DOI:10.19652/j.cnki.femt.2305549

基于幅值滤波与分层特征融合策略的语音情感识别*

喻永振 刘大明
(上海电力大学计算机科学与技术学院 上海 200090)

摘 要:针对语音情感识别在多语言联合数据集上识别准确率低的问题,提出了一种基于幅值滤波与分层特征融合策略的语音情感识别方法。该方法首先对梅尔谱图内幅值分布规律进行幅值滤波,通过概率叠加扩大梅尔谱图内相近幅值之间的差异,实现谱图内的高频强增益、低频弱增益;同时,通过概率相乘缩小梅尔谱图内相远幅值之间的差异,以显示谱图内中频的细节部分。在此基础上,使用矩形卷积提取音频信号的时间动态特征,生成梅尔谱图动态特征图,并将其作为分层特征融合策略的输入。分层特征融合策略通过压缩特征图来提取不同尺度的时间动态特征,并提取不同深度中的时间动态特征。在多语言联合数据集 CER 上取得了 84.44%的分类准确率。

关键词:语音情感识别;幅值滤波;分层特征融合策略;梅尔谱图动态特征图
中图分类号:TN912.3 **文献标识码**:A **国家标准学科分类代码**:520.20

Speech emotion recognition based on amplitude filtering and hierarchical feature fusion strategy

Yu Yongzhen Liu Daming
(School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: A speech emotion recognition method based on amplitude filtering and hierarchical feature fusion strategy is proposed in response to the problem of low accuracy of speech emotion recognition on multi-language joint datasets. The method first applies amplitude filtering to the amplitude distribution pattern in the Mel spectrogram, enlarging the differences between similar amplitudes and achieving high frequency strong gain and low frequency weak gain within the spectrogram. Meanwhile, by multiplying probabilities, it reduces the differences between distant amplitudes in the Mel spectrogram, displaying the detailed middle frequency components. Based on this, the method uses rectangular convolution to extract the temporal dynamic features of the audio signal, generating dynamic feature maps of the Mel spectrogram, which serve as inputs to the hierarchical feature fusion strategy. The hierarchical feature fusion strategy compresses the feature maps to extract temporal dynamic features of different scales and from different depths. The proposed method achieves a classification accuracy of 84.44% on the multi-language joint dataset CER.

Keywords: speech emotion recognition; amplitude filtering; hierarchical feature fusion strategy; dynamic feature map of Mel spectrogram

0 引 言

语音情感识别(speech emotion recognition,SER)[1]是指机器能够自动从语音信号中识别人的情感状态,如高兴、悲伤等。随着人工智能的快速发展,语音情感识别技术在机器人技术、移动服务、呼叫中心、电脑游戏以及心理测试等领域中得到广泛应用。然而,语音情感识别的准确率常常受到情感表达的多样性、复杂性以及不同个体之间的情感表达差异等因素的影响,导致其准确率相对较低。

通常情况下,大多数语音情感识别模型包含特征提取和分类器两个部分。特征提取可以根据需要进行单模态或多模态处理。传统的单模态是从原始的语音信号中提取出情感信息,李翔等[2]将声学参数、音质参数和 Teager 能量算子的非线性特征相结合,采用隐藏马尔科夫模型

收稿日期:2023-09-14
* **基金项目**:上海市科技计划项目(23010501500)资助

(hidden Markov models, HHM)作为分类器,设计智能服务机器人的情感交互框架。然而,直接将手工特征输入分类器的方法仍不成熟,因为这些特征中包含了一些与情感无关的信息,会干扰情感识别。许良凤等[3]将语音样本转换成语谱图,并采用 Log-Gabor 滤波器[4]和完全局部二值模式构建新的图特征。Zhao 等[5]通过将原始音频信号转换成梅尔谱图。梅尔谱图中含有低电平信号,通过几个卷积层可以映射到潜在的时间序列上。对于多模态处理方式,Wang 等[6]指出过于随意的使用文本信息可能会导致误导性的预测,因为相同的文本内容在被不同的情感激发时具有完全不同的意义。因此,进一步充分探索音频信号的全部潜力是必要的。

随着深度学习的不断发展,循环神经网络(recurrent neural network, RNN)和卷积神经网络(convolutional neural network,CNN)已经被广泛应用于情感识别。研究表明,基于 CNN 的模型在情感分类方面展现出更好的性能,因此受到了更多的关注和改进。张雄等[7]使用 CNN 从语谱图中提取卷积特征并融合统计学特征。Tursunov 等[8]采用修改后的卷积核和池化策略(Deep-Net),从谱图中提取出具有更强区分性和可靠性的深度频率特征。在情感识别中,时间动态特征能够反映情绪的变化,对准确识别情感起到重要作用。Zhao 等[9]通过计算 log-Mel 的 deltas 和 delta-deltas 两种时间动态特征,构建具有时间动态特征的 3D 频谱图,然后将其直接输入带有 CTC 损失的膨胀卷积方法中,学习高层次的情感表示。虽然这种方法增加了手动工作量和数据存储需求,但选择合适的初始时间动态特征也非常重要的。

