ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 542－554
DOI:10. 16337/j. 1004-9037. 2022. 03. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 基于深度残差收缩网络多特征融合语音情感识别

李瑞航，吴红兰，孙有朝，吴华聪

(南京航空航天大学民航学院,南京 211106)

**摘　要**：针对语音情感识别任务中说话者的差异性,计算谱特征的一阶差分、二阶差分组成三通道的特征集输入二维网络。结合卷积神经网络、双向长短时记忆网络以及注意力机制建立基线模型,引入深度残差收缩网络分配二维网络中的通道权重,进一步提高语音情感识别的精度。为提升模型的学习效果,采取特征层融合(特征向量并行和特征向量拼接两种方式)和决策层融合(平均得分和最大得分两种方式)等不同信息融合机制。结果表明:(1)特征层融合中的特征向量并行策略是更有效的方式;(2)本文提出模型在CASIA和EMO-DB数据库下分别取得了84.93%和86.83%的未加权平均召回率(Unweighted average recall, UAR),相较于基线模型,引入深度残差收缩网络后的模型在CASIA和EMO-DB数据库上的未加权召回率分别提高5.3%和6.2%。

**关键词**：深度学习;语音情感识别;深度残差收缩网络;注意力机制;多特征融合

**中图分类号**：TP391　　**文献标志码**:A

## Multi-feature Fusion Speech Emotion Recognition Based on Deep Residual Shrinkage Network

LI Ruihang, WU Honglan, SUN Youchao, WU Huacong

(College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

**Abstract**: Aiming at the difference of speakers in speech emotion recognition task, calculate the first-order difference and second-order difference of spectral features to form three-channel feature sets and input the feature sets to the two-dimensional network. The convolutional neural network, bidirectional short and long memory network and attention mechanism were combined to establish a baseline model, and the deep residual shrinkage network was introduced to allocate channel weights in the two-dimensional network to further improve the accuracy of speech emotion recognition. In order to improve the learning effect of the model, two different information fusion mechanisms, feature layer fusion (Add and Concatenate) and decision layer fusion (Average and Maximum), were adopted. The results show that :(1) Add strategy in feature layer fusion is more effective; (2) The proposed model achieves 84.93% and 86.83% of unweighted average recall (UAR) in CASIA and EMO-DB databases respectively. Compared with the baseline model, the unweighted recall rates of CASIA and EMO-DB are increased by 5.3% and 6.2% respectively after introducing deep residual shrinkage network.

**Key words:** deep learning; speech emotion recognition(SER); deep residual shrinkage network(DRSN); attention mechanism; multi-feature fusion

---

**基金项目**:国家自然科学基金(U2033202,52172387,U1333119)。
**收稿日期**:2021-12-28;**修订日期**:2022-03-25

## 引　言

语音是人类最广泛使用的交流方式之一,不仅传达了显性的语言内容,还隐含着说话人的情感信息[1]。目前语音情感识别的主要特征包括韵律特征与谱特征[2]。

韵律特征主要包括基频相关特征以及能量相关特征等[3],韵律特征按照全局特征和局部特征,可提取最大值、最小值、均值及方差组成高维特征集。

谱特征能够反映信号在频域上的差异性,谱特征分为线性频谱特征和倒谱特征[4]。常用的线性谱特征有线性预测系数(Linear prediction coefficients, LPC)、对数频率功率系数(Log frequency power coefficients, LFCC)。常用的倒谱特征有梅尔频率倒谱系数(Mel-frequency cepstral coefficient, MFCC)、线性预测倒谱系数(Linear predictive cepstral coefficient, LPCC)等。Bou-Ghazale等[5]的研究表明倒谱特征对语音情感的区分能力明显优于线性谱特征。

2017年,Han等[6]提出了一种基于高斯核非线性近端支持向量机,采用了16维特征,其中前9维为韵律特征,后7维为谱特征,包括能量、共振峰、谐波噪声比等特征。2018年,Hsiao等[7]提取了语音帧中基频、频率微扰、过零率、能量、谐波噪声比、MFCC等特征,计算极值、方差、中值、均值、偏度、最小值、最大值、峰度、线性回归系数,采用深度RNN模型在FAU-Aibo任务动态建模框架中将未加权平均召回率(Unweighted average recall, UAR)从37.00%提高到了46.30%。2021年,胡德生等[8]提取了语音帧中的平均过零率、能量、基音频率、共振峰、MFCC等特征,采取主辅网络特征融合的方式,在IEMOCAP数据集上取得72.50%的非加权准确率。

上述研究所提出的模型,将不同特征输入多个网络中,但采用固定的信息融合方式。本文针对语音特征的多样性,采取不同的融合策略对多特征进行信息融合,应用此方法能够更有效地利用语音信号的特征。

语音情感识别模型可分为基于传统机器学习和基于深度学习[9]。传统应用于语音情感识别的机器学习的模型有高斯混合模型[10]、隐马尔可夫模型[11]、支持向量机[12]、多层感知器[13]和递归神经网络[14]等。

近年来,随着深度学习框架的发展,深度学习模型以较高的识别精度优势在语音情感识别领域得到大量应用。其中,卷积神经网络(Convolutional neural network, CNN)和长短期记忆网络(Long short-term memory, LSTM)在语音情感识别领域得到了大量的应用[15-16]。在最近的研究中,进一步表明引入注意力机制[17]和使用双向长短时记忆网络[18]能进一步提高识别准确率。针对语音帧所蕴含情感信息量不同的特点,引入注意力机制,增强有效语音帧的权重,减弱无效语音帧的权重。当前语音信号蕴含情感与前后语音帧皆相关,采用双向长短时记忆网络可获取前后时间依赖特征。

