



研究与开发

一种基于 DN-ResNet11 的语音情感识别算法

应娜, 邹雨鉴, 杨雪滢, 孙文胜, 叶学义, 蒋银河
(杭州电子科技大学通信工程学院, 浙江 杭州 310018)

摘要: 为解决网络训练复杂度高的问题并改进语音情感特征提取, 提出了基于双嵌套残差网络 (DN-ResNet11) 与通道注意残差网络 (CRNet) 的双支路特征提取模型。首先, 设计了低复杂度的 DN-ResNet11 以高效提取语谱图的融合情感特征, 提升情感识别率; 然后, 结合多尺度引导滤波和局部二值模式 (local binary pattern, LBP) 算法对语谱图进行细节增强; 最后, 融合两组特征进行情感分类, 形成双支路加权融合模型 (weighted fusion model based on dual nested residual and channel residual network, WFDN_CRNet), 进一步提升情感表征能力。在 CASIA、EMO-DB、IEMOCAP 等语音情感数据集上情感识别率分别达到 94.58%、85.59%、65.72%, 所提方法在情感识别率优于 ResNet18 等基准方法的同时, 显著降低了计算成本, 验证了模型的有效性。

关键词: 情感识别; 双嵌套残差网络; 细节增强; 加权融合

中图分类号: TP18

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2025042

A speech emotion recognition algorithm based on DN-ResNet11

YING Na, ZOU Yujian, YANG Xueying, SUN Wensheng, YE Xueyi, JIANG Yinhe
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: To address the high complexity of network training and improve speech emotion feature extraction, a dual-branch feature extraction model based on DN-ResNet11 and a channel attention residual network (CRNet) was proposed. Firstly, the low-complexity DN-ResNet11 was designed to efficiently extract fused emotional features from spectrograms, enhancing emotion recognition accuracy. Secondly, multiscale guided filtering and the local binary pattern (LBP) algorithm were incorporated to enhance spectrogram details. Finally, the two sets of features were fused for emotion classification, forming a dual-branch weighted fusion model (weighted fusion model based on dual nested residual and channel residual network, WFDN_CRNet), further enhancing emotional representation ability. Experiments on the CASIA, EMO-DB, and IEMOCAP speech emotion datasets show emotion recognition rates of 94.58%, 85.59%, and 65.72%, respectively. The proposed method not only achieves superior emotion recognition rates compared to baseline models such as ResNet18, but also significantly reduces computational cost, demonstrating the

收稿日期: 2024-09-27; 修回日期: 2024-12-05

通信作者: 应娜, yingna@hdu.edu.cn

基金项目: 浙江省科技计划项目 (No.LGF21F010003); 浙江省“尖兵”“领雁”项目 (No.2022C03065)

Foundation Items: The Zhejiang Provincial Natural Science Foundation of China (No.LGF21F010003), The Key Research and Development Program in Zhejiang Province (No.2022C03065)



model's effectiveness.

Key words: emotion recognition, dual nested residual network, detail enhancement, weighted fusion

0 引言

语音是人们表达情感的一种媒介,也是人类日常生活中最方便和最有效的一种交流方式。在人机交互普及的时代,“如何让机器拥有识别人类情感波动的能力”的研究逐渐展开,语音情感识别技术开始进入人们的视野,通过对说话人情绪状态的感知,机器可以有效地调整交流方式,使得人机交流更加自然和有效。因此,语音情感识别具有重要的研究意义。

传统的语音特征主要包括短时对数频率功率系数(log frequency power coefficients, LFPC)特征、梅尔倒谱系数(Mel-frequency cepstral coefficients, MFCC)特征^[1]、调制谱特征(modulation spectral feature, MSF)^[2]等,早期主要采用隐马尔可夫模型(hidden Markov model, HMM)^[3]、支持向量机(support vector machine, SVM)^[4]等算法作为分类器得到情感识别结果。相比传统语音特征,语谱图能够将语音中包含的谐波、共振峰、能量等信息通过二维图像展现出来,是用来提取情感特征最有效的方式之一。

随着深度学习的发展,Yenigalla等^[5]提出将语音序列与传统语谱图分别送入两个卷积神经网络(convolutional neural network, CNN)提取特征,该方法使得识别率得到4.0%以上的提升;Li等^[6]提出将深度残差网络(deep residual network, ResNet)作为特征提取网络、SVM作为分类器的模型,达到90.0%的识别率;Wang等^[7]提出双序列的长短期记忆(long short-term memory, LSTM)模型同时处理同一语音、不同分辨率的两幅语谱图得到一组特征,同时使用传统LSTM网络提取语音的梅尔倒谱系数,最后进行分类,但该模型在提升识别率的同时,增加了网络复杂

度,不利于后续模型的改进;Zhang等^[8]将梅尔语谱图输入三层卷积以及胶囊网络提取空间特征,然后采用动态路由算法识别情感,使得识别率得到2.0%以上的提升;金俊林等^[9]通过在卷积层之间进行集成以提取更加密切的语谱图相关特征,使得识别率得到1.31%以上的提升。

近年来大型预训练模型如Transformer^[10]、wav2vec2.0^[11]、HuBERT^[12]等开始被推广应用于语音情感识别任务当中,通常会考虑语义信息,但需要大量的显存和算力支持,适合于数据中心和服务器环境,但难以在资源有限的边缘设备上部署,同时大模型通常需要大量数据、较长的训练时间、较高的复杂度。针对改善模型复杂度和特征提取问题,目前主要的方法是模型压缩和轻量化,例如,剪枝^[13]和量化^[14]技术可以去除冗余参数,深度可分离卷积^[15]可以减少计算量。另外,对于高维特征,主要使用特征选择方法或降维技术来减少特征数量,从而降低计算复杂度,常见的技术包括主成分分析、线性判别分析和自动编码器等。此外,通常还进行数据增强,通过如加噪声、变换、拉伸等获得更大的数据量。此外,语义信息会受不同语言或文化在语言表达上差异的影响,而非语言特征直接使语音情感识别更具有鲁棒性。

根据上述研究现状,为降低模型复杂度、同时提升语音情感识别率以及算法的训练效率,本文重点研究语谱图以及深度学习网络模型,主要从改进模型降低复杂度、特征提取和数据增强等方面入手,提出一种不考虑语义信息、低复杂度、增强情感特征细节、高效提取融合特征的双支路情感识别模型,在3个语音情感数据集CASIA、EMO-DB、IEMOCAP上进行实验。

本文创新如下:(1)根据残差网络结构设计

了 DN-ResNet11 模型提取语谱图的融合情感特征,降低网络复杂度的同时提升模型的识别性能;(2)采用多尺度引导滤波对语谱图进行细节增强,得到加权语谱图,并将其作为 DN-ResNet11 模型的输入,提取语谱图增强情感特征;(3)引入局部二值模式(local binary pattern, LBP)算法辅助增强加权语谱图纹理方向的情感特征,并设计了 CRNet 模型提取传统语谱图的 LBP 特征;(4)将增强情感特征与 LBP 特征加权融合后得到新的特征图,再进行分类。实验结果显示,本文算法在降低网络复杂度的同时,情感识别率有稳定提升。

1 双支路情感识别模型

1.1 基于 DN-ResNet11 的情感特征提取模型

残差结构^[16]如图 1 所示,通过输入特征 x 与提取特征 $F(x)$ 的累加结合,使得输出特征 $Y(x)$ 既包含了新的情感信息又保留了原始情感信息;同时,由于残差连接打破了网络的对称性^[17],在反向传播过程中,相比非残差连接的网络结构,残差结构对网络参数的调整更加敏感,有利于网络的参数更新与性能改善^[18]。