为解决上述问题,一些研究将 CNN 与其他能够学习时间动态特征的网络进行集成。通过融合不同网络的优势,更好地捕捉时间动态特征。Meyer 等[10]提出了一种集成了 CNN 和 RNN 的方法来学习高层次特征,通过 RNN 层获取时间动态特征。考虑到 RNN 的计算效率较低,Liu 等[11]使用时间卷积网络(time convolutional network, TCN)代替 RNN 来提取时间动态特征,提高学习效率。Mustaqeem 等[12]采用扩张卷积神经网络来捕捉频率特征,应用多头注意机制来捕捉深层特征,从而在较短的计算时间内展现出较好的识别性能。虽然这些方法都考虑了时间动态特征,但它们忽视了不同尺度和深度上的时间动态特征。

Wu 等[13]通过对时间序列多周期的观察,提出了二维结构的时间序列。通过将时间变化扩展到二维空间,可以将周期内和周期间的变化嵌入到二维张量的行和列中。然后,通过计算每个频率的幅值,提取出最显著的频率信息。为了避免无意义的高频噪声,还对显著频率进行了非归一化处理。另外,Chen 等[14]研究了不同核大小的卷积在不同尺度下对目标检测性能的影响,提出了一种具有分层特征融合策略的新型块 MS-Block,以增强实时目标检测器对多尺度特征的提取能力,并保持快速的推理速度。

因此,本文结合梅尔谱图的幅值分布规律和分层特征融合策略的优势,提出了一种基于幅值滤波与分层特征融合策略的语音情感识别方法(amplitude filtering and hierarchical feature fusion strategy ,AF-HFFS),用于增强梅尔谱图并提取不同尺度和深度中的时间动态特征,提高识别的准确率。

1 AF-HFFS 模型

AF-HFFS 模型如图 1 所示。该模型以梅尔谱图作为输入特征,通过幅值滤波动态特征(amplitude filtering dynamic feature,AFDF)网络提取音频信号中的时间动态特征,生成梅尔谱图动态特征图。然后,通过分层特征融合策略(hierarchical feature fusion strategy,HFFS)提取不同尺度、不同深度的时间动态特征。在分类过程中,采用联合损失函数(中心损失、交叉熵损失、性别损失)来训练 AF-HFFS。

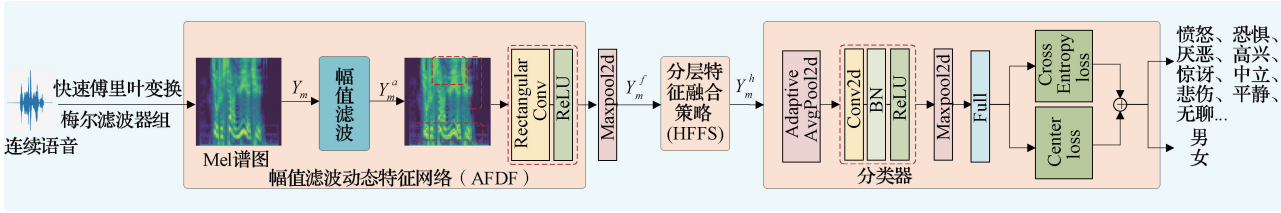

图 1 AF-HFFS 模型

1.1 幅值滤波动态特征网络

1)梅尔谱图特征提取

谱图是用于分析语音和音频波形的二维图形表示,展示了不同频率信号强度随时间变化的特征。在谱图上,横轴表示时间,纵轴表示频率,每个点的颜色则对应着特定时间和频率的幅值。值得注意的是,谱图上的频率通常是线性分布的,而人耳对频率的感知却不是线性关系,即对于低频段的变化更加敏感,而对于高频段的变化更加迟钝。为此,研究者提出了梅尔谱图。梅尔谱图通过对频率进行转换,使得频率的刻度更符合人耳的感知特性。梅尔谱图频率与信号频率的转换公式为:

$$Mel(k) = 2\ 595 \times \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

式中:k 表示梅尔尺度的频率;f 表示普通频率。

梅尔谱图提取过程如图2所示。

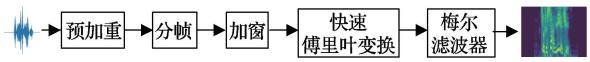

图2 梅尔谱图提取过程

首先将语音信号 x_t 进行预加重，实现对原始语音信号高频成分的补偿，然后通过分帧、加窗（汉宁窗）得到离散的语音信号 x_n，对 x_n 的每一帧进行快速傅里叶变换（FFT），将其转化为频域信号 x_k，再利用梅尔滤波器组对 x_k 进行处理，输出相应的梅尔谱图 x_m。

