基于小波散射变换和MFCC的双特征语音情感识别融合算法

应娜，吴顺朋，杨萌，邹雨鉴
（杭州电子科技大学通信工程学院，浙江 杭州 310018）

摘　要：为了充分挖掘语音信号频谱包含的情感信息以提高语音情感识别的准确性，提出了一种基于小波散射变换和梅尔频率倒谱系数（Mel-frequency cepstral coefficient，MFCC）的排列熵加权和偏差调整规则的语音情感识别融合算法（PEW-BAR）。算法首先获取语音信号的小波散射特征和梅尔频率倒谱系数的相关特征；然后按尺度维度扩展小波散射特征，利用支持向量机得到情感识别的后验概率并获得排列熵，并使用排列熵对后验概率进行加权；最后采用一种偏差调整规则进一步融合MFCC的相关特征的识别结果。实验结果表明，在EMODB、RAVDESS和eNTERFACE05数据集上，与传统的基于小波散射系数的语音情感识别方法相比，该算法将ACC分别提高了2.82%、2.85%和5.92%，将UAR分别提升了3.40%、2.87%和5.80%，IEMOCAP上提高了6.89%。

关键词：语音情感识别；小波散射变换；排列熵；MFCC；模型融合
中图分类号：TP18
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024088

Dual-feature speech emotion recognition fusion algorithm based on wavelet scattering transform and MFCC

YING Na, WU Shunpeng, YANG Meng, ZOU Yujian
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: A fusion algorithm named permutation entropy weighted and bias adjustment rule fusion (PEW-BAR) was proposed to enhance the accuracy of speech emotion recognition by exploiting the emotional information in the spectral characteristics of speech signals. The algorithm was based on the integration of wavelet scattering transform and Mel-frequency cepstral coefficients (MFCC). Firstly, wavelet scattering features and MFCC-related features from speech signals were extracted. Then, the wavelet scattering features were expanded in the scale dimension and applied support vector machines to obtain posterior probabilities for emotion recognition. And permutation entropy was calcu-

收稿日期：2023−11−30；**修回日期：**2024−04−10
通信作者：应娜，yingna@hdu.edu.cn
基金项目：浙江省自然科学基金资助项目（No.LTGS23F010001）；浙江省属高校基本科研业务费专项资金资助项目（No.GK239909299001-406）
Foundation Items: The Natural Science Foundation of Zhejiang Province (No.LTGS23F010001), Zhejiang Province Higher Education Institutions' Basic Research Fund Special Project Funding (No.GK239909299001-406)

lated and a weighted fusion based on this entropy was subsequently applied. Finally, a bias adjustment rule was utilized to refine the integration results obtained from the MFCC-related features. Experimental results on various datasets, including EMODB, RAVDESS, and eNTERFACE05, demonstrate notable improvements. The proposed algorithm outperforms traditional wavelet scattering coefficient-based methods, achieving accuracy improvements of 2.82%, 2.85%, and 5.92%, respectively. Additionally, it shows enhancements of 3.40%, 2.87%, and 5.80% in terms of unweighted average recall (UAR), and a 6.89% improvement on the IEMOCAP dataset.

Key words: speech emotion recognition, wavelet scattering transform, permutation entropy, MFCC, model fusion

0　引言

人类的情感可以通过各种信息源来表达和识别，如语音、文字、面部表情、生理信号等，利用两种或多种信息源进行情感识别的方法称为多模态情感识别[1]。在这些信息源中，语音是表达情感最有前景的方法之一，也是实现人机交互的重要基础[2]。因此，语音情感识别（speech emotion recognition，SER）通过计算机实现对人类语音中所蕴含的情感进行感知和理解，但是由于情感的复杂性和主观性，如何为SER提取具有情感特异性的特征仍然是一项具有挑战性的任务。