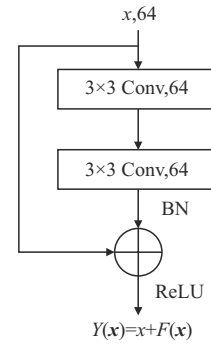


图1 残差结构

受残差结构对特征提取作用的启发,为充分利用残差连接对特征融合的有效性,考虑将残差结构做进一步改进。首先,根据残差连接的结构有效提取融合情感特征的思路,将残差结构作为小尺度残差模块,并在该结构基础上嵌套一层大尺度残差连接,提出的双嵌套残差结构如图 2 所示。

在该结构中,假设输入特征为 x ,经过两层 3×3 的卷积计算得到特征 $y_1 = G_1(x)$,再将 y_1 同时输入两层 3×3 卷积计算后累加的小尺度残差结构与一层 1×1 卷积,分别得到特征 $G_2(x)$ 与特征 $S_1(y_1)$,二者累加得到融合特征 $y_2 = S_1(y_1) + G_2(y_1)$;最后,通过两层 3×3 卷积计算得到特征 $G_3(y_2)$;

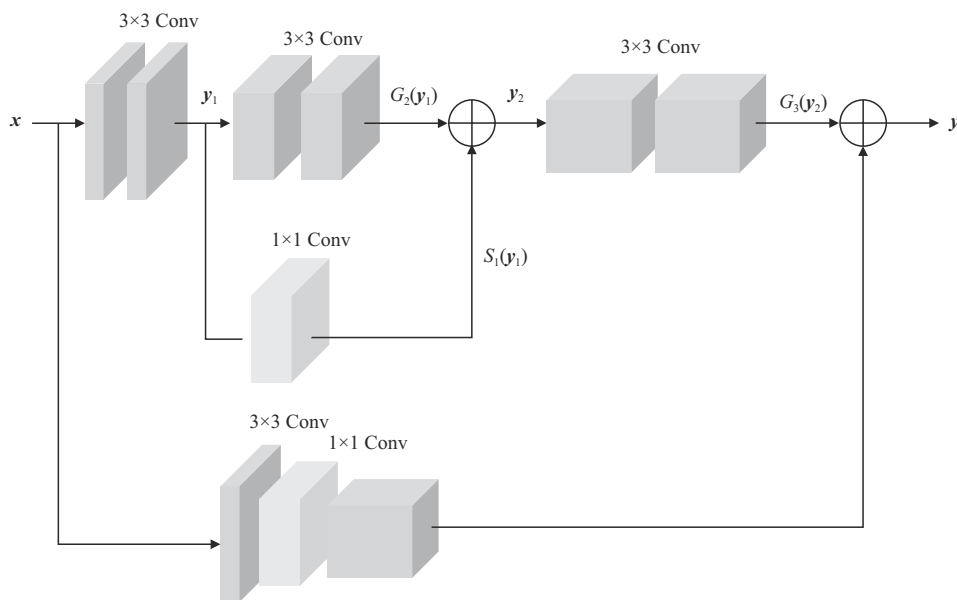


图2 双嵌套残差结构



同时, 输入特征 x 在经过大尺度残差中 3×3 卷积、 1×1 卷积与 3×3 卷积的三层卷积计算后得到 $S_2(x)$, 并最终与特征 $G_3(y_2)$ 累加融合, 得到双嵌套残差结构的输出特征 $y = S_2(x) + G_3(y_2)$ 。相比 ResNet18 网络, 既节省参数又提高效率。

基于 DN-ResNet11 的情感特征提取模型如图 3 所示。双嵌套残差结构通过将最大池化后的特征图与平均池化前的特征图叠加生成新特征图。每次卷积后进行批标准化 (BN) 和 ReLU^[19]激活。DN-ResNet11 的前两层采用与常规 ResNet 不同的步长设置: 7×7 卷积步长为 2 以快速捕捉全局信息, 3×3 最大池化步长为 1 以减少有用特征丢失。网络最后通过自适应平均池化和全连接层进行情感分类, 输入为细节增强后的加权语谱图。

1.2 基于 GF 的情感特征增强算法

引导滤波 (guided filter, GF)^[20]最早在 2010 年的计算机视觉会议上被提出, 它能够实现对图像的边缘保持与去雾, 且该算法与同样有较好边缘保持效果的双边滤波^[21] (bilateral filter, BF) 算法进行比较。GF 算法在假设输出图像 q 与引导图像 I 满足线性关系 $q = aI + b$ (其中, a 和 b 是在局部窗口内拟合这个线性关系的两个系数) 的前提

下, 将输入图像 p 中非边缘且不平整的区域视为噪声 n , 则有 $q = p - n$ 。设置 $I = p$ 。将图像 p 分成若干边长为 $2r$ 的局部窗口, 记为 w_k , 生成局部滤波图像后顺序拼接得到输出图像 q 。采用引导滤波算法对滤波后的图像 q 在像素点 i 的像素值, 如式 (1) 所示:

$$q_i = \frac{1}{|w|} \sum_{k, i \in w_k} (a_k^* p_i + b_k^*) = \bar{a}_i^* p_i + \bar{b}_i^* \quad (1)$$

其中, $|w|$ 为滤波窗口 w_k 内的像素点数量, p_i 为输入图像 p 在像素点 i 的像素值, (a_k^*, b_k^*) 表示滤波后求得的像素点 i 处的线性系数, 其值为集合 $\{(a_k^*, b_k^*), k \in w_k\}$ 中元素 (a_k^*, b_k^*) 的均值。 (a_k^*, b_k^*) 的取值如式 (2):

$$\begin{cases} a_k^* = \frac{\sigma_k^2}{\sigma_k^2 + \varepsilon} \\ b_k^* = \frac{\varepsilon \mu_k}{\sigma_k^2 + \varepsilon} \end{cases} \quad (2)$$

其中, μ_k 和 σ_k^2 为滤波窗口 w_k 内的均值和方差。

由于引导滤波在窗口 w_k 内进行, w_k 的大小对图像在窗口内的均值、方差等参数有直接影响, 从而影响滤波效果; 此外, 引入的正则化系数 ε 决定滤波的阈值, 对 (a_k^*, b_k^*) 的取值有不可忽略的影响。因此, 可以通过调整 ε 与窗口半径 r 来