然而,语音情感特征具有个体差异性。为降低说话者差异的影响,对谱特征进行差分得到特征集形成多个通道输入到2D-CNN,3D-CNN网络中。2018年,Chen等[19]提出3D-CNN-LSTM模型,将MFCC系数以及其一阶、二阶差分作为多通道一起输入3D-CNN模型,在IEMOCAP和EMO-DB两个数据集上将未加权平均召回率分别达到64.74%和82.82%。2019年,Zhao等[20]将对数梅尔特征(Log-Mel)以及其一阶、二阶差分作为多通道输入2D-CNN-LSTM模型,在IEMOCAP数据库中,基于说话人依赖实验和基于说话人独立实验的识别准确率分别为89.16%和52.14%。2021年,徐华南等[21]计算语音信号的对数梅尔特征和其一阶差分和二阶差分特征,合并成3D Log-Mel特征集,在IEMOCAP和EMO-DB数据库上分别得到61.22%和85.69%的平均准确率。

上述方法为降低说话者差异性,都将语音信号谱特征进行了差分处理,输入到多通道的卷积网络中,但未对通道权重进行考虑。而不同阶差分处理的特征对语音情感的区分度可能出现差异性,需要分配不同的通道权重以增加语音情感识别的准确度。深度残差收缩网络(Deep residual shrinkage network, DRSN)[22]将软阈值化作为非线性层,易于提高在噪声数据的深度学习效果,适用于给重要特征分配更大的权重,滤除不重要特征。因此,本文引入深度残差收缩分配特征通道权重,以提高有效特征权重,降低冗余特征权重,进一步提高语音识别精度。

## 1 语音情感特征

韵律特征和谱特征都是描述情绪状态的有效特征[23]。为减小训练计算量,本文选取韵律特征中语音短时能量、短时平均幅度、短时过零率等典型特征。鉴于倒谱特征区分能力明显优于线性谱特征,本文选取梅尔频率倒谱系数、线性频率倒谱系数、线性预测倒谱系数等谱特征。

### 1.1 韵律特征

(1)短时能量

不同情感的表达在语音信号的幅度值上的体现有所不同,将语音短时能量作为判断语音情感的特征之一。语音短时能量$E_n$是一个表征语音信号幅度值变化的函数

$$E_n=\sum_{m=0}^{N-1}x_n^2(m) \tag{1}$$

(2)短时平均幅度

短时能量$E_n$对高电平非常敏感,因此,可使用短时平均幅度$M_n$度量语音信号幅度值变化

$$M_n=\sum_{m=0}^{N-1}\left|x_n(m)\right| \tag{2}$$

(3)短时过零率

短时过零率表示分帧后语音信号中一帧语音波形穿过横轴的次数

$$Z_n=\frac{1}{2}\sum_{m=0}^{N-1}\left|\mathrm{sgn}\left[x_n(m)\right]-\mathrm{sgn}\left[x_n(m-1)\right]\right| \tag{3}$$

语音低频部分和高频部分分别具有较低和较高的平均过零率,可以以此区分轻音和浊音,进而反映声带振动情况,作为区分情感的特征之一。

### 1.2 谱特征

(1)梅尔频率倒谱系数

MFCC可有效表征声道共振信息。在1980年,MFCC由Davis等提出。此后在语音识别领域,MFCC成为运用最广泛的特征参数[24]。MFCC的计算步骤如下:

①通过高通滤波器对语音信号进行预加重处理,提高高频部分使信号频谱变得平缓

$$H(z)=1-\mu z^{-1} \tag{4}$$

②对语音信号按照帧长为25 ms,移帧为10 ms进行分帧并采用汉明窗进行加窗。

③进行快速傅里叶变换,将时域信号转化为到频域信号,得到能量分布

$$X_a(k)=\sum_{n=0}^{N-1}x(n)\mathrm{e}^{-\frac{\mathrm{j}2\pi k}{N}}\quad 0\leqslant k\leqslant N \tag{5}$$

④将能量谱通过Mel尺度的三角滤波器,对谱进行平滑化

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m-1)\leqslant k<f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m)\leqslant k<f(m+1)\\ 0 & k\geqslant f(m+1)\end{cases} \tag{6}$$

⑤计算从滤波器输出的对数能量,进行离散余弦变换,得到MFCC系数

$$s(m)=\ln\left(\sum_{k=0}^{N-1}\left|X_a(k)\right|^2 H_m(k)\right)\quad 0\leqslant m\leqslant M \tag{7}$$

$$C(n)=\sum_{m=0}^{N-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right)\quad n=1,2,\cdots,L \tag{8}$$

(2)线性频率倒谱系数

LFCC与梅尔频率倒谱特征提取过程相同,但其滤波器组频率按照线性频率分布。

(3)线性预测倒谱系数

LPCC利用线性预测分析获得倒谱系数。该特征描述元音效果较好,描述辅音效果较差。LPCC的提取过程如下:

①通过线性预测分析得到全极点模型为

$$H(z)=\frac{G}{1-\sum_{i=1}^{p}a_i z^{-1}} \tag{9}$$

②浊音的激励模型可表示为

$$U(z)=E(z)G(z)=\frac{A_v}{1-Z^{-1}}\cdot\frac{1}{(1-\mathrm{e}^{-cT}z^{-1})^2} \tag{10}$$

③输入输出的关系表示为

$$s(n)=\sum_{k=1}^{P}a_k s(n-k)+Gu(n) \tag{11}$$

④如果采样点$n$输出$s(n)$可用前面$p$个样本的线性组合来表示,则$a_1,a_2,\cdots,a_p$为常数值,称为线性预测系数

$$\tilde{s}(n)\approx a_1 s(n-1)+a_2 s(n-2)+\cdots+a_p s(n-p) \tag{12}$$

⑤线性预测倒谱系数$c(n)$为

$$c(n)=\begin{cases}0 & n<0\\ \ln G & n=0\\ a_n+\sum_{k=1}^{n-1}\left(\dfrac{k}{n}\right)c(k)a_{n-k} & 0<n\leqslant p\\ \sum_{k=n-p}^{n-1}\left(\dfrac{k}{n}\right)c(k)a_{n-k} & n>p\end{cases} \tag{13}$$