一直以来，梅尔频率倒谱系数（Mel-frequency cepstral coefficient，MFCC）作为一种模拟人耳听觉特性的声音特征表示方法，被广泛应用于SER研究。随着数据量的增加和计算能力的提高，深度方法在语音信号处理中也展现出强大的性能，例如自动语音识别[3]、说话人确认[4]和语音情感识别。在SER相关研究中，常见的算法输入是原始波形、语谱图和声学特征等。例如，文献[5]通过重构相空间（reconstruction phase space，RPS）将原始语音信号转换为三维张量，输入三维卷积神经网络（convolutional neural network，CNN）结构，证明了三维张量能够有效地捕捉说话人的情感特征。在此基础上，文献[6]使用数据增强和脱落等技术降低训练模型的局部最小值和过拟合风险，实现了情感分类精度的提升。而文献[7]采用迁移学习的策略，把原始波形RPS的三维投影输入预训练的VGG 16模型，并使用灰狼优化算法调整超参数，进一步提高了情感识别的精度。以语谱图作为输入的算法模型越来越成为语音识别算法的一个分支，例如文献[8]结合视觉理论，进一步提取语谱图的颜色、亮度、方向和纹理特征，并根据听觉敏感度对显著图设置权重，送入微调的深层混合网络，相比于基准模型更具情感区分能力；文献[9]采用迁移学习的方法，引入ResNet架构和残差适配器结合的技术，研究了基于语谱图的多语料库SER问题。研究者们也以声学特征作为模型的主要输入进行研究，文献[10]使用并行CNN从MFCC中学习时间和频谱信息，在保证精度的同时简化了模型复杂度；常数Q变换（constant-Q transform，CQT）具有可变的时频分辨率[11]，文献[12]使用多个不同的深度神经网络（deep neural network，DNN）架构作为后端分类器，在4个公开数据集上对比分析了短期梅尔频率系数和基于CQT特征的表现，结果表明常数Q滤波器组具有更高的低频分辨率，从而改善了SER性能。文献[13]分别基于原始波形、语谱图和时—空动态建模一维、二维和三维CNN模型，在分数级上融合多模型，证明多模深度音频特征具有互补性。在每个任务都需要定制化的神经网络架构的同时，小波散射变换（wavelet scattering transform，WST）作为一种新兴的特征提取技术出现，优点在于它在数学上具有严格的理论证明，并且在实际应用中被证明小样本情境下可以获得优于CNN的识别率，因此在生物信号，尤其是声音信号[14]中可以为特定分类器或CNN提供优质输入。

目前，基于WST的应用研究主要有两种方式：直接使用散射系数特征和对散射系数特征进行尺度维度扩展[15]。例如，文献[16]首次将WST应用于SER，该研究在常用数据集上的实验结果表明WST特征比基于MFCC的特征更能捕获情感的相关信息；文献[17]结合散射系数特征和音质特征，基于支持向量机（support vector machine，SVM）分类器验证了两种特征相互补充对语音情感识别的有效性；在声学采集设备分类任务中，基于散射系数的特征输入SVM分类器时，基于WST的均匀权重投票策略相比CNN可以更好地工作[18]；文献[19]同样使用散射系数的特征输入SVM分类器对煤矿微震事件进行识别，实现了数据的降维和精度的提升。由此说明，散射系数特征能够很好地进行信号分类，性能甚至不输于CNN。在进一步提高散射系数特征在多尺度扩展应用中的表达能力上，现有的方法还存在不足之处。例如，文献[15]在心音信号分类任务中，对散射系数使用尺度平均方法，降低了特征维度和分类精度；文献[20]针对心律失常分类，结合尺度维度投票和主成分分析方法来降低特征数量、筛选分类精度较高的尺度特征，结果表明具有更大振幅和波动的尺度特征包含更多更清晰的心跳细节，而均匀权重投票的方法容易受精度较低的尺度特征的影响，忽略了尺度特征的差异性；文献[21]基于多时间尺度小波散射特征，分别使用子频谱混洗法和时间混洗法对声音场景进行分类，结果表明单一时间尺度小波散射特征具有局限性，不能获取所有的声学特性。而已有的研究证实，基于WST特征并组合多个弱分类器的结果以获得比从单个有偏分类器更准确[22-23]的结果。但是，基于小波散射尺度特征差异性的探索局限于特征级，尚未发现分数级的相关文献，融合语音WST特征和其他时频尺度特征具有进一步提升声音信号分类性能的潜力。