2）幅值滤波

每张梅尔谱图的横轴表示时间 t，纵轴表示经过 n 个梅尔滤波器生成的频率，n 设定为64。梅尔谱图尺寸为64×126 pixels。在梅尔谱图中，幅值越大，颜色越亮。从时间轴来看，同一时间帧内不同频域段的幅值变化不同；从频率轴来看，不同频率在时间域内也呈现出不同的变化。为了将不同位置的幅值在两个坐标轴上的变化规律趋于一致，采取softmax函数[15]来计算不同位置幅度值的概率。每个时间帧内或频率上的幅值概率总和为1。由于整个梅尔谱图中所有幅值均为负数，在计算概率时，幅值大的区域，概率值相对较小，同时，幅值权重图上的值趋近于0，这可能导致滤波后，梅尔谱图中所有幅值变得统一，减少了差异性特征。因此，采用1减去概率值的方式，既能增强高幅值区域的特征，又能对低幅值区域进行弱增益。单纯从时间轴或频率轴对幅值进行增益会忽略它们之间的关联性，因此，将双轴上获得了概率值进行组合。有3种组合方法：(1)将时间轴和频率轴的概率值进行相乘；(2)将时间轴和频率轴的概率值进行叠加；(3)通过自学习参数[16]将组合(1)与组合(2)进行融合。幅值滤波示意图如图3所示。

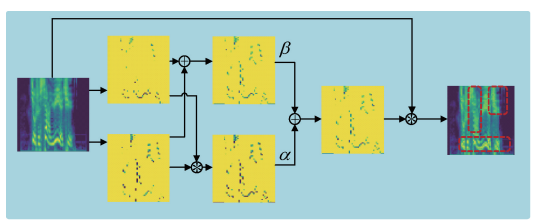

图3 幅值滤波

概率叠加的方法可以扩大梅尔谱图内相近特征值之间的差异，实现对明亮特征的强增益以及对暗淡区域特征的弱增益；概率相乘的方法可以缩小梅尔谱图内相远特征值之间的差异，显示中间特征值，从而实现对细节部分的增益补充。

自学习参数使用AdamW[17]优化，步骤如下：

步骤1）先将自学习参数初始化为1；

步骤2）根据模型在训练集上的损失函数值对参数求导，计算出梯度值；

步骤3）将梯度衰减添加到梯度中；

步骤4）通过指数加权平均，计算出当前梯度的一阶矩估计和二阶矩估计；

步骤5）根据一阶矩估计和二阶矩估计，对自学习参数进行更新。

3种组合对梅尔谱图的增益效果如图4所示，计算公式如下：

$$\boldsymbol{M}_t = |1-softmax(\boldsymbol{m}_t)| \tag{2}$$

$$\boldsymbol{M}_n = |1-softmax(\boldsymbol{m}_n)| \tag{3}$$

$$Mel(\boldsymbol{m}_a) = (\boldsymbol{M}_t+\boldsymbol{M}_n)\times Mel(\boldsymbol{k}) \tag{4}$$

$$Mel(\boldsymbol{m}_m) = (\boldsymbol{M}_t\times\boldsymbol{M}_n)\times Mel(\boldsymbol{k}) \tag{5}$$

$$Mel(\boldsymbol{m}_{am}) = \frac{\alpha\times Mel(\boldsymbol{m}_m)+\beta\times Mel(\boldsymbol{m}_a)}{2} \tag{6}$$

式中：$\boldsymbol{m}_t$ 表示在时间轴方向的幅值；$\boldsymbol{m}_n$ 表示在频率轴方向的幅值；$\boldsymbol{M}_t$ 表示幅值在时间轴方向的概率值；$\boldsymbol{M}_n$ 表示幅值在频率轴方向的概率值；$\boldsymbol{k}$ 表示梅尔谱图的幅值；α、β 为自学习参数。