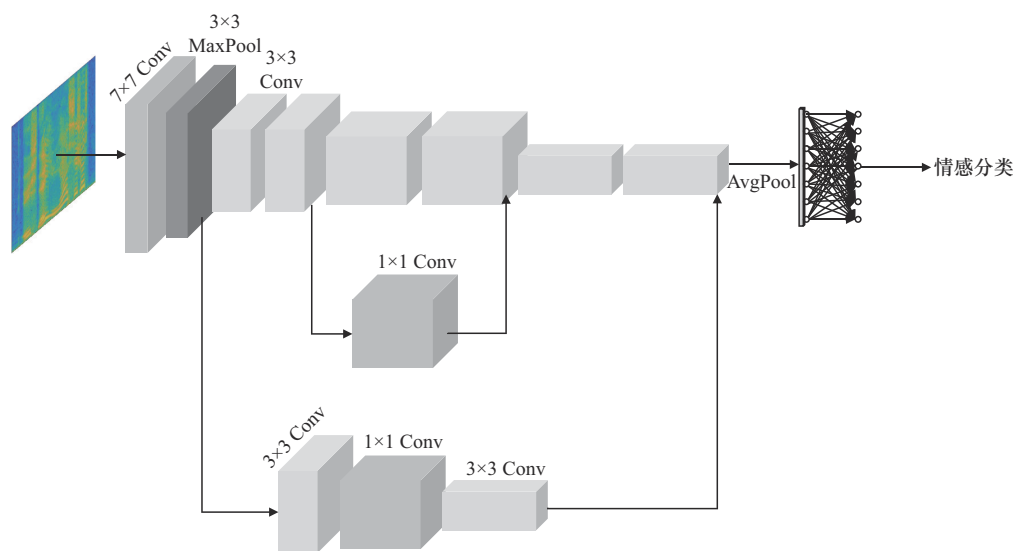


图3 基于 DN-ResNet11 的情感特征提取模型

改变滤波强度。

具体算法流程如下。

(1) 将输入语谱图记为 $\mathbf{X}$ ，选取小尺度引导滤波与大尺度引导滤波两组滤波参数：($r_{\min}, \varepsilon_{\min}$) 与 ($r_{\max}, \varepsilon_{\max}$)，分别对语谱图进行小尺度 GF 与大尺度 GF，得到不同尺度的两组滤波图像，分别记为 $\mathbf{Y}_{\min}$ 与 $\mathbf{Y}_{\max}$ 。

(2) 通过对 $\mathbf{Y}_{\min}$ 与 $\mathbf{Y}_{\max}$ 像素点的差值来描述语谱图细节特征，其距离函数如下表示：

$$\begin{aligned} |E_{q_{\min}} - E_{q_{\max}}| &= \left| \frac{1}{|w_1|} \sum_{k_{\min}, i \in w_{k_{\min}}} (a_{k_{\min}}^* X_i + b_{k_{\min}}^*) - \right. \\ &\quad \left. \frac{1}{|w_2|} \sum_{k_{\max}, i \in w_{k_{\max}}} (a_{k_{\max}}^* X_i + b_{k_{\max}}^*) \right| = \\ &\quad |\bar{a}_{i_{\min}} - \bar{a}_{i_{\max}}| X_i + |\bar{b}_{i_{\min}} - \bar{b}_{i_{\max}}| \end{aligned} \quad (3)$$

其中， X_i 表示 $\mathbf{X}$ 在像素点 i 的像素值； $w_{k_{\min}}$ 、 $w_{k_{\max}}$ 分别表示窗口半径为 $r_{\min}$ 与 $r_{\max}$ 的局部窗口； $|w_1|$ 、 $|w_2|$ 分别表示在 $w_{k_{\min}}$ 、 $w_{k_{\max}}$ 内的像素点数量；($a_{k_{\min}}^*, b_{k_{\min}}^*$) 与 ($a_{k_{\max}}^*, b_{k_{\max}}^*$) 分别表示在 $w_{k_{\min}}$ 与 $w_{k_{\max}}$ 内求得的局部线性系数；($\bar{a}_{i_{\min}}^*, \bar{b}_{i_{\min}}^*$) 与 ($\bar{a}_{i_{\max}}^*, \bar{b}_{i_{\max}}^*$) 分别表示在 $w_{k_{\min}}$ 与 $w_{k_{\max}}$ 内 ($a_{k_{\min}}^*, b_{k_{\min}}^*$) 与 ($a_{k_{\max}}^*, b_{k_{\max}}^*$) 求得的所有的均值， $|\bar{a}_{i_{\min}} - \bar{a}_{i_{\max}}|$ 和 $|\bar{b}_{i_{\min}} - \bar{b}_{i_{\max}}|$ 是距离函数在像素点 i 处的线性系数。

(3) 引入增强系数 λ 对细节特征进行加权，小尺度引导滤波保留了语谱图更多的非边缘区域的细节，故选其作为基底图像，与加权细节特征累加结合，得到特征增强的加权语谱图，记为 $\mathbf{Y}$ ， Y_i 表示 $\mathbf{Y}$ 在像素点 i 的局部语谱图， $\bar{a}_{i_{\min}}^*$ 、 $\bar{b}_{i_{\min}}^*$ 是 Y_i 关于 X_i 的线性系数：

$$Y_i = (\bar{a}_{i_{\min}}^* X_i + \bar{b}_{i_{\min}}^*) + \lambda |E_{q_{\min}} - E_{q_{\max}}| \quad (4)$$

1.3 基于 LBP 纹理特征提取的 CRNet 模型

为增强加权语谱图纹理方向的情感信息，引入局部二值模式^[22] (local binary pattern, LBP) 算法，

以其旋转不变性与灰度不变性等优点处理原始语谱图，并设计了 CRNet 提取 LBP 图像的纹理特征。

LBP 算法处理前后的情感为生气的语谱图如图 4 所示。LBP 图像充分描述了语谱图的纹理信息，提取 LBP 纹理特征能够辅助增强第 1.2 节提取加权语谱图的增强情感特征在纹理方向上的情感信息，进而有助于提升模型的识别性能。

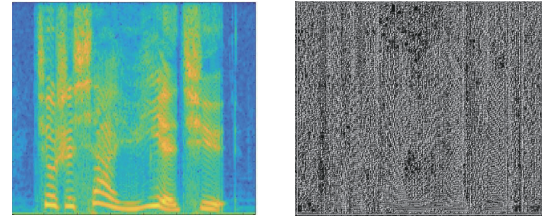


图 4 LBP 算法处理前后的情感为生气的语谱图

为有效提取 LBP 图像中的纹理特征，设计了 CRNet 模型作为 LBP 纹理特征提取模型。在该网络中，为保证网络简单有效，引入高效通道注意力 (efficient channel attention, ECA) 机制，模型能够有效进行跨通道交互等^[23]。

CRNet 模型如图 5 所示，在通道注意残差网络结构中，输入 $\mathbf{X}$ 表示 LBP 图像，经过 7×7 卷积与 3×3 最大池化后得到特征 $\mathbf{x}$ ，通道注意残差网络将 $\mathbf{x}$ 送入 ECA 模块实现数据间的跨通道交互，得到 ECA 加权后的特征 $S(\mathbf{x})$ 。然后，对残差结构进行微调，采用 3 组微调的残差结构对 $S(\mathbf{x})$ 进一步提取纹理特征，经过 3 次累加融合后得到的融合特征经过全局平均池化 (global average pooling, GAP)，得到最终的输出特征 $\mathbf{F}_2$ 。

1.4 基于特征加权融合的 WFDN_CRNet 模型

为增强语音情感的差异性，本节将第 1.2 节 DN-ResNet11 提取的增强情感特征与第 1.3 节 CRNet 提取的 LBP 纹理特征相结合，提出融合两种不同特征的加权融合算法。其中，GF 加权语谱图通过 DN-ResNet11 得到的增强情感特征 $\mathbf{F}_1$ 与 CRNet 所提取的 LBP 纹理特征 $\mathbf{F}_2$ 分别进行不同尺