## 2 模型构建

### 2.1 基线模型

本文引入注意力机制增强有效语音帧的权重,减弱无效语音帧的权重,结合CNN、BLSTM模型构建基线模型。

(1)CNN模型

将韵律特征输入1D-CNN网络。计算谱特征以及其一阶和二阶差分形成3个通道的特征集,输入到2D-CNN网络。1D-CNN设置两个卷积层,卷积核的数目为128,大小为5,步长为1,激活函数使用Relu。设置两个池化层,池化大小分别为5和3。2D-CNN设置3个卷积层,卷积核数目为64,卷积核大小分别为$5\times5$,$5\times5$,$3\times3$,步长为1,激活函数使用Relu。设置3个池化层,池化大小皆为$2\times2$。

(2)双向LSTM模型

当前语音信号蕴含情感不仅与前面的语音帧相关,还与后面的语音帧相关。所以需要使用BLSTM,用独立的两个LSTM网络从两个方向处理语音序列。BLSTM在时间$t$时刻隐藏状态输出结果为

$$h_t=[\overrightarrow{h_t},\overleftarrow{h_t}] \tag{14}$$

式中:$\overrightarrow{h_t}=\overrightarrow{\mathrm{LSTM}}(x_t,\overrightarrow{h_{t-1}})$为前向LSTM在$t$时刻的隐藏输出结果,$\overleftarrow{h_t}=\overleftarrow{\mathrm{LSTM}}(x_t,\overleftarrow{h_{t-1}})$为后向LSTM在$t$时刻的隐藏输出结果。

最终隐藏层输出为

$$\boldsymbol{H}=(h_1,h_2,\cdots,h_t) \tag{15}$$

(3)注意力机制

将BLSTM层输出的隐藏层$\boldsymbol{H}=(h_1,h_2,\cdots,h_t)$作为注意力层的输入,$\boldsymbol{H}\in\mathbf{R}^{t\times d}$,$t$为语音帧数,$d$为BLSTM隐藏层的大小。

$$e_i=\tanh(W_jh_i+b_j) \tag{16}$$

$$\alpha_i=\frac{\exp(e_i)}{\sum_i\exp(e_i)} \tag{17}$$

$$h_i'=\alpha_ih_i \tag{18}$$

式中:$\alpha_i$为注意力权重,$h_i'$为按照语音帧加权后的特征值。

在卷积神经网络后加入引用注意力机制的双向长短期记忆网络,BLSTM网络每个方向包含64个节点,输出一个128维的序列。通过注意力机制,对包含情感信息较为丰富的语音帧分配较大的权重。

## 2.2 信息融合

信息融合的目的是将不同模型的优势结合一起,起到互补缺点的作用。信息融合分为特征层融合和决策层融合[25]。

特征层融合是将原始特征输入到多个深度学习网络中,得到多个降维特征向量,融合得到单个特征向量,然后输入分类器中。特征层融合常采用的方式有两种,特征向量并行(Add)方式和特征向量拼接(Concatenate)方式。并行方式需要所有网络输出相同维度的降维特征向量进行叠加,能够增加每一维特征的信息量,而不改变特征向量的维数。拼接方式则是将降维特征向量进行串联,能够增加特征向量的维数,而不增加每一维的信息量。

决策层融合是通过代数组合规则对多个网络识别结果进行融合,每个网络都会有一个预测评分。这种方式每个网络的分类结果都是独立的。常见的决策层融合方法有取分数的平均值(Average)、最大值(Maximum)等。

## 2.3 深度残差收缩网络

深度残差收缩网络引入软阈值化作为非线性层。软阈值化的本质是设计滤波器将噪声信号转化为接近为零的特征,是信号去噪方法的一个关键步骤。在深度残差网络结构中应用软阈值化的构建深度残差收缩网络,提高含噪数据或复杂数据上的特征学习效果。

软阈值的作用可表示为

$$y=\begin{cases}x-\tau & x>\tau\\ 0 & -\tau\leqslant x\leqslant\tau\\ x+\tau & x<-\tau\end{cases} \tag{19}$$

式中:$x$为输入特征;$y$为输出特征;$\tau$为阈值,即一个正参数。

输出对输入的导数为1或0,可有效防止梯度消失和爆炸问题。其导数可表示为

$$\frac{\partial y}{\partial x}=\begin{cases}1 & x>\tau\\ 0 & -\tau\leqslant x\leqslant\tau\\ 1 & x<-\tau\end{cases} \tag{20}$$

堆叠多个各通道不同阈值的残差收缩模块(Residual shrinkage building unit with channel-wise thresholds, RSBU-CW)则可得到深度残差收缩网络,如图1所示构建RSBU-CW,在每个通道上都应用一个单独的阈值。对特征中的每个元素取绝对值,利用全局平均池化将特征$x$映射为一个一维向量,然后输入到两个全连接(Fully connected, FC)层。其中第二个全连接层的神经元数量为输入特征映射的特征通道数量,全连接层的输出则被缩放到(0,1),即

$$\alpha_c=\frac{1}{1+\mathrm{e}^{-z_c}} \tag{21}$$

式中:$c$为特征通道序号,$\alpha_c$为第$c$个通道的缩放参数,$z_c$是第$c$个通道的输出特征。

下一步可计算第$c$个通道的阈值为

$$\tau_c=\alpha_c\cdot\operatorname*{average}_{i,j}\left|x_{i,j,c}\right| \tag{22}$$