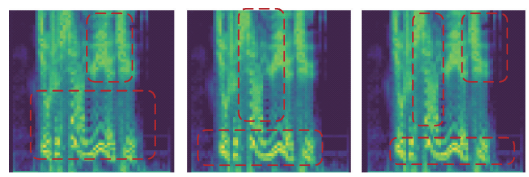

图4 双轴概率组合增益图

3）动态特征表示

幅值沿着时间轴方向的变化具有音频的时间动态特征，因此，使用了矩形卷积[18]来获取幅值前后两帧的幅值变化。设定频率轴方向核尺寸为1，时间轴方向上的核尺寸为3，通过将卷积层中的填充和扩张设置为1，使卷积层将当前时刻以及相邻的幅值信息都考虑在内。每一列频率 $\boldsymbol{X}$ 中不同位置的时间动态特征，计算公式如下：

$$H[n] = \sum_{i=0}^{k-1} X[n+i]\cdot C[i],\ k\geqslant 1 \tag{7}$$

式中：$H[n]$ 是卷积运算第 n 个位置输出的时间动态特征；$C[i]$ 是卷积核内第 i 个参数值；$X[n+i]$ 是输入某一列频率对应的时间序列上第 $n+i$ 个位置上的值；k 为时间轴上卷积核尺寸。

将滤波增益后的梅尔谱图 $\boldsymbol{Y}_m$ 输入到一个3×1的矩形卷积层中，生成基本的动态特征 $\boldsymbol{X}\in\boldsymbol{R}^{C\times T\times F}$，其中 C 表示通道数，在实验中 $C=32$。使用ReLU激活函数拟合输出的时间动态特征，激活部分神经元。使用最大池化过滤动态特征图中的噪声，生成梅尔频谱动态特征图 $\boldsymbol{Y}_m^f\in\boldsymbol{R}^{C\times T\times F}$，将其作为HFFS模块的输入。

1.2 分层特征融合策略

不同尺度、不同深度上的时间动态特征影响语音情感的识别效果。因此，在AF-HFFS中，通过HFFS学习不同尺度和深度上的时间动态特征。

分层特征融合策略的结构如图5所示。以梅尔谱图动态特征图$\boldsymbol{Y}_m^f \in \boldsymbol{R}^{C\times T\times F}$首次输入HFFS为例。首先，使用一个3×3卷积层，步长为2，通道数$C=64$，经过BN[19]层进行正则化加速，然后使用ReLU激活函数对输出的时间动态特征进行拟合，激活部分神经元，输出时间动态特征图$\boldsymbol{Y}_r$，尺寸由$T\times F$缩减为$T/2\times F/2$，实现动态特征图的尺寸压缩。然后，采用分层融合策略[14]，通过一个1×1卷积层，步长为1，将通道数C从64扩展为$64\times n$，实现通道数的扩充。将通道数C分成n组，$\boldsymbol{X}_i$表示第i组，$i\in 1,2,3,\cdots,n$，在实验中i设置为3，实验中除了$\boldsymbol{X}_1$之外，其他每一组都经过一个反向瓶颈层。$\boldsymbol{X}_1$作为跨级连接，保留前一层提取到的时间动态特征$\boldsymbol{Y}_1$。最后，将所有组的时间动态特征连接在一起，使用1×1卷积层将通道压缩为64，实现不同通道、不同尺度之间的交互。

$$\boldsymbol{Y}_r = ReLU(BN(Conv(\boldsymbol{Y}_m^f))) \tag{8}$$

$$\boldsymbol{X} = Conv(\boldsymbol{Y}_r) \tag{9}$$

$$\boldsymbol{Y}_i = \begin{cases} \boldsymbol{X}_i, & i=1 \\ IB_{3\times3}(\boldsymbol{Y}_{i-1}+\boldsymbol{X}_i), & i>1 \end{cases} \tag{10}$$

$$\boldsymbol{Y}_m{}^h = Conv\left(\sum_{i=1}^{n}\boldsymbol{Y}_i\right), \quad n\geqslant 1 \tag{11}$$

式中：$IB_{3\times3}(\cdot)$表示反向瓶颈层。

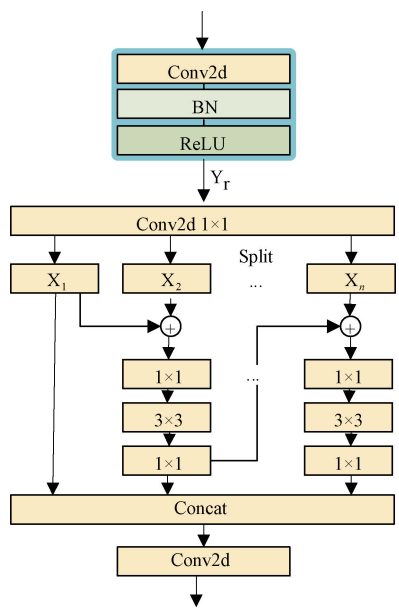

图5 分层特征融合策略

在实验中，分层特征融合策略的使用次数设置为3。由于提取的梅尔谱图尺寸为64×126，在经过3次尺度变换后，尺寸变为8×16，此时，时间动态特征已具有较强的判别性，再次压缩可能会导致尺寸不足和信息丢失。