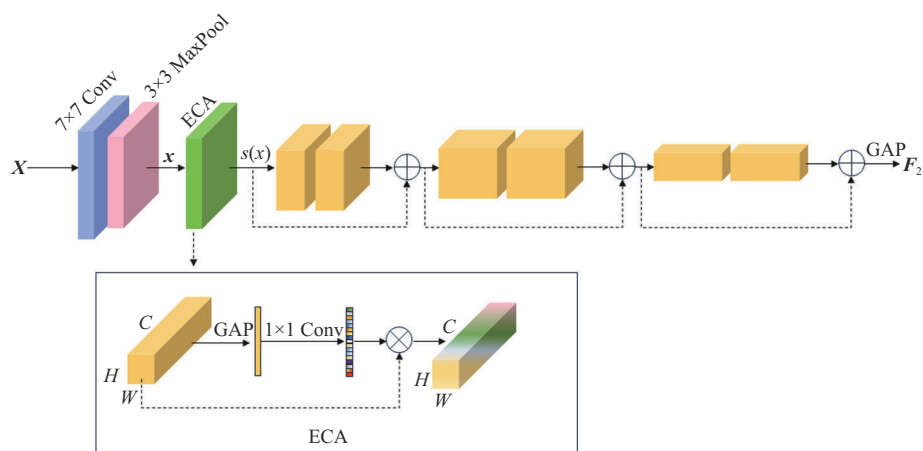


图5 CRNet模型

度的乘性加权后进行累加融合操作，得到加权融合后的情感特征。由此引入增强情感特征加权参数 α 与LBP纹理特征加权参数 β ，加权融合计算式如下所示：

$$F = \alpha F_1 + \beta F_2 \quad (5)$$

特征加权融合算法能够有效融合增强情感特征与LBP纹理特征，使得部分有效情感信息得以加强。最后，将加权融合特征送入全连接层进行分类，得到的WFDN_CRNet模型如图6所示。

2 实验结果与分析

为验证DN-ResNet11、GF情感特征增强算法及WFDN_CRNet模型在语音情感识别中的有效性，基于CASIA、EMO-DB和IEMOCAP数据集设计3组实验。

实验1：比较DN-ResNet11与其他网络的语音谱图识别效果与复杂度，证明其在降低复杂度的

同时提取更具表征力的情感特征。

实验2：验证情感增强算法对加权语谱图细节特征的提升及其对网络识别性能的增强。

实验3：调参确定加权系数，比较基线与改进模型，验证纹理方向上情感特征增强的有效性。

2.1 实验数据集及参数设置

为验证本文模型提取情感特征的有效性，基于CASIA、EMO-DB与IEMOCAP 3个数据集进行验证。

(1) CASIA语音情感数据集是由4位演员（2男2女）录制的平衡数据集，包括6种情感：生气（angry）、害怕（fear）、高兴（happy）、中性（neutral）、伤心（sadness）、惊奇（surprise）。该数据集的语音样本采用16 kHz采样，16 bit量化，有50句相同文本，共1 200条音频。每种情感包含200条音频。

(2) EMO-DB为柏林工业大学由10位演员（5男5女）对10个语句（5长5短）录制的天平

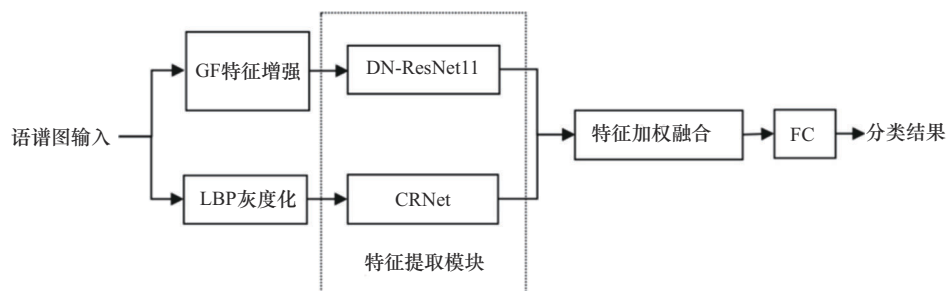


图6 WFDN_CRNet模型

衡数据集, 包括7种情感: 中性 (neutral)、生气 (angry)、恐惧 (fear)、高兴 (happy)、伤心 (sadness)、厌恶 (disgust)、无聊 (boredom)。该数据集的语音样本采用 16 kHz 采样, 16 bit 量化, 共包含 535 条音频。

(3) IEMOCAP 语音数据集是由 10 位演员 (5 男 5 女) 录制的大型语音情感数据集, 平衡不同情感类别的数据后, 包含 4 类: 生气 (angry)、高兴 (happy)、中性 (neutral)、伤心 (sadness)。该数据集语音样本采用 16 kHz 采样, 16 bit 量化, 共包含 5 531 条音频。

实验设置初始学习率为 0.001, 批处理为 16, 迭代次数为 400 次, 设置数据集的 80% 为训练集, 20% 为测试集, 模型输入为语谱图, 设置输入维度为 (3×224×224)。

2.2 实验 1

为验证 DN-ResNet11 的低复杂度和双嵌套残差结构的情感表征效果, 与 ResNet18 和 ResNeXt 进行对比实验, 输入同一语谱图, 输出类别为 6, DN-ResNet11、ResNet18 与 ResNeXt 复杂度对比见表 1, DN-ResNet11 的 FLOPs 为 ResNet18 和 ResNeXt 的 0.36 倍和 0.17 倍, 参数量为 0.12 倍和 0.06 倍, MemR+W 为 0.47 倍和 0.12 倍。结果表明, DN-ResNet11 在时间和空间复杂度上显著降低, 减少了训练时间并提高了效率。

表 1 DN-ResNet11、ResNet18 与 ResNeXt 复杂度对比

对比项	FLOPS	总参数量	MemR+W
ResNet18	1.82×10^9	11 179 590	95.09 MB
ResNeXt	3.81×10^9	22 996 038	383.84 MB
DN-ResNet11	0.66×10^9	1 286 692	44.53 MB

以 CASIA 数据集为例, 3 种网络训练时准确率随着迭代次数变化曲线如图 7 所示, 其中 DN-ResNet11 模型在前 50 次迭代中, 模型识别率增长迅速, 在迭代次数为 100 次左右识别率开始趋于

稳定, 在小范围内波动, 模型进入收敛状态, 训练过程中优化器以阶梯式衰减的方式逐渐减小学习率, 在训练后期, 模型能够在已学到的特征基础上进行微调, 从而保持收敛稳定性。对比 ResNet18 和 ResNeXt, DN-ResNet11 模型收敛速度最快且识别精度最高。

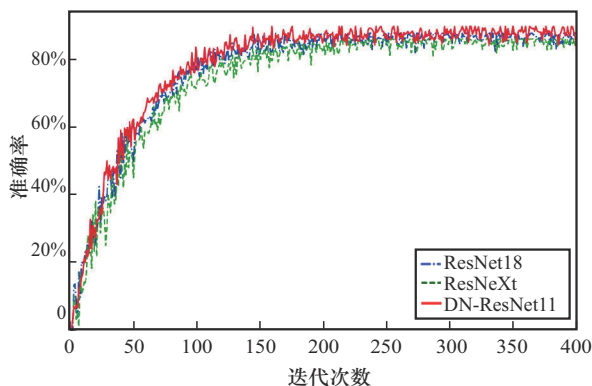


图 7 3 种网络训练时准确率随着迭代次数变化曲线

CASIA 数据集下的情感识别率见表 2, EMO-DB 数据集下的情感识别率见表 3, IEMOCAP 数据集下的情感识别率见表 4。

表 2 CASIA 数据集下的情感识别率

情感类别	ResNet18	ResNeXt	DN-ResNet11
生气	90.00%	92.50%	87.50%
恐惧	92.50%	72.50%	87.50%
高兴	77.50%	87.50%	87.50%
中性	85.00%	85.00%	92.50%
伤心	80.00%	77.50%	82.50%
惊奇	100%	100%	100%
平均识别率	87.50%	85.83%	89.58%