综合上述研究，为了减少WST尺度特征差异性及单一时间尺度局限性对SER的影响，并提高情感识别的准确率，本文提出一种基于WST和MFCC的排列熵（permutation entropy，PE）加权和偏差调整规则的语音情感识别融合（permutation entropy weighted and bias adjustment rule fusion，PEW-BAR）算法。该算法使用了两种特征提取方法，分别为基于WST的尺度扩展特征和基于MFCC的相关特征，并使用两个SVM对两类特征分别预测各种情感类别的后验概率，然后根据排列熵对小波散射变换特征的预测概率进行加权，最后使用偏差调整系数来修正基于两类特征的预测后验概率，从而削弱两种特征提取方法之间的偏差，提高情感识别的准确率。

1 WST基本原理

WST是一种复值卷积神经网络[24]，它用小波代替数据驱动滤波器，用复数模运算代替非线性。WST能够从实值时间序列中提取具有低方差、时间平移不变性和扭曲变形稳定性的特征，同时克服了MFCC在长时间尺度信息和高频信息上的缺陷[14]。

小波散射变换以迭代的方式生成特征。首先，采用低通滤波函数对输入信号$x(t)$进行卷积，得到零阶散射系数如下：

$$S_0x(t)=x(t)*\phi_T(t) \tag{1}$$

其中，$x(t)$为输入信号，$\phi_T(t)$表示低通滤波器，T为$\phi_T(t)$的时间窗长，符号“$*$”表示卷积操作。由于低通滤波过程去除了高频，通过小波模变换恢复高频信息。将$x(t)$与一阶小波滤波器组卷积并求复数模，生成一阶尺度图系数如下：

$$x_1(t,\lambda_1)=\left|x(t)*\psi_{\lambda_1}(t)\right| \tag{2}$$

其中，“$|\cdot|$”为取模操作，$\psi_{\lambda_1}(t)$为一阶小波滤波器组，该滤波器组由母小波$\psi(t)$扩展得到，其傅里叶变换$\hat{\psi}(\omega)$集中在无量纲频率区间$\left[1-2^{\frac{1}{2Q}},1+2^{\frac{1}{2Q}}\right]$

上，其中Q是质量因子（每个滤波器组的每倍频程小波滤波器的数量）。通过对母小波进行伸缩，产生一簇以$\lambda_1=2^{j_1+\frac{k}{Q}}$为中心频率的低通滤波器，其中$j_1\in Z$，$k\in(1,2,\cdots,Q)$分别表示每倍频程和色度。为了确保信号的时移不变性，使用低通滤波器对一阶尺度图系数进行平均，获得一阶散射系数如下：

$$S_1x(t,\lambda_1)=x_1(t,\lambda_1)*\phi_T(t)=\left|x(t)*\psi_{\lambda_1}(t)\right|*\phi_T(t) \tag{3}$$

将$x_1(t,\lambda_1)$与$\psi_{\lambda_2}(t)$进行卷积和复数模运算，得到二阶尺度图系数：

$$x_2(t,\lambda_1,\lambda_2)=\left|x_1(t)*\psi_{\lambda_2}(t)\right| \tag{4}$$

通过与式（3）相同的方式，捕获第一层每个频段发生的高频幅度调制的二阶系数由式（5）获得：

$$S_2x(t,\lambda_1,\lambda_2)=x_2(t,\lambda_1,\lambda_2)*\phi_T(t)=\left|\left|x(t)*\psi_{\lambda_1}(t)\right|*\psi_{\lambda_2}(t)\right|*\phi_T(t) \tag{5}$$

其中，$\psi_{\lambda_2}(t)$表示二阶小波滤波器组，λ_2表示其中心频率。散射系数是通过任意阶迭代的操作来定义的，令r表示迭代阶数，对于任意$r>1$，迭代的小波模卷积由式（6）表示：

$$x_r(t,\lambda_1,\cdots,\lambda_r)=\left|\left|\left|x(t)*\psi_{\lambda_1}(t)\right|*\cdots\right|*\psi_{\lambda_r}(t)\right| \tag{6}$$

