级别特征、识别准确率较高的深度学习模型。在多任务语音情感特征提取、个性化神经网络模型设计和成人自然情感语料库设计等方面开展了相关的研究,通过实验验证,本文模型的识别准确度较高。

在未来的研究过程中,将从语音识别入手,寻求一种通用的网络结构,结合显著性区域特征,实现对于语音情感识别任务的泛化能力和效率的提升;考虑到长语音中可能夹杂多种不同的情感,将考虑通过模型的调整实现多标签的语音情感识别。

## 参 考 文 献


[1] Rajasekhar B, Kamaraju M, Sumalatha V. Glowworm swarm based fuzzy classifier with dual features for speech emotion recognition[J]. Evolutionary Intelligence, 2022, 15: 939 – 953.

[2] Mohammed A, Carlos B. Domain adversarial for acoustic emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2423 – 2435.

[3] Ramanarayanan V, Pugh R, Qian Y, et al. Suendermann-Oeft: Automatic Turn-Level language identification for Code-Switched Spanish-English dialog[C]//International Workshop on Spoken Dialog Systems, 2018.

[4] Mao Q, Ming D, Huang Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. IEEE Transactions on Multimedia, 2014, 16(8): 2203 – 2213.

[5] Juvela L, Bollepalli B, Tsiaras V, et al. GlotNet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(6): 1019 – 1030.

[6] Zhang X L, Wang D L. Boosting contextual information for deep neural network based voice activity detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(2): 252 – 264.

[7] Wang X, Tang M, Yang S. Automatic hypernasality detection in cleft palate speech using CNN[J]. Circuits Systems and Signal Processing, 2019, 38(8): 3521 – 3547.

[8] Zhang T, Wu J. Discriminative frequency filter banks learning with neural networks[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2019(1): 1 – 16.

[9] Ma X, Wu Z, Jia J, et al. Emotion recognition from variable-length speech segments using deep learning on spectrograms[C]//Interspeech, 2018: 3683 – 3687.

[10] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language Resources & Evaluation, 2008, 42(4): 335 – 359.

[11] Ribas D, Vincent E. An improved uncertainty propagation method for robust I-vector based speaker recognition[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 2019.

[12] You L, Guo W, Dai L R. Multi-Task learning with High-Order statistics for x-Vector based Text-Independent speaker verification[C]//Interspeech, 2019.

[13] Korba M C A, Messadeg D, Bourouba H, et al. Noise robust features based on MVA post-processing[J]. IFIP Advances in Information and Communication Technology, 2015, 456: 155 – 166.

[14] Bandela S R, Kumar T K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC[C]//2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017.

[15] Şişman B, Li H, Tan K C. Transformation of prosody in voice conversion[C]//2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017.

[16] Shon S, Tang H, Glass J. Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model[C]//2018 IEEE Spoken Language Technology Workshop(SLT), 2018.


(上接第 107 页)


[10] Ramaji I J, Memari A M, Messner J I. Product-oriented information delivery framework for multistory modular building projects[J]. Journal of Computing in Civil Engineering, 2017, 31(4): 1 – 17.

[11] 金杰,夏超,肖士利,等. 基于数字孪生的火箭起飞安全系统设计[J]. 计算机集成制造系统,2019,25(6):1337 – 1347.

[12] 罗家文. 数字化车间实时三维可视化监控关键技术研究[D]. 南京:南京航空航天大学,2019.

[13] 彭成吉,何文雪,刘赫. 基于 PLC 的船舶机舱监测报警系统改进方案[J]. 工业控制计算机,2014,27(4):38 – 39.

[14] 张勇亮,张均东,张志政. 三维船舶轮机虚拟实验室的设计和实现[J]. 计算机应用与软件,2019,36(1):177 – 181.