表 3 EMO-DB 数据集下的情感识别率

情感类别	ResNet18	ResNeXt	DN-ResNet11
生气	92.31%	80.77%	88.46%
恐惧	70.59%	88.24%	76.47%
高兴	70.00%	50.00%	80.00%
中性	92.86%	78.57%	92.86%
无聊	66.67%	80.00%	60.00%
厌恶	75.00%	81.25%	87.50%
伤心	76.92%	84.61%	84.15%
平均识别率	79.28%	79.28%	81.98%



表4 IEMOCAP数据集下的情感识别率

情感类别	ResNet18	ResNeXt	DN-ResNet11
生气	74.31%	66.06%	70.64%
高兴	55.06%	44.55%	55.06%
中性	63.30%	43.10%	56.03%
伤心	57.80%	64.22%	69.75%
平均识别率	62.62%	54.48%	62.87%

结合表1至表4可证明, DN-ResNet11在降低网络复杂度的同时提升了情感识别性能, 并具备一定的泛化性。然而, 部分情感如中性、无聊、伤心等识别率较低, 容易被混淆(中性和无聊的声调单调且能量较低, 生气和厌恶具有高能量和快语速)。此外, 不同数据集的表现差异表明数据来源、录制环境等因素影响了模型性能, 如IEMOCAP数据集上的效果有待提升。总体来看, ResNet18的计算复杂度比DN-ResNet11高约2.76倍, 且后者情感识别率略高, 证明了算法的有效性。

2.3 实验2

经多次调参实验, 选取滤波参数为: $((r_{\min}=1, \varepsilon_{\min}=0.005), (r_{\max}=3, \varepsilon_{\max}=0.015))$, 增强系数 $\lambda=3$ 。不同变换后的中性情感语谱图如图8所示, 表示情感为中性的原始语谱图经过不同尺度滤波进行特征增强后加权, 得到最后的加权语谱图。小尺度滤波后的语谱图模糊度较小, 保留了原始语谱图的较多细节, 因此在第2.2节算法介绍中选取其作为基底图像大尺度滤波后的语谱图表现较为平滑; 经过特征增强后的加权语谱图, 图像细节描述更加清晰明显, 有利于情感特征的提取。

将传统语谱图与加权语谱图作为模型输入, 并将BF算法作为对比算法, 不同数据集下DN-ResNet11+BF与DN-ResNet11+GF的混淆矩阵如图9所示, CASIA数据集下不同模型的情感识别率见表5, EMO-DB数据集下不同模型的情感识

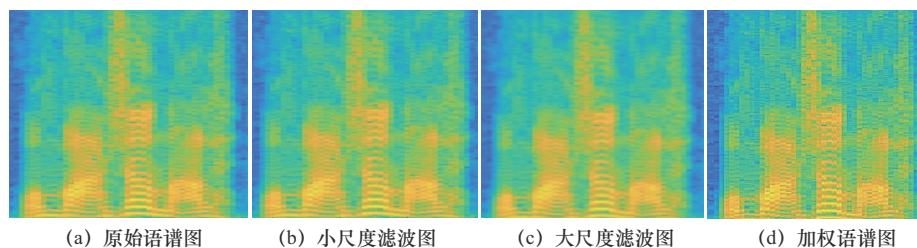
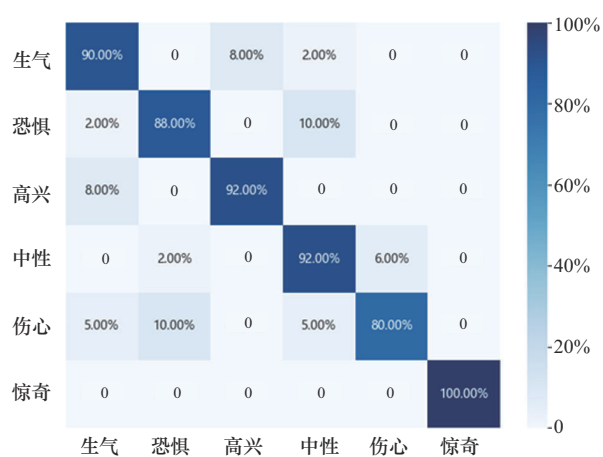


图8 不同变换后的中性情感语谱图

表5 CASIA数据集下不同模型的情感识别率

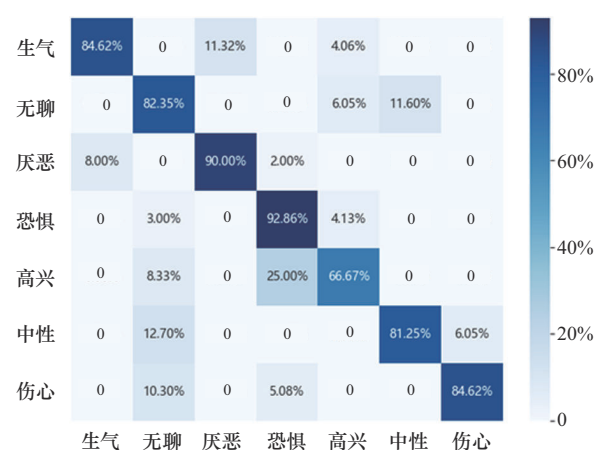
模型	生气	恐惧	高兴	中性	伤心	惊奇	平均识别率
ResNet18	90.00%	92.50%	77.50%	85.00%	80.00%	100%	87.50%
ResNet18+BF	85.00%	85.00%	92.50%	92.50%	82.50%	100%	89.58%
ResNet18+GF	87.50%	82.50%	95.00%	92.50%	82.50%	100%	90.00%
ResNeXt	92.50%	72.50%	87.50%	87.50%	77.50%	100%	85.83%
ResNeXt+BF	72.50%	87.50%	97.00%	95.00%	75.00%	100%	87.83%
ResNeXt+GF	90.00%	82.50%	90.00%	85.00%	82.50%	100%	88.33%
DN-ResNet11	87.50%	87.50%	87.50%	92.50%	82.50%	100%	89.58%
DN-ResNet11+BF	90.00%	88.00%	92.00%	92.00%	80.00%	100%	90.40%
DN-ResNet11+GF	90.00%	90.00%	97.00%	95.00%	85.00%	100%	92.92%