其中，$\psi_{\lambda_r}(t)$表示r阶小波滤波器组，其中心频率由λ_r表示，倍频程分辨率为Q_r。在$x_r(t,\lambda_1,\cdots,\lambda_r)$上应用低通滤波函数$\phi_T(t)$进行卷积，可得$r$阶散射系数如下：

$$S_rx(t,\lambda_1,\cdots,\lambda_r)=x(t,\lambda_1,\cdots,\lambda_r)*\phi_T(t)=\left|\left|\left|x(t)*\psi_{\lambda_1}(t)\right|*\cdots\right|*\psi_{\lambda_r}(t)\right|*\phi_T(t) \tag{7}$$

2　基于WST的PE加权和偏差调整规则的双特征融合算法

本文提出的PEW-BAR双特征语音情感识别融合算法流程如图1所示。算法首先对原始语音数据进行预处理，接着分为两个支路进行特征提

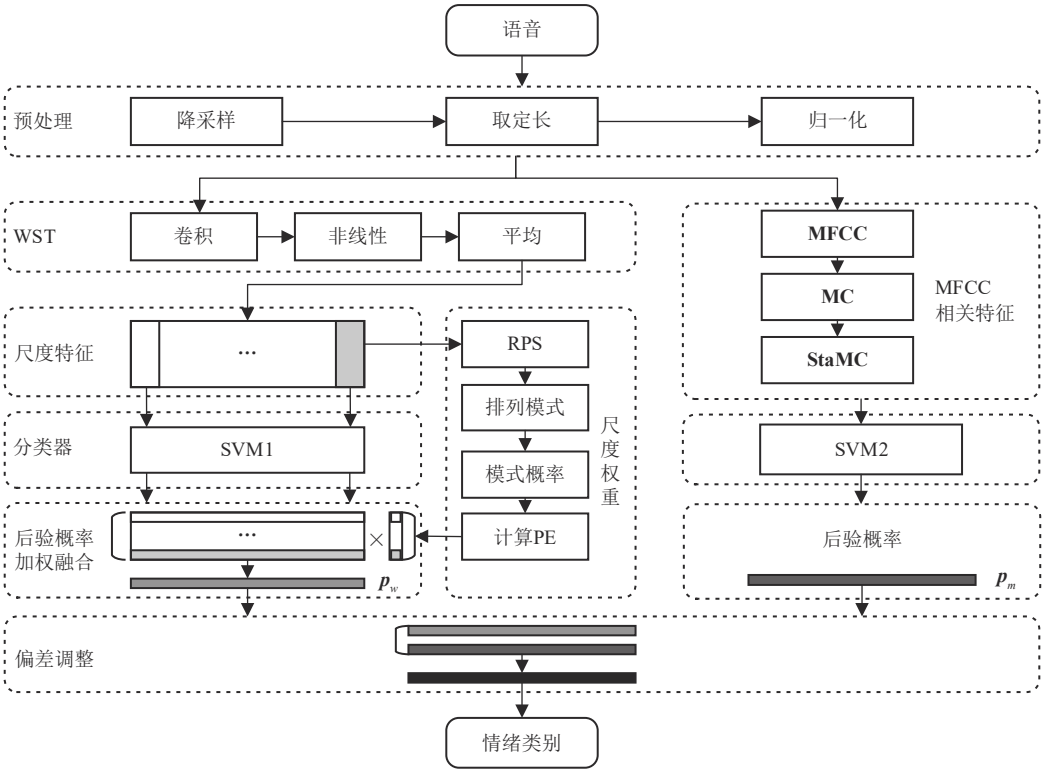

图1　PEW-BAR双特征语音情感识别融合算法流程

取和情感分类：一个支路在获取WST尺度特征的基础上，由尺度特征估计各情感类别的后验概率并获得PE，再使用各尺度下的PE对预测结果进行加权，赋予振幅和波动较大的尺度维度更高权重；另一个支路根据MFCC对人耳听觉特性的模拟性，提取出基于MFCC的特征并得到各情感类别的后验概率，再利用WST和MFCC对情感语音表征的互补性，使用偏差调整规则融合两条支路，以得到更高的情感分类准确率。

2.1 预处理

为了便于并行计算和保持散射变换特征矩阵的统一尺寸，文献[2]采用了以下方法：先把采样频率高于16 kHz的语音文件降采样为16 kHz；再截取每个语音文件的前3 s，对于持续时间小于3 s的语音进行零填充；最后为了消除响度的影响，对波形幅值进行归一化。