式中:$\tau_c$为第$c$个通道的阈值,$x_{i,j,c}$则为特征$x$的通道$c$下坐标为$(i,j)$的特征。

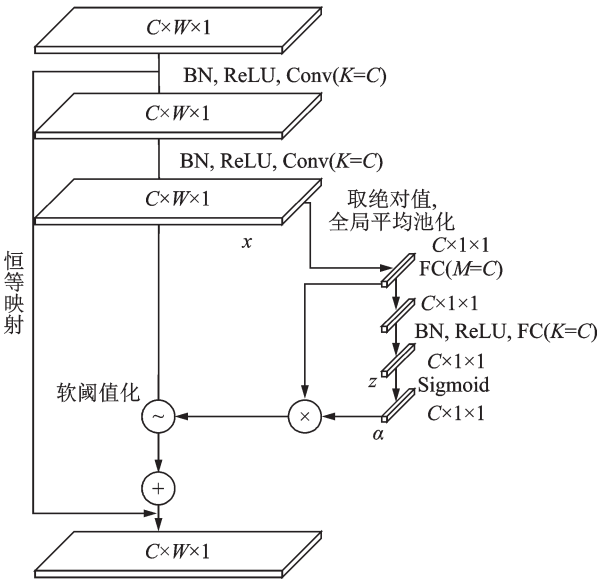



图1 RSBU-CW

Fig.1 RSBU-CW

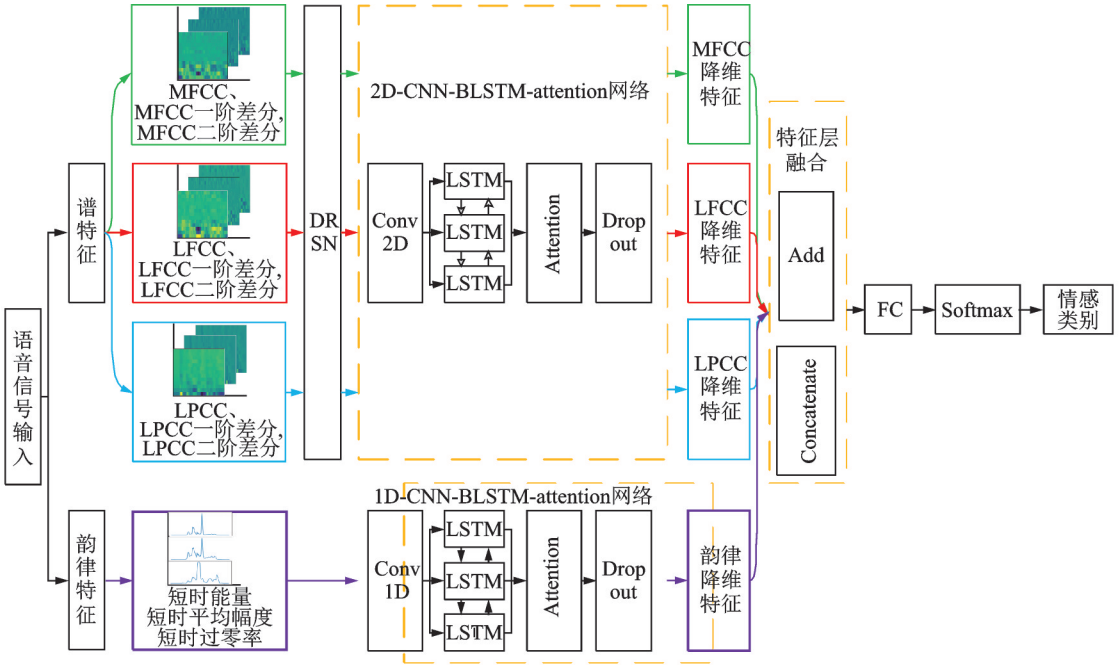
深度残差收缩网络也适用于非噪声数据,因为其阈值是由样本自适应确定。样本中若不含冗余信息,阈值可被训练得非常接近于零,从而软阈值化几乎不会对模型造成影响。本文将谱特征经过一阶、二阶差分得到3个通道,在输入二维网络之前,通过深度残差收缩网络获得通道权重。通过这种方式,每组训练样本都可有自己独特的一组通道权重,结合样本自身特点,对特征通道进行加权调整,从而得到具有通道权重的卷积神经网络,提升深度学习结果。