1.3 分类器

在情感分类模型中将性别作为辅助任务，因此全连接层的输出神经元分别为M（情感类别）和2（性别类别）。为了综合考虑情感分类和性别分类的训练效果，采用中心损失[20]和Softmax交叉熵损失的联合损失函数。在训练过程中，使用AdamW优化器并应用权重衰减技术来减轻过拟合问题。

模型的损失函数定义如下：

$$L_m = (L_s+\lambda\times L_c)\times(\gamma\times L_e+L_g) \tag{12}$$

式中：L_s、L_c、L_e和L_g分别表示Softmax交叉熵损失、中心损失、情感类别分类损失和性别分类损失；λ用于表示中心损失在联合损失函数中所占比重；γ表示情感分类任务占分类任务的比重。

2 实验结果与分析

2.1 数据集描述

为了评估AF-HFFS的性能，采用了CASIA[21]、Emo-DB[22]、RAVDESS[23]以及自行构建的联合数据集CER，实现对AF-HFFS在不同背景和情境下的性能评估。

CASIA是由中国科学院录制的汉语情感语料库。该数据集由4位专业发音人（2男、2女）演绎的6种不同情绪，包括愤怒、恐惧、高兴、中立、悲伤和惊讶。公开的CASIA数据集共包含1 200条语句，每个音频文件的平均时长约为1.9 s。

EMO-DB是由柏林大学录制的德语语音情感数据集。该数据集由10位专业演员（5男、5女）演绎的7种不同情感，包括愤怒、无聊、厌恶、恐惧、高兴、中立和悲伤。EMO-DB共有535条语句，每个音频文件的平均时长约为2.7 s。

RAVDESS是一个被广泛应用于歌曲和对话的英语情感数据集。该数据集包含了24位专业演员（12男、12女）演绎的8种不同情绪，包括悲伤、平静、高兴、愤怒、惊讶、中立、恐惧和厌恶。RAVDESS数据集共录制了1 440个音频文件。

联合数据集CER是将CASIA、Emo-DB和RAVDESS 3个数据集合并。如果某类情感只在单个数据集中存在，那么将其作为联合数据集中的一类。如果是共有的情感类别则直接合并为联合数据集中的一类情感。合并后的数据集包含愤怒、恐惧、厌恶、高兴、惊讶、中立、悲伤、平静和无聊9种不同的情绪，共有3 175个音频文件。

2.2 实验设置

首先，使用第3方库Torchaudio来读取每个.wav文件。然后，对音频数据进行重新采样，将采样率设置为16 kHz。为了固定音频长度为4 s，采用零填充和截断的方法，同时，将多声道立体声转换为单声道。接下来，从4 s的音频中提取梅尔谱图。采用汉宁窗（Hanning window）作为窗口函数，窗口长度为64 ms（1 024点），帧移为32 ms（512点），采用64个Mel滤波器组。

由于本文所使用的CASIA、Emo-DB、RAVDESS和CER 4个数据集的数据量较小，而针对于小数据集（数据

量在万以下),主流的数据集划分方法是将数据集划分为训练集、测试集和验证集,其中,验证集主要用于调节卷积核个数、卷积核大小等超参数。由于本文实验主要在同一平台上进行纵向比较以确定最佳的改进方向,在这个过程中使用验证集调节超参数无法控制单一变量,因此在进行实验评估时,将数据集按照 8∶2 的比例随机划分为训练集和测试集。使用未加权准确率(unweighted accuracy,UA)、精准率(precision,P)、召回率(recall,R)和 F1-得分(F1-score)作为评估指标。对于多分类任务,将每一类单独视为正样本,所有其他类型视为负样本。*UA* 为正确预测的样本数占总样本数的比重;*P* 为预测为正的样本数占真实为正的样本数的比重;*R* 为预测为正的样本数占预测为正的样本数的比重;F1-score 为 *P* 和 *R* 的调和平均值。

$$UA = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$F1 - \text{score} = 2 \cdot \frac{P \cdot R}{P + R} \tag{16}$$

式中:*TP* (true positive)表示预测为正样本实际也为正样本的数量;*TN* (true negative)表示预测为负样本实际也为负样本的数量;*FP* (false positive)表示预测为正样本实际为负样本的数量;*FN* (false negative)表示预测为负样本实际为正样本的数量。

此外,还使用混淆矩阵来分析每个情感类别的识别精度。在实验中,将学习率设置为 5×10^{-4},批大小设置为 16,γ 设置为 1,λ 设置为 0.01,迭代次数设置为 100 次。

实验所采用的硬件环境包括 Intel(R) Xeon(R) Platinum 8255C 的 CPU 和 RTX 2080 11 G 显卡。开发语言为 Python 3.8.10,深度学习框架为 Pytorch。