2.2 基于WST的尺度扩展特征

由于散射系数的能量随着层数的增加而减少，且前两层包含99%的能量[20]，本文应用两阶散射网络来提取语音信号的小波散射特征。对于给定输入$x(t)$，两阶散射特征矩阵**SC**由零阶、一阶和二阶散射系数共同组成，维度为$N_q \times N_s$的**SC**及其尺度特征可由式（8）表示：

$$\mathbf{SC}=\begin{bmatrix}\boldsymbol{S}_0\\ \boldsymbol{S}_1\\ \boldsymbol{S}_2\end{bmatrix}=\left[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_{N_s}\right] \tag{8}$$

其中，N_q表示散射路径数，N_s表示尺度维度数，$\boldsymbol{S}_0$是由式（1）得到的零阶散射系数，$\boldsymbol{S}_1$是由式（3）得到的一阶散射系数，$\boldsymbol{S}_2$是由式（5）得到的二阶散射系数，$\boldsymbol{f}_n$表示重塑**SC**得到的维度为N_q的列向量，$n=1,2,\cdots,N_s$。两阶WST的特征提取原理如图2所示。

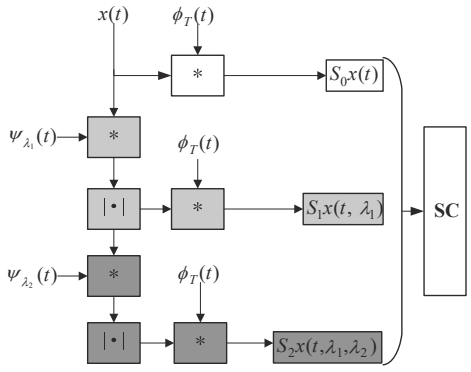

图2 两阶WST特征提取原理

本文根据实验和对文献[17]的参考，设定一阶滤波器组的每倍频程小波数Q_1为4，二阶滤波器组的Q_2为1，散射变换的不变性尺度T为0.75 s。通过这些参数的设置，得到了$N_q=200$、$N_s=24$的小波散射特征矩阵，并按照尺度维度扩展，形成24个长度为200的尺度特征，标记为相同的类别训练SVM1分类器。

2.3 基于MFCC的统计特征

MFCC是对语音的梅尔谱对数进行倒谱分析而生成的。在快速傅里叶变换（fast Fourier transform，FFT）将原始音频数据转换为声谱图的基础上，使用梅尔频率滤波器组模拟人耳听觉特性，最后使用对数尺度，并利用离散余弦变换（discrete cosine transform，DCT）进行分析和计算。MFCC特征提取原理如图3所示。

提取MFCC特征的采样率设置为16 kHz，以窗口长度为30 ms的汉明窗分帧，帧移为10 ms，每帧取13个系数和对数能量，然后合并**MFCC**及其一阶差分Δ**MFCC**和二阶差分ΔΔ**MFCC**，构建一个特征矩阵**MC**如下：

$$\mathbf{MC}=\left[\mathbf{MFCC},\Delta\mathbf{MFCC},\Delta\Delta\mathbf{MFCC}\right] \tag{9}$$

接着，对**MC**中的时序特征计算最小值、最

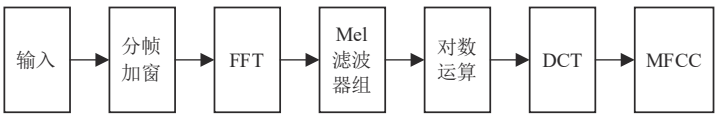

图3 MFCC特征提取原理

大值、均值、中值和标准差5类统计值作为基于MFCC的统计特征，最终获得一个训练SVM2长度为210维的特征向量**StaMC**如下：

$$\mathbf{StaMC}=\left[\overline{\mathbf{MFCC}},\overline{\Delta\mathbf{MFCC}},\overline{\Delta\Delta\mathbf{MFCC}}\right] \tag{10}$$

其中，$\overline{\mathbf{MFCC}}$、$\overline{\Delta\mathbf{MFCC}}$和$\overline{\Delta\Delta\mathbf{MFCC}}$分别表示**MFCC**、$\Delta$**MFCC**和$\Delta\Delta$**MFCC**的5个统计值构成的特征向量。