### 2.4 DRSN-MF模型

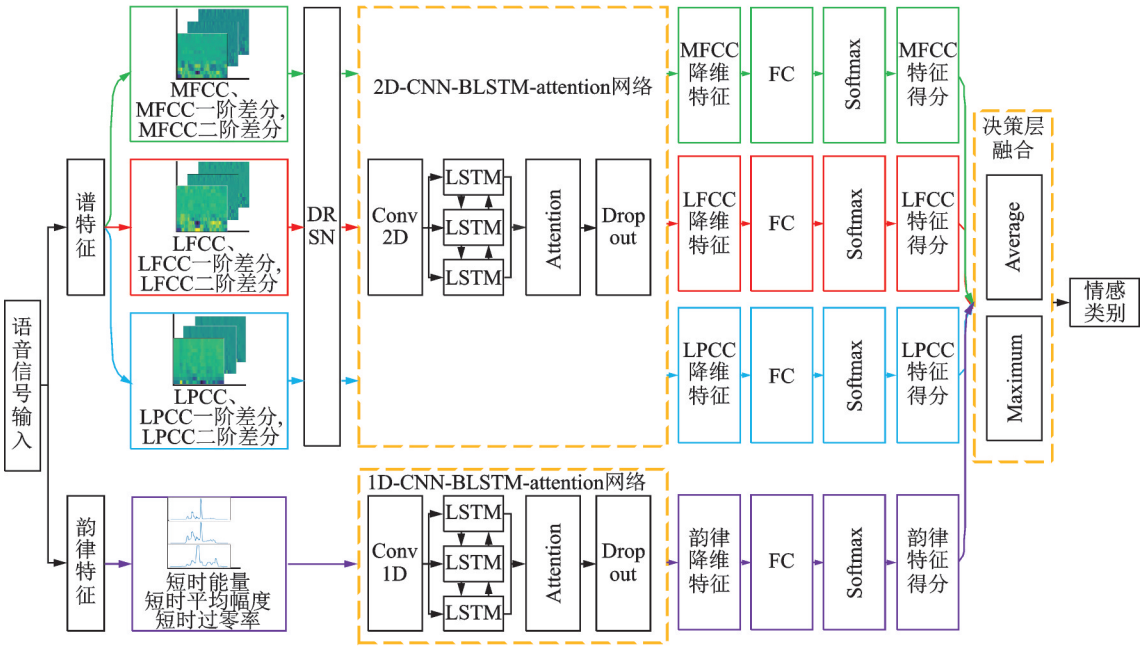
按照第1节的计算流程对预处理后的语音信号提取语音情感特征。深度残差收缩网络多特征融合(Deep residual shrinkage network with multi-feature fusion, DRSN-MF)模型包含一维网络(1D-CNN-BLSTM-attention)和二维网络(2D-CNN-BLSTM-attention)。两个网络中都引入了注意力机制,提高有效语音帧的权重,以提高情感识别效果。两个网络最后都引入一个Dropout层,防止过拟合,提升模型泛化能力。二维网络与一维网络的差异是采用了2D-CNN,其输入输出相比1D-CNN多出了一个维度,可利用这个维度实现多通道的输入。将语音信号韵律特征输入到一维网络。将语音信号谱特征输入二维网络之前,先计算谱特征的一阶和二阶差分形成3个通道的特征集,引入深度残差收缩网络获得二维网络3个通道的权重,再将谱特征输入。MFCC、LFCC、LPCC都采用这种方式输入二维网络。

不同特征经过深度学习网络后可得到对应的降维特征,为更好地利用降维特征,研究其在特征层

融合和在决策层融合的差异。在特征层融合中采取特征向量并行方式和拼接方式，对降维特征进行融合后通过全连接层，采用Softmax函数对情感进行分类，如图2(a)所示。决策层融合采取了取平均得分和最大得分方式，先通过全连接层和Softmax函数得到每一类降维特征对情感的分类预测分数，最后通过代数组合规则输出情感分类结果，如图2(b)所示。




(a) Feature layer fusion




(b) Decision layer fusion

图2 DRSN-MF模型

Fig.2 DRSN-MF model

## 3 数据库及实验结果

### 3.1 情感数据库

为验证本文所提出模型以及特征融合策略的有效性,本文选用两个公开数据集CASIA中文数据集和EMO-DB德语数据集进行实验。

CASIA中文数据集由中国科学院自动化研究所录制,共9 600条情感数据,包括6种情感,分别为生气、害怕、开心、中性、悲伤、惊讶。采样率为16 kHz,采用16 bit量化级数据[26]。由于该数据库并不完全对外开放,故本文只选用了其中1 200条语音。

EMO-DB德语数据集由德国柏林工业大学录制,共535条情感数据,包括7种情感,分别为生气、无聊、恶心、害怕、开心、中性、悲伤。采样率为48 kHz,采用16 bit量化级数据[27]。

### 3.2 关键指标

语音情感识别本质上为一个多分类任务,其评价指标包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)、$F_1$值($F_1$ score)[28]。对于二分类任务,样本只存在两个类别,即正样本和负样本。对于一个随机样本,根据其预测类别和实际类别可分为以下4种情况:(1)真阳性(True positive, TP),预测为正样本实际也为正样本;(2)假阴性(False negative, FN),预测为负样本实际为正样本;(3)真阴性(True negative, TN),预测为负样本实际也为负样本;(4)假阳性(False positive, FP),预测为正样本实际为负样本。

(1)准确率

准确率为正确预测的样本数量占总样本数量的比例。

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{23}$$

(2)精确率

精确率为预测为正的样本数量占真实为正的样本数量的比例。

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{24}$$

(3)召回率

召回率为预测为正的样本数量占预测为正的样本数量的比例。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{25}$$

(4)$F_1$值

$F_1$值为精确率和召回率的调和平均值。

$$F_1=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{26}$$

多分类任务中计算每个类别的精确率和召回率时,将每个类别单独视为“正”,所有其他类型视为“负”。所有类别计算得到的精确率$\boldsymbol{R}$和召回率$\boldsymbol{P}$可表示为一个$n$维的向量,$n$表示样本的类别数量,$R_i$和$P_i$分别表示第$i$个类别的精确率和召回率。

$$\boldsymbol{R}=(R_1,\cdots,R_i,\cdots,R_n) \tag{27}$$

$$\boldsymbol{P}=(P_1,\cdots,P_i,\cdots,P_n) \tag{28}$$

对各分类的精确率求平均值即为多分类任务的未加权平均精确率$R$,对各分类的召回率求平均值即为多分类任务的未加权平均召回率$P$。

$$R=\sum_{i=1}^{n}R_i \tag{29}$$

$$P=\sum_{i=1}^{n}P_i \tag{30}$$

### 3.3 参数设置

本文实验在TensorFlow深度学习框架上完成。本文采用未加权平均召回率(Unweighted average recall, UAR)作为模型的主要评价指标,同时计算模型的准确率、未加权平均精准率和F1值。。按照25 ms帧长和10 ms移帧,将CASIA语音库中所有语音划分成200帧等长的语音帧,将EMO-DB语音库中所有语音划分成250帧等长的语音帧。这种划分方式能够保留大部分语音的有效信息,其中长度未能达到指定帧数的语音采取零值补充至指定帧数。学习率设置为$10^{-4}$,衰减率设置为$10^{-6}$,训练迭代次数设置为150。训练集和测试集的比例为4∶1,进行5次实验取平均值以减小实验的偶然性。