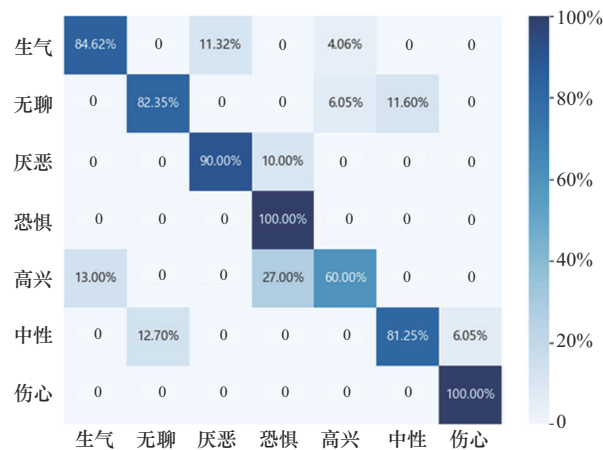
(a) CASIA数据集下的DN-ResNet11+BF



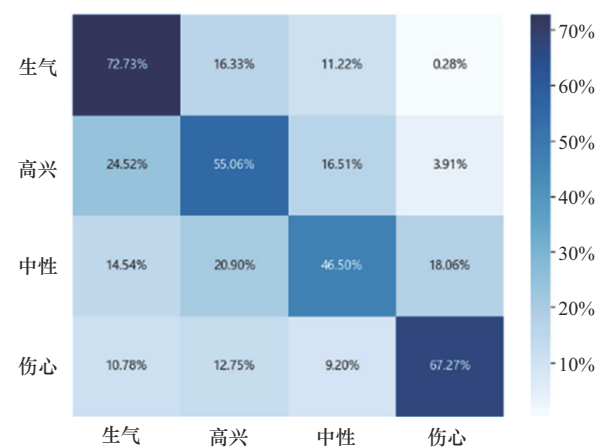
(b) CASIA数据集下的DN-ResNet11+GF



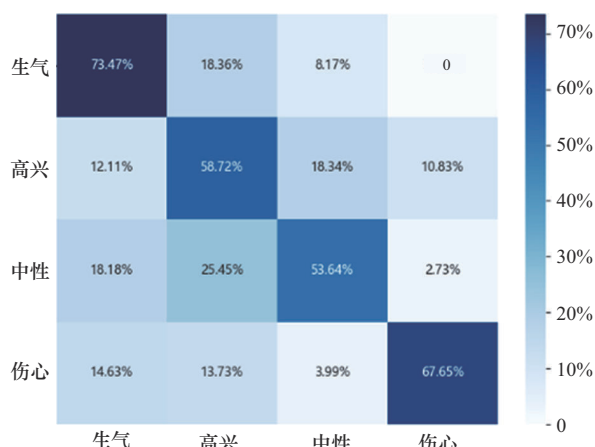
(c) EMO-DB数据集下的DN-ResNet11+BF



(d) EMO-DB数据集下的DN-ResNet11+GF



(e) IEMOCAP数据集下的DN-ResNet11+BF



(f) IEMOCAP数据集下的DN-ResNet11+GF

图9 不同数据集下DN-ResNet11+BF与DN-ResNet11+GF的混淆矩阵

别率见表6, IEMOCAP数据集下不同模型的情感识别率见表7。

对比CASIA数据集和EMO-DB数据集下的

情感识别率, DN-ResNet11+GF模型的平均识别率最高。对比IEMOCAP数据集下的情感识别率, DN-ResNet11+GF模型与其他模型相比, 平



表6 EMO-DB数据集下不同模型的情感识别率

模型	生气	无聊	厌恶	恐惧	高兴	中性	伤心	平均识别率
ResNet18	92.31%	66.67%	75.00%	70.59%	70.00%	92.86%	75.92%	79.28%
ResNet18+BF	88.46%	76.47%	80.00%	92.68%	60.00%	87.50%	84.62%	80.68%
ResNet18+GF	84.62%	82.35%	80.00%	92.68%	60.00%	87.50%	100%	83.78%
ResNeXt	80.77%	80.00%	81.25%	88.24%	50.00%	78.57%	84.61%	79.28%
ResNeXt+BF	84.62%	47.06%	80.00%	71.43%	66.67%	87.50%	84.62%	74.78%
ResNeXt+GF	84.62%	71.43%	60.00%	85.71%	60.00%	93.75%	100%	80.18%
DN-ResNet11	88.64%	60.00%	87.50%	76.47%	80.00%	92.86%	84.15%	81.98%
DN-ResNet11+BF	84.62%	82.35%	90.00%	92.86%	66.67%	81.25%	84.62%	82.88%
DN-ResNet11+GF	84.62%	82.35%	90.00%	100%	60.00%	81.25%	100%	84.68%

均识别率与 ResNet18+GF 基本相当, 且高于其他模型。综上所述, DN-ResNet11+GF 在语音情感识别中有着较好的识别性能, 验证了基于 GF 的情感增强算法能够提高模型对情感的表征能力。

表7 IEMOCAP数据集下不同模型的情感识别率

模型	生气	高兴	中性	伤心	平均识别率
ResNet18	74.31%	55.06%	63.30%	57.80%	62.62%
ResNet18+BF	71.43%	55.05%	55.45%	58.82%	60.19%
ResNet18+GF	71.43%	57.79%	62.73%	62.75%	63.68%
ResNeXt	66.06%	44.55%	43.10%	64.22%	54.48%
ResNeXt+BF	72.45%	45.87%	47.27%	52.94%	54.18%
ResNeXt+GF	66.33%	49.54%	48.18%	62.75%	56.33%
DN-ResNet11	70.64%	55.06%	56.03%	69.75%	62.87%
DN-ResNet11+BF	72.73%	55.06%	46.50%	67.27%	60.39%
DN-ResNet11+GF	73.47%	58.72%	53.64%	67.65%	63.37%

2.4 实验3

语谱图和LBP灰度图代表了语音信号特征的不同方面。语谱图提供了关于语音信号频谱分布的信息, 而LBP灰度图则更强调捕捉语音信号的纹理和局部结构特征。通过将这两种不同类型的特征进行融合, 可以综合利用它们的优势, 从而提高情感识别的性能。

调节不同特征的加权系数可以根据实际情况来平衡两种特征对情感识别的贡献。在 CASIA 数据集下对 WFDN_CRNet 模型的加权融合参数(α, β)进行调整。首先, 设置加权参数满足 $\alpha + \beta = 1$, 将

增强情感特征加权参数用 $\alpha = 1 - \beta$ 替代, 不同加权参数下的情感识别率如图 10 所示。当 ($\alpha = 0.8, \beta = 0.2$) 时, 识别效果最佳。

实验中不同模型网络复杂度对比见表 8。可以看出, 由于 CRNet 在 WFDN_CRNet 模型中仅辅助基于 GF 的情感特征增强算法完善纹理信息的作用。对比独立的基线模型, 改进后的 WFDN_CRNet 算法模型比 DN-ResNet11 增加了约 2/5 的时间复杂度, 但相比 ResNet18 与 ResNeXt 的复杂度, WFDN_CRNet 模型的网络复杂度仍有很大降低。

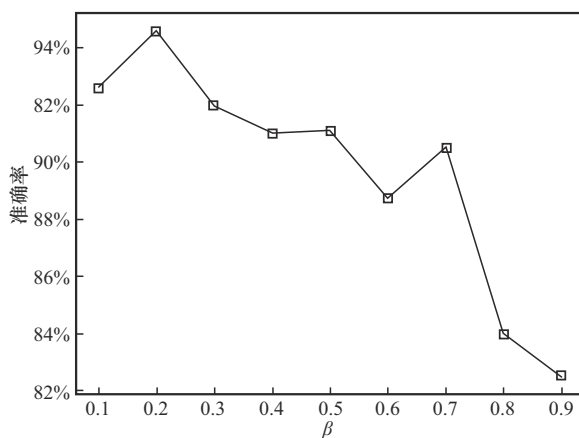


图10 不同加权参数下的情感识别率

为了突出 LBP 灰度化的作用, 增加一组同样使用 WFDN_CRNet 模型构架但 CRNet 支路不再使用 LBP 灰度化而使用单通道的语谱图代替的方法 (记为 WFDN_CRNet0), 和使用 LBP 灰度化

表 8 不同模型网络复杂度对比

模型	FLOPS	总参数量	MemR+W
ResNet18	1.82×10^9	11 179 590	95.09 MB
ResNeXt	3.81×10^9	22 996 038	383.84 MB
CRNet	0.25×10^9	439 081	34.41 MB
DN-ResNet11	0.66×10^9	1 286 692	44.53 MB
WFDN_CRNet	0.91×10^9	1 724 231	78.94 MB

的 WFDN_CRNet 进行对比。经过训练与识别，不同数据集下 DN-ResNet11+GF 与 WFDN_CRNet 的混淆矩阵如图 11，CASIA 数据集下的情感识别率见表 9，EMO-DB 数据集下的情感识别率见表 10，IEMOCAP 数据集下的情感识别率见表 11。