2.4　基于WST的PE加权算法

排列熵基于输入的有序模式来计算序列的随机性，它是一种非参数方法，能够鲁棒地估计输入情感语音特征的不规则性信息。考虑到尺度维度均匀权重的投票方法，对不同尺度特征赋予相同权重，忽略了各尺度维度特征之间的差异，本文算法提出基于小波散射的PE加权算法。首先，由式（8）中**SC**扩展得到尺度特征$\boldsymbol{f}_n$作为计算排列熵的输入，然后获得尺度特征的RPS，RPS基于时延和嵌入维度表示，根据经验分别取1和3；接着，将每个嵌入向量按升序排列，得到相应的排列模式；最后计算每个排列模式出现的概率。算法具体步骤如下。

步骤1　将式（8）由**SC**按尺度维度重塑得到的N_s个N_q维特征向量$\boldsymbol{f}_n$，通过式（11）计算尺度特征的排列熵：

$$\mathrm{PE}\left(\boldsymbol{f}_n\right)=-\sum_{i=1}^{d!}p\left(\pi_i\right)\mathrm{lb}\,p\left(\pi_i\right) \tag{11}$$

其中，d表示尺度特征的RPS嵌入维度，π_i表示$\boldsymbol{f}_n$的RPS的模式，$p(\cdot)$表示概率。由于$\mathrm{PE}\in[0,\mathrm{lb}\,d!]$，可以得到归一化排列熵：

$$\mathrm{PE}^{\mathrm{norm}}\left(\boldsymbol{f}_n\right)=\frac{1}{\mathrm{lb}\,d!}\mathrm{PE}\left(\boldsymbol{f}_n\right)=-\frac{1}{\mathrm{lb}\,d!}\sum_{i=1}^{d!}p\left(\pi_i\right)\mathrm{lb}\,p\left(\pi_i\right) \tag{12}$$

计算所得的PE值作为尺度权重，用$\boldsymbol{w}=(w_1,w_2,\cdots,w_{N_\mathrm{s}})$来表示。

步骤2　使用SVM1分类器分别估计各尺度特征$\boldsymbol{f}_n$的情感类别后验概率，由$\boldsymbol{p}_n=(p_{n1},p_{n2},\cdots,p_{n\mathrm{C}})$表示，其中，$C$表示情感类别数。后验概率采用二次优化算法进行估计，该方法通过求解一个二次优化问题，找到最佳概率分布。由步骤1获取的尺度权重$\boldsymbol{w}$对$\boldsymbol{p}_n$进行加权，得到尺度加权预测概率$\dot{\boldsymbol{p}}_n=\left(\dot{p}_{n1},\dot{p}_{n2},\cdots,\dot{p}_{nC}\right)$由式（13）计算：

$$\dot{p}_{nl}=w_n\times p_{nl} \tag{13}$$

其中$l=1,2,\cdots,C$。

步骤3　计算步骤2得到的各维度$\dot{\boldsymbol{p}}_n$之和，得到加权预测概率$\boldsymbol{p}_w=(p_{w1},p_{w2},\cdots,p_{wC})$由式（14）计算：

$$p_{wl}=\sum_{n=1}^{N_t}\dot{p}_{nl} \tag{14}$$

为方便模型融合，对$\boldsymbol{p}_w$归一化，得到归一化概率$\boldsymbol{p}_w^{\mathrm{norm}}=(p_{w1}^{\mathrm{norm}},p_{w2}^{\mathrm{norm}},\cdots,p_{w\mathrm{C}}^{\mathrm{norm}})$由式（15）计算：

$$p_{wl}^{\mathrm{norm}}=\frac{p_{wl}}{\sum_{l=1}^{C}p_{wl}} \tag{15}$$

2.5　基于偏差调整规则的融合判决

基于偏差调整规则的决策是由Pham等[22]提出的，用于将两个分类器的输出结合起来，以得到最终的分类。考虑到本文两条支路单独预测时，对不同的情感识别准确率差异显著，本文算法进一步提出基于偏差调整规则的融合策略，在基于WST尺度特征输入SVM1并由