### 3.4 实验结果及分析

(1) 信息融合策略

分别在CASIA和EMO-DB数据库上采取4种信息融合方式,结果如表1、2所示。

**表1 CASIA数据库不同融合方式结果**

**Table 1 Results of different fusion methods in CASIA database** %

| 融合策略 | 准确率 | 精确率 | 召回率 | $F_1$ |
|---|---|---|---|---|
| 特征层融合(Concatenate) | 81.25 | 80.55 | 80.16 | 80.35 |
| 特征层融合(Add) | **85.00** | **84.73** | **84.93** | **84.83** |
| 决策层融合(Average) | 79.58 | 81.17 | 79.05 | 80.10 |
| 决策层融合(Maximum) | 63.33 | 69.62 | 62.85 | 66.06 |

**表2 EMO-DB数据库不同融合方式结果**

**Table 2 Results of different fusion methods in EMO-DB database** %

| 融合策略 | 准确率 | 精确率 | 召回率 | $F_1$ |
|---|---|---|---|---|
| 特征层融合(Concatenate) | 85.98 | 86.06 | 82.73 | 84.36 |
| 特征层融合(Add) | **86.92** | **87.73** | **86.83** | **87.28** |
| 决策层融合(Average) | 79.44 | 77.61 | 74.79 | 76.17 |
| 决策层融合(Maximum) | 71.03 | 75.09 | 68.34 | 71.56 |

从两个数据库语音情感识别结果可看出模型在特征层融合比在决策层融合效果更好,说明先将信息进行整合后再进行分类是更有效的策略。在特征层融合中,采取特征向量并行方式优于拼接方式。在决策层融合中,取最大得分方式出现过拟合,在验证集中表现较差。

(2) RSBU-CW数量设置

深度残差收缩网络可由多个RSBU-CW堆叠,本文对RSBU-CW数量进行了对比。以特征层融合特征向量并行方式比较不同RSBU-CW对整个模型的影响。

从表3的对比可看出,RSBU-CW数量为1时模型就达到最佳了,更多的追求RSBU-CW数量并不能带来召回率的提升。原因是输入深度残差收缩网络的特征量并不大,堆叠RSBU-CW数量反而会造成训练过拟合。

**表3 不同RSBU-CW数量模型未加权平均召回率**

**Table 3 UAR under different numbers of RSBU-CW** %

| RSBU-CW数量 | CASIA | EMO-DB |
|---|---|---|
| 1 | **84.93** | **86.83** |
| 2 | 80.07 | 83.18 |
| 3 | 78.45 | 80.39 |

(3) 训练过程

迭代次数设置为150次,并设置在测试集上准确率40次未提升则提前结束训练,以减小过拟合影响。准确率—迭代次数变化曲线如图3所示。

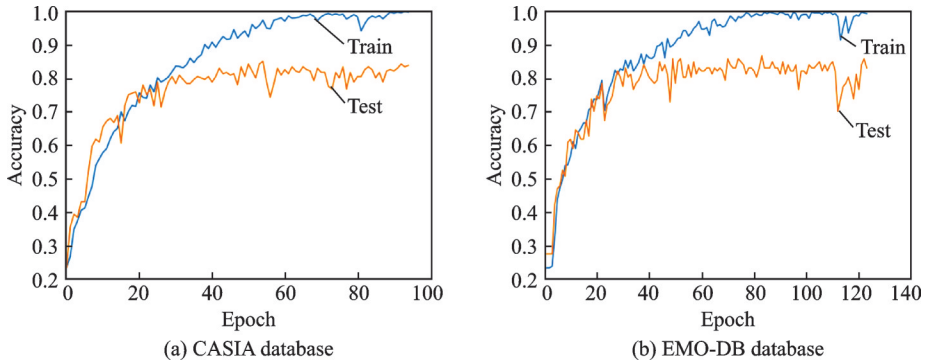



图3 准确率随迭代次数变化曲线

Fig.3 Curve of accuracy with the number of iterations

图3中显示两个数据库都未能完整训练预定义的150次迭代,因为本次试验仅使用了CASIA数据集中1 200句语料,而EMO-DB数据集所含语料也较少,样本较小容易造成过拟合,使用完整CASIA数据集可进一步提高模型的识别精度。同时,当模型训练达到40次左右时模型开始收敛,可证明本文所提出方法的有效性。

(4) 混淆矩阵

CASIA数据库和EMO-DB数据库的混淆矩阵由图4所示。

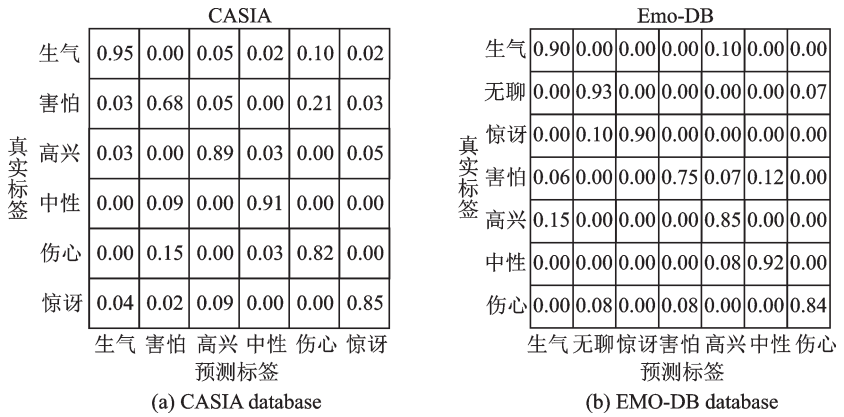



图4 混淆矩阵

Fig.4 Confusion matrix

通过混淆矩阵可看出所提出模型对情感标签“中性”的召回率较高,对情感标签“害怕”和“伤心”的召回率较低,这两种情感在两个数据集上都有一定比例相互混淆,说明提取的特征对于这两种情感的区分度有限,可针对这两种情感的区分进一步深入研究。而另外几种情感都具有较高的召回率,说明所提出模型能够有效区分语音情感。