在 CASIA 数据集下，WFDN_CRNet 的平均情感识别率最高，达到 94.58%；相比其他 4 种模型平均识别率均有所提升。在 EMO-DB 数据集

下，WFDN_CRNet 模型的情感识别率最高，达到 85.59%，对比未经过 LBP 辅助支路的 DN-ResNet11+GF，无聊、厌恶、高兴、中性 4 种情感的识别率均提升。在 IEMOCAP 数据集下，WFDN_CRNet 的平均情感识别率为 65.72%，高于其他模型，其中，对比 DN-ResNet11+GF，平均识别率小幅上涨。

综上，从 3 种情感数据集中各模型情感识别率的对比可以看出，WFDN_CRNet 识别率最高。在 CASIA 数据集和 IEMOCAP 数据集下，WFDN_CRNet 和 WFDN_CRNet0 的情感识别率均高于 DN-ResNet11+GF，验证了 LBP 纹理特征提取支路对上支路的辅助效果，同时，WFDN_CRNet 的情感识别率较 WFDN_CRNet0 也有小幅

表 9 CASIA 数据集下的情感识别率

情感类别	ResNet18+GF	ResNeXt+GF	DN-ResNet11+GF	WFDN_CRNet0	WFDN_CRNet
生气	87.50%	90.00%	90.00%	90.00%	97.50%
恐惧	82.50%	82.50%	90.00%	90.00%	90.00%
高兴	95.00%	90.00%	97.50%	97.00%	95.00%
中性	92.50%	85.00%	95.00%	100%	95.00%
伤心	82.50%	82.50%	85.00%	85.00%	90.00%
惊奇	100%	100%	100%	100.00%	100%
平均识别率	90.00%	88.33%	92.92%	93.75%	94.58%

表 10 EMO-DB 数据集下的情感识别率

情感类别	ResNet18+GF	ResNeXt+GF	DN-ResNet11+GF	WFDN_CRNet0	WFDN_CRNet
生气	84.62%	84.62%	84.62%	89.98%	84.62%
无聊	82.35%	71.43%	82.35%	77.80%	88.24%
厌恶	80.00%	60.00%	90.00%	100%	100%
恐惧	92.86%	85.71%	100%	81.38%	85.71%
高兴	60.00%	60.00%	60.00%	67.20%	80.00%
中性	87.50%	93.75%	81.25%	87.50%	87.50%
伤心	100%	100%	100%	76.92%	76.92%
平均识别率	83.72%	80.18%	84.68%	83.97%	85.59%

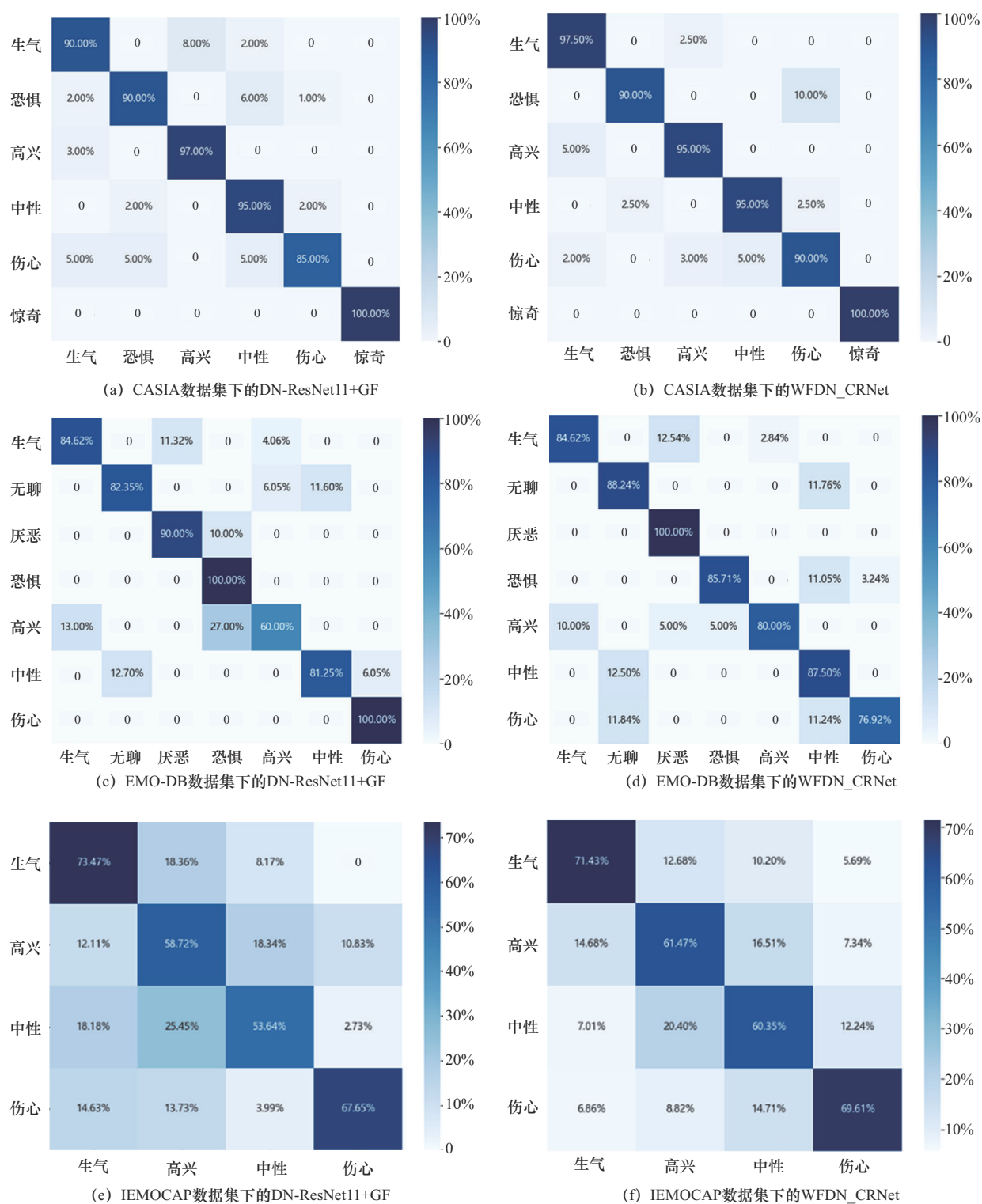


图11 不同数据集下 DN-ResNet11+GF 与 WFDN_CRNet 的混淆矩阵

提升,验证了LBP灰度化的有效性。

3个数据集下其他模型的情感平均识别率见表12,在CASIA数据集和EMO-DB数据集下,

WFDN_CRNet的平均识别率均优于其他模型,在IEMOCAP数据集下,虽然略低于TIM-Net和swim-Transformer模型,但仍优于其他模型。

表 11 IEMOCAP 数据集下的情感识别率

情感类别	ResNet18+GF	ResNeXt+GF	DN-ResNet11+GF	WFDN_CRNet0	WFDN_CRNet
生气	71.43%	66.33%	73.47%	65.15%	71.43%
高兴	57.79%	49.54%	58.72%	59.69%	61.47%
中性	62.73%	48.18%	53.64%	65.11%	60.35%
伤心	62.75%	62.75%	67.65%	68.77%	69.61%
平均识别率	63.68%	56.33%	63.37%	64.68%	65.72%