2.3 实验分析

实验主要包括如下 3 个方面:1)在不同数据集上验证 AFDF 对性能的影响;2)在联合数据集上验证 AF-HFFS 的有效性;3)与其他基线方法进行比较。

从表 1 可以看出,在 Emo-DB 和 RAVDESS 数据集上,采用组合 b 获得了最高的准确率(Emo-DB 为 87.38%,RAVDESS 为 80.28%);在 CASIA 和 CER 数据集上,采用组合 c 获得了最高的准确率(CASIA 为 88.33%,CER 为 80.48%)。使用未处理的梅尔谱图的平均耗时最短,主要因为谱图的处理步骤是在 CPU 上进行的。结果表明,在 Emo-DB 和 RAVDESS 数据集上,判别性特征主要集中在幅值的均值附近,而在 CASIA 和 CER 数据集上,判别性特征在幅值的均值和差异值附近都存在。

表 1 不同 AFDF 组合的识别结果

方法	UA/%				平均耗时/μs
	CASIA	Emo-DB	RAVDESS	CER	
梅尔谱图	86.67	83.5	77.82	75.71	**237**
组合 a	86.67	82.52	79.23	77.94	246
组合 b	86.25	**87.38**	**80.28**	76.98	245
组合 c	**88.33**	82.52	78.17	**80.48**	279

从表 2 可以看出,组合 a 对应的 α 值小于组合 b 对应的 β 值。结果表明,在梅尔谱图上,具有强判别性的特征主要来自于幅值的均值较远的部分,考虑到均值附近也存在具有判别性的特征,因此,需要通过自学习参数来调整出最佳的特征表达。

表 2 AFDF 组合 c 的学习参数值变化

自学习参数值	CASIA	Emo-DB	RAVDESS	CER
α	0.956 6	0.972 4	0.950 7	0.817
β	1.019 7	0.995 6	1.039 1	1.123

AF-HFFS 中不同模块在联合数据集上的识别效果如图 6 所示,可以看出,使用 AFDF 和 HFFS 可以提升联合数据集 CER 上的所有性能指标。尤其是联合使用 AFDF 和 HFFS 可以获得最佳的识别效果(UA 为 84.44%,*P* 为 83.40%,*R* 为 86.48%,*F*1-score 为 84.60%)。图 7 和 8 所示分别为仅使用梅尔谱图和 AF-HFFS 的混淆矩阵,可以看出,除了愤怒情感的准确度下降(77.78%),模型对其他情感的识别率都有显著的提升,说明在不同尺度、不同通道之间都存在具有判别性的时间动态特征。

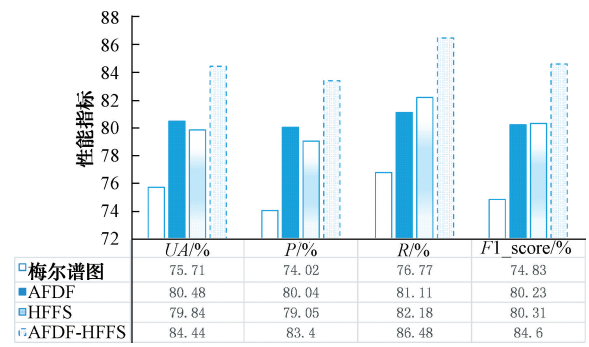

图 6 AF-HFFS 的识别效果

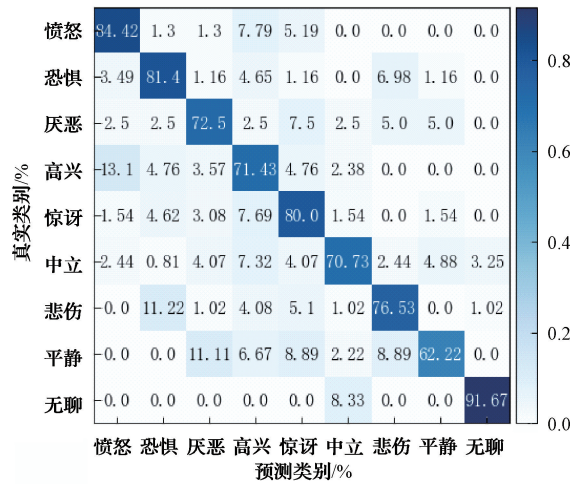

图 7　梅尔谱图的混淆矩阵

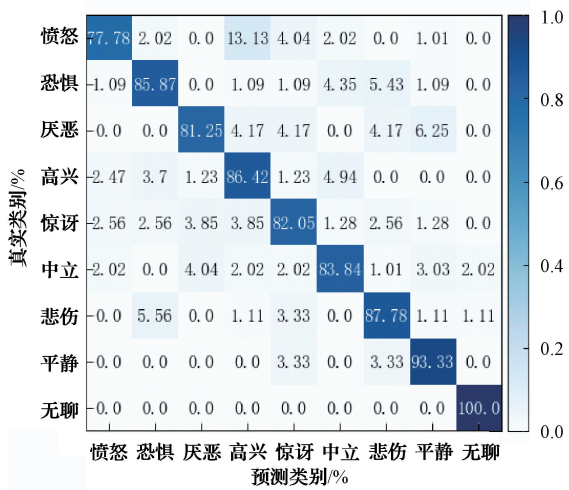

图 8　AF-HFFS 的混淆矩阵

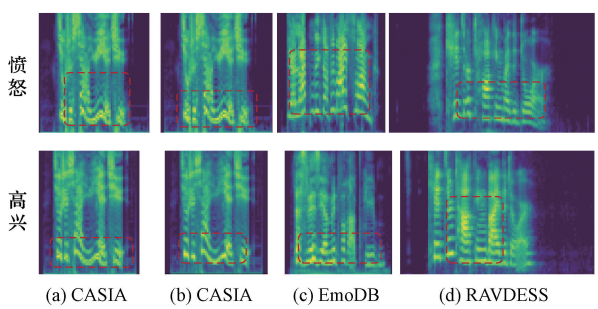

图 9　愤怒与高兴的梅尔谱图

愤怒与高兴在不同数据集上的梅尔谱图如图 9 所示，其中，图 9(a)是经过幅值滤波增益的梅尔谱图。在 EmoDB 和 RAVDESS 两个数据集上愤怒和高兴的梅尔谱图呈现出相互交叉的特征趋势，幅值滤波在 CASIA 数据集上虽然有效的拉开了两种情感的特征变化趋势，然而，又与其他两个数据集上的相同情感的变化趋势相近，可以看出这主要由于不同数据集的差异所引起，AF-HFFS 在联合数据集 CER 上，通过幅值滤波和探索不同尺度、通道之间的时间动态特征，缓和不同数据集之间引起的差异。

实验主要将本文提出的 AF-HFFS 与其他基线方法进行性能比较。

1)Trumpet-6[24]，通过添加高斯白噪来增加样本量实现数据扩充，并搭建一种轻量级卷积神经网络模型，在 CASIA 数据集和 Emo-DB 数据集上分别得到了 95.70%和 83.40%的准确率，模型结构简单，参数量少，对中文语音识别效果好。

2)DRN-A-CASA[25]，通过添加高斯白噪和时频掩码来增加样本量实现数据扩充，使用膨胀卷积替换残差网络卷积层，引入注意力机制和辅助分类器，从梅尔对数频谱图中提取情感特征，在 Emo-DB 数据集和 RAVDESS 数据集上分别得到了 89.15%和 92.91%的准确率，模型结构较深，待训练参数量较大。

3)DeepESN[26]，使用梅尔频率倒谱系数作为音频描述方法，引入扩张卷积和多头注意力机制，在 Emo-DB 数据集和 RAVDESS 数据集上分别得到了 85.57%和 77.02%的准确率。

4)CNN[27]，通过选取 10 个低水平特征(基音频率、频谱质心和谱对比度等)的均值、标准差、最小值和最大值，构建一个手工特征集并输入到卷积神经网络，在 CASIA 数据集、Emo-DB 数据集和 RAVDESS 数据集上分别得到了 82.50%、75.80%和 74.30%的准确率。

5)SA-CGRU[28]，通过将自注意力机制引入卷积门控循环网络，从语谱图的时间轴方向捕获语音信号的时序特征，在 CASIA 数据集、Emo-DB 数据集和 RAVDESS 数据集上分别得到了 87.58%、88.41%和 83.87%的准确率。

比较结果如表 3 所示，可以看出，AF-HFFS 方法相比于 CNN 方法在 CASIA、Emo-DB 和 RAVDESS 数据集上有一定的提升，并取得了与 SA-CGRU 方法相似的识别效果。与 Trumpet-6、DRN-A-CASA 和 DeepESN 方法相比，识别效果较稳定，且保持了 Trumpet-6 的轻量型特性。

表 3　与其他基线方法的比较

方法	CASIA	Emo-DB	RAVDESS	参数量/($\times 10^6$)
Trumpet-6[24]	95.70	83.40	—	0.18
DRN-A-CASA[25]	—	89.15	92.91	14.90
DeepESN[26]	—	85.57	77.02	—
CNN[27]	82.50	75.80	74.30	—
SA-CGRU[28]	87.58	88.41	83.87	—
AF-HFFS	87.92	87.38	81.34	0.27

3 结 论

本文提出了一种基于AF-HFFS语音情感识别方法，用于增强梅尔谱图并提取不同尺度和深度中的时间动态特征。AFDF网络可以根据谱图的幅值分布规律，对相应位置进行增益，并生成梅尔谱图时间动态特征图。HFFS能够同时获取不同尺度和深度中提取的时间动态特征，并实现通道间的交互。实验结果表明，AF-HFFS在语音情感识别方面能够有效分类，并在联合数据集上依旧保持稳定的识别性能。未来的工作将集中在如何将模型轻量化，以实现在计算能力有限的嵌入式芯片上运行。

参 考 文 献

[1] AL-DUJAILI M J, EBRAHIMI-MOGHADAM A. Speech emotion recognition: A comprehensive survey[J]. Wireless Personal Communications, 2023, 129(4): 2525-2561.

[2] 李翔，李昕，胡晨，等. 面向智能机器人的Teager语音情感交互系统设计与实现[J]. 仪器仪表学报，2013，34(8)：1826-1833.

[3] 许良凤，刘泳海，胡敏，等. 语谱图改进完全局部二值模式的语音情感识别[J]. 电子测量与仪器学报，2018，32(5)：25-32.

[4] GU Y, POSTMA E, LIN H X. Vocal emotion recognition with log-gabor filters[C]. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, 2015: 25-31.

[5] 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.

[6] WANG J, XUE M, CULHANE R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]. ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6474-6478.

[7] 张雄，刘蓉，刘明. 基于卷积特征提取与融合的语音情感识别研究[J]. 电子测量技术，2018，41(16)：138-142.

[8] TURSUNOV A, KHAN M, KWON S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features[J]. Sensors, 2020, 20(18): 5212.

[9] ZHAO Z, LI Q, ZHANG Z, et al. Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition[J]. Neural Networks, 2021, 141: 52-60.

[10] MEYER P, XU Z, FINGSCHEIDT T. Improving convolutional recurrent neural networks for speech emotion recognition[C]. 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021: 365-372.

[11] LIU J, LIU Z, WANG L, et al. Temporal attention convolutional network for speech emotion recognition with latent representation[C]. INTERSPEECH, 2020: 2337-2341.

[12] MUSTAQEEM K, EL SADDIK A, ALOTAIBI F S, et al. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network[J]. Knowledge-Based Systems, 2023, 270: 110525.

[13] WU H, HU T, LIU Y, et al. TimesNet: Temporal 2d-variation modeling for general time series analysis[J]. arXiv preprint arXiv:2210.02186, 2022.

[14] CHEN Y, YUAN X, WU R, et al. YOLO-MS: Rethinking multi-scale representation learning for real-time object detection[J]. arXiv preprint arXiv: 2308.05480, 2023.

[15] GAO B, PAVEL L. On the properties of the softmax function with application in game theory and reinforcement learning[J]. arXiv preprint arXiv: 1704.00805, 2017.

[16] LU Z, LI J, LIU H, et al. Transformer for single image super-resolution[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 457-466.

[17] LOSHCHILOV I, HUTTER F. Fixing weight decay regularization in Adam[J]. arXiv preprint arXiv: 1711.05101, 2018.

[18] OORD A, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[19] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. International Conference on Machine Learning, 2015: 448-456.

[20] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C]. Computer Vision-ECCV 2016: 14th European Conference. Springer, 2016: 499-515.

[21] LI Y, TAO J, CHAO L, et al. CHEAVD: A Chinese natural emotional audio-visual database[J]. Journal of Ambient Intelligence and Humanized Computing, 2017, 8: 913-924.

[22] BURKHARDT F, PAESCHKE A, ROLFES M, et al. A database of German emotional speech[C].

Interspeech, 2005: 1517-1520.
[23] LIVINGSTONE S R, RUSSO F A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English[J]. PloS one, 2018, 13(5): e0196391.
[24] 乔栋,陈章进,邓良, 等.基于改进语音处理的卷积神经网络中文语音情感识别方法[J].计算机工程,2022,48(2):281-290.
[25] 周佳鑫,焦亚萌,王彦斌, 等.融合注意力和辅助分类器的膨胀残差网络语音情感识别研究[J].国外电子测量技术,2023,42(8):19-25.
[26] MUSTAQEEM K, EL SADDIK A, ALOTAIBI F S, et al. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network [J]. Knowledge-Based Systems, 2023, 270: 110525.
[27] 谷泽月,边巴旺堆,祁晋东.基于多特征融合的藏语语音情感识别[J].现代电子技术,2023,46(21):129-133.
[28] 孙韩玉,黄丽霞,张雪英, 等.基于双通道卷积门控循环网络的语音情感识别[J].计算机工程与应用,2023,59(2):170-177.

作者简介

喻永振,硕士研究生,主要研究方向为深度学习、语音情感识别。

刘大明,博士, 副教授,硕士生导师,主要研究方向为物联网技术、嵌入式系统等。
E-mail:ldm@shiep.edu.cn