(5)模型对比

本文以引入注意力机制的CNN-BLSTM作为基线模型。基线模型均采用特征层融合特征向量并行方式,在CASIA和EMO-DB数据库上基线模型与所提出模型的对比如分别表4、5所示。

**表4 CAISA数据库中基线模型与DRSN-CNN-ABLSTM对比**

**Table 4 Comparison of baseline model and DRSN-CNN-ABLSTM in CAISA database** %

| 模型 | 准确率 | 精确率 | 召回率 | $F_1$ |
|---|---|---|---|---|
| 基线模型 | 80.42 | 80.95 | 79.60 | 80.27 |
| DRSN-MF | **85.00** | **84.73** | **84.93** | **84.83** |

**表5 EMO-DB数据库中基线模型与DRSN-CNN-ABLSTM对比**

**Table 5 Comparison of baseline model and DRSN-CNN-ABLSTM in EMO-DB database** %

| 模型 | 准确率 | 精确率 | 召回率 | $F_1$ |
|---|---|---|---|---|
| 基线模型 | 81.31 | 82.64 | 80.59 | 81.60 |
| DRSN-MF | **86.92** | **87.73** | **86.83** | **87.28** |

如表4、5所示,所提出模型在准确率,精确率,召回率和$F_1$值4个指标上均优于基线模型。为进一步验证本文引入深度残差收缩网络以及采取特征层融合特征向量并行方式的有效性,和其他论文中主要使用CNN、LSTM及注意力机制建立模型的结果进行对比,如表6所示。结果表明,深度残差收缩网络可以有效提高基于CNN和LSTM深度学习的语音情感识别精度,同时采取特征层融合特征向量并行方式能够有效提升模型识别的效果。

**表6 与其他模型UAR结果对比**

**Table 6 UAR comparison with other models** %

| 模型 | CASIA | EMO-DB |
|---|---|---|
| DCNN_LSTM[29] | — | 80.60 |
| 3D ACRNN[19] | — | 82.28 |
| DCNN_DTPM[1] | — | 80.39 |
| CNN_LSTM[30] | 79.70 | 80.10 |
| LSTM-self-Attention[31] | 83.21 | 81.18 |
| DRSN-MF(本文) | **84.73** | **86.83** |

## 4 结束语

本文以采用注意力机制的CNN-BLSTM为基线模型,引入深度残差收缩网络,设置卷积通道的权重进行语音情感识别,并考虑了特征融合机制,得到在本模型中使用特征层融合特征向量并行方式能够更有效地训练模型的结论。在CASIA数据库和EMO-DB数据库上进行实验,与本文基线模型和其他论文中以CNN、LSTM及注意力机制建立模型进行对比分析,验证了所提出方法的有效性。

**参考文献:**

[1] ZHANG S, ZHANG S, HUANG T, et al. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching[J].IEEE Transactions on Multimedia,2018,20(6): 1576-1590.

[2] 张会云, 黄鹤鸣, 李伟, 等.语音情感识别研究综述[J].计算机仿真, 2021, 38(8): 7-17.
ZHANG Huiyun, HUANG Heming, LI Wei, et al. A survey of speech emotion recognition[J]. Computer Simulation, 2021, 38(8): 7-17.

[3] 冯亚琴, 沈凌洁, 胡婷婷, 等.利用语音与文本特征融合改善语音情感识别[J].数据采集与处理, 2019, 34(4): 625-631.
FENG Yaqin, SHEN Lingjie, HU Tingting, et al. Using speech and text features fusion to improve speech emotion recognition [J]. Journal of Data Acquisition and Processing, 2019, 34(4): 625-631.

[4] LAUKKA P, NEIBERG D, FORSELL M, et al. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation[J]. Computer Speech & Language, 2011, 25(1): 84-104.

[5] BOU-GHAZALE S E, HANSEN J. A comparative study of traditional and newly proposed features for recognition of speech under stress[J]. Speech & Audio Processing IEEE Transactions on, 2000, 8(4): 429-442.

[6] HAN Z, JIAN W. Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine[C]// Proceedings of 2017 Chinese Automation Congress (CAC). Jinan: IEEE, 2017: 2513-2516.

[7] HSIAO P W, CHEN C P . Effective attention mechanism in dynamic models for speech emotion recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 2526-2530.

[8] 胡德生, 张雪英, 张静, 等.基于主辅网络特征融合的语音情感识别[J].太原理工大学学报,2021,52(5): 769-774.
HU Desheng, ZHANG Xueying, ZHANG Jing, et al. Speech emotion recognition based on primary and secondary network feature fusion[J]. Journal of Taiyuan University of Technology, 2021, 52(5): 769-774.

[9] 韩文静, 李海峰, 阮华斌, 等.语音情感识别研究进展综述[J].软件学报, 2014, 25(1): 37-50.
HAN Wenjing, LI Haifeng, RUAN Huabin, et al. Review on speech emotion recognition[J]. Journal of Software, 2014, 25 (1): 37-50.

[10] SIDDIQI M H. An improved gaussian mixture hidden conditional random fields model for audio-based emotions classification [J]. Egyptian Informatics Journal, 2020, 22(1): 45-51.

[11] GERAVANCHIZADEH M, FOROUHANDEH E, BASHIRPOUR M. Feature compensation based on the normalization of vocal tract length for the improvement of emotion-affected speech recognition[J]. EURASIP Journal on Audio Speech and Music Processing, 2021, 2021(1): 2-19.

[12] KE X, ZHU Y, WEN L, et al. Speech emotion recognition based on SVM and ANN[J]. International Journal of Machine Learning and Computing, 2018, 8(3): 198-202.

[13] 潘国壮.基于实时陆空通话情感识别的管制员疲劳状态快速监测研究[D].南京:南京航空航天大学,2020.
PAN Guozhuang. Research on rapid monitoring of controller fatigue based on real-time land air communication emotion recognition[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2020.

[14] FARHOUDI Z, SETAYESHI S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition[J]. Speech Communication, 2021, 127: 92-103.

[15] DANGOL R, ALASDOON A, PRASAD P, et al. Speech emotion recognition using convolutional neural network and long-short term memory[J]. Multimedia Tools and Applications, 2020, 79(43/44): 32917-32934.

[16] 胡婷婷, 冯亚琴, 沈凌洁, 等.基于注意力机制的LSTM语音情感主要特征选择[J].声学技术, 2019, 38(4): 414-421.
HU Tingting, FENG Yaqin, SHEN Lingjie, et al. Main feature selection of LSTM speech emotion based on attention mechanism[J]. Technical Acoustics, 2019, 38(4): 414-421.

[17] 邦锦阳, 孙蒙, 张雄伟, 等.融合卷积网络与残差长短时记忆网络的轻量级骨导语音盲增强[J].数据采集与处理, 2021, 36 (5): 921-931.
BANG Jinyang, SUN Meng, ZHANG Xiongwei, et al. Lightweight model for bone-conducted speech enhancement based on convolution network and residual long short-time memory network[J]. Journal of Data Acquisition and Processing, 2021, 36(5): 921-931.

[18] 姜特, 陈志刚, 万永菁.基于注意力机制的多任务3D CNN-BLSTM情感语音识别[J].华东理工大学学报(自然科学版), 2021, 49(1): 1-8.
JIANG Te, CHEN Zhigang, WAN Yongjing. Multi-task 3D CNN-BLSTM emotional speech recognition based on attention mechanism[J]. Journal of East China University of Science and Technology, 2021, 49(1): 1-8.

[19] CHEN M, HE X, JING Y, 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.

[20] ZHAO J, MAO X, CHEN L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical Signal Processing and Control, 2019, 47: 312-323.

[21] 徐华南, 周晓彦, 姜万, 等.基于3D和1D多特征融合的语音情感识别算法[J].声学技术, 2021, 40(4): 496-502.
XU Huanan, ZHOU Xiaoyan, JIANG Wan, et al. Speech emotion recognition algorithm based on 3D and 1D multi-feature fusion[J]. Technical Acoustics, 2021, 40(4): 496-502.

[22] ZHAO M, ZHONG S, FU X, et al. Deep residual shrinkage networks for fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2020, 16(7): 4681-4690.

[23] BANSE R, SCHERER K R. Acoustic profiles in vocal emotion expression[J]. Journal of Personality & Social Psychology, 1996, 70(3): 614-636.

[24] SWAIN M, SAHOO S, ROUTRAY A, et al. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition[J]. International Journal of Speech Technology, 2015, 18(3): 387-393.

[25] GUNES H, PICCARDI M. Affect recognition from face and body: Early fusion vs. late fusion[C]//Proceedings of Systems, Man and Cybernetics, 2005 IEEE International Conference on. Waikoloa, HI, USA: IEEE, 2005: 3437-3443.

[26] WANG K, AN N, LI B N, et al. Speech emotion recognition using fourier parameters[J]. IEEE Transactions on Affective Computing, 2017, 6(1): 69-75.

[27] ZAO L, CAVALCANTE D, RUI C. Time-frequency feature and AMS-GMM mask for acoustic emotion classification[J]. IEEE Signal Processing Letters, 2014, 21(5): 620-624.

[28] GUPTA A, ANJUM, GUPTA S, et al. InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray[J]. Applied Soft Computing, 2020, 99: 1-13.

[29] KIM J, SAUROUS R A. Emotion recognition from human speech using temporal information and deep learning[C]//Proceedings of Interspeech 2018. Hyderabad, India: ISCA, 2018: 937-940.

[30] 卢官明, 袁亮, 杨文娟, 等. 基于长短期记忆和卷积神经网络的语音情感识别[J]. 南京邮电大学学报:自然科学版, 2018, 38(5): 63-69.
LU Guanming, YUAN Liang, YANG Wenjuan, et al. Speech emotion recognition based on short-term memory and convolutional neural network[J]. Journal of Nanjing University of Posts and Telecommunications, 2018, 38(5): 63-69.

[31] 陈巧红, 于泽源, 孙麒, 等. 基于注意力机制与LSTM的语音情绪识别[J]. 浙江理工大学学报:自然科学版, 2020, 43(6): 815-822.
CHEN Qiaohong, YU Zeyuan, SUN Lin, et al. Speech emotion recognition based on attention mechanism and LSTM[J]. Journal of Zhejiang University of Technology: Natural Science Edition, 2020, 43(6): 815-822.

**作者简介:**

**李瑞航**(1997-),男,硕士研究生,研究方向:语音情感识别、驾驶舱人机交互,E-mail: liruihang@nuaa.edu.cn。

**吴红兰**(1969-),**通信作者**,女,高级工程师,研究方向:交通信息工程及控制、飞机适航审定与验证技术,E-mail: wuhonglan@126.com。

**孙有朝**(1965-),男,博士,教授,研究方向:航空器适航性设计与验证技术、飞行安全与人为因素。

**吴华聪**(2002-),男,本科生,研究方向:飞行器适航技术。

(编辑:夏道家)