表 12 3 个数据集下其他模型的情感平均识别率

数据集	模型	平均识别率
CASIA	CNN+RNN+Attention	93.27%
	GM-TCNet	92.50%
	CNN+ChannelAttention	88.67%
	TIM-Net ^[24]	91.08%
	TBDM-Net ^[25]	91.01%
EMO-DB	WFDN_CRNet	94.58%
	Dual-TBNet ^[26]	84.10%
	wav2vec2.0	82.06%
	CNN+LSTM	83.23%
	GM-TCNet	85.07%
IEMOCAP	CNN+ChannelAttention ^[27]	84.57%
	WFDN_CRNet	85.59%
	FCN+Attention	63.9%
	wav2vec2.0	63.11%
	HuBERT	63.87%
	data2vec	54.19%
	TIM-Net	68.7%
	CNN+LSTM	61.70%
	TF-Transformer ^[28]	62.90%
	swim-Transformer ^[29]	65.94%
	WFDN_CRNet	65.72%

综上表明对传统语谱图提取 LBP 特征对上支路增强情感特征在纹理方向的特征增强的有效性，同时加权融合特征的提取增强了情感表征力。此外 WDFN_CRNet 模型由于其低复杂度更适用于中小型数据集，在中小型数据集上达到快速收敛和较高的精度，适用于在资源有限的边缘设备上部署。

3 结束语

本文提出了一种基于 DN-ResNet11 的语音情感识别算法，包括 GF 情感特征增强的 DN-ResNet11 模型和结合 LBP 纹理特征的 WFDN_CRNet 模型。通过语谱图细节增强与情感特征提取，提高了情感表征力，并通过与 CRNet 提取的 LBP 特征加权融合进一步提升性能。在 CASIA 数据集、EMO-DB 数据集和 IEMOCAP 数据集上的实验表明，DN-ResNet11 显著降低了模型复杂度且有效提取了融合情感特征；WFDN_CRNet 在低复杂度下展现了更好的情感区分度和泛化性，验证了细节增强和纹理方向信息增强对情感识别性能的提升。

参考文献：

[1] NWE T L, FOO S W, DE S L C. Speech emotion recognition using hidden Markov models[J]. Speech Communication, 2003, 41(4): 603-623.

[2] WU S, FALK T H, CHAN W Y. Automatic speech emotion recognition using modulation spectral features[J]. Speech Communication, 2011, 53(5): 768-785.

[3] SCHULLER B, RIGOLL G, LANG M. Hidden Markov model-based speech emotion recognition[C]//Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2003, 2: II-1.

[4] SUN L, ZOU B, FU S, et al. Speech emotion recognition based on DNN-decision tree SVM model[J]. Speech Communication, 2019, 115: 29-37.

[5] YENIGALLA P, KUMAR A, TRIPATHI S, et al. Speech emotion recognition using spectrogram & phoneme embedding[C]//

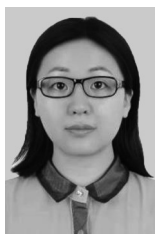


- Proceedings of the 2018 IEEE International Conference on Interspeech. Piscataway: IEEE Press, 2018: 3688-3692.
- [6] LI Z, LI J, MA S, et al. Speech emotion recognition based on residual neural network with different classifiers[C]//Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). Piscataway: IEEE Press, 2019: 186-190.
- [7] WANG J, XUE M, CULHANE R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 6474-6478.
- [8] ZHANG W, JIA Y. A study on speech emotion recognition model based on Mel-spectrogram and CapsNet[C]//Proceedings of the 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST). Piscataway: IEEE Press, 2021: 231-235.
- [9] 金俊林, 于玲, 周晓群. 基于图卷积神经网络的语音情感识别[J]. 信息技术与信息化, 2022(8): 202-205.
- JIN J L, YU L, ZHOU X Q. Speech emotion recognition based on graph convolutional neural network[J]. Information Technology and Informatization, 2022(8): 202-205.
- [10] MRUNAL P G, ABHISHEK V. Automatic recognition of emotions in speech with large self-supervised learning transformer models[C]//Proceedings of the IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings). Piscataway: IEEE Press, 2023.
- [11] CHEN L W, RUDNICKY A. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition[C]//Proceedings of the 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.
- [12] ZI H Z, YAN F W, YU W. Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition[C]//Proceedings of the 2022 IEEE International Conference on Interspeech. Piscataway: IEEE Press, 2022: 725-729.
- [13] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks[J]. In Advances in Neural Information Processing Systems (NeurIPS), 2015: 1135-1143.
- [14] COURBARIAUX M, BENGIO Y, DAVID J P. BinaryConnect: training deep neural networks with binary weights during propagations[J]. In Advances in Neural Information Processing Systems (NeurIPS), 2015: 3123-3131.
- [15] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 1251-1258.
- [16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [17] GREFF K, SRIVASTAVA R K, SCHMIDHUBER J. Highway and residual networks learn unrolled iterative estimation[J]. arXiv preprint arXiv:1612.07771, 2016.
- [18] ORHAN A E, PITKOW X. Skip connections eliminate singularities[J]. arXiv preprint arXiv:1701.09175, 2017.
- [19] XU J, LI Z, DU B, et al. Reluplex made more practical: leaky ReLU[C]//Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC). Piscataway: IEEE Press, 2020: 1-7.
- [20] HE K, SUN J, TANG X. Guided image filtering[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(6): 1397-1409.
- [21] WU H, ZHENG S, ZHANG J, et al. Fast end-to-end trainable guided filter[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1838-1847.
- [22] PRAKASA E. Texture feature extraction by using local binary pattern[J]. INKOM Journal, 2016, 9(2): 45-48.
- [23] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11534-11542.
- [24] JIA X Y, XIN C W, YU J W, et al. Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023.
- [25] VLAD S C, COSMIN S C, ADRIANA S. TBDM-Net: bidirectional dense networks with gender information for speech emotion recognition[C]//Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing. Piscataway: IEEE Press, 2024.
- [26] ZHENG L, XIN K, FU J R. Dual-TBNet: improving the robustness of speech features via dualtransformer-BiLSTM for speech emotion recognition[C]//Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing. Pis-

cataway: IEEE Press, 2023: 2193 – 2203.

- [27] ZHU R F, SUN C X, WEI X P, et al. Speech emotion recognition using channel attention mechanism[C]//Proceedings of the IEEE International Conference on Computer Engineering and Application (ICCEA). Piscataway: IEEE Press, 2023: 680-684.
- [28] YONG W, CHENG L, YUAN Z, et al. Time-frequency transformer: a novel time frequency joint learning method for speech emotion recognition[C]//Proceedings of the International Conference on Neural Information Processing (ICONIP). Changsha: Central South University Press, 2023.
- [29] LIAO Z, SHEN S. Speech emotion recognition based on swin-transformer[J]. Journal of Physics: Conference Series, 2023: 2508(1), 012056.

[作者简介]



应娜 (1978–), 女, 博士, 杭州电子科技大学通信工程学院副教授、硕士生导师, 主要研究方向为智能信号处理与应用。



邹雨鉴 (1998–), 男, 杭州电子科技大学通信工程学院硕士生, 主要研究方向为多模态情感识别。



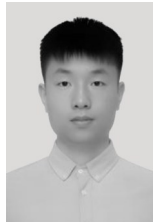
杨雪滢 (1997–), 女, 杭州电子科技大学通信工程学院硕士生, 主要研究方向为语音情感识别。



孙文胜 (1966–), 男, 现就职于杭州电子科技大学通信工程学院, 主要研究方向为网络通信。



叶学义 (1973–), 男, 博士, 杭州电子科技大学通信工程学院副教授、硕士生导师, 主要研究方向为图像处理、模式识别、信息隐藏。



蒋银河 (1999–), 男, 杭州电子科技大学通信工程学院硕士生, 主要研究方向为多模态情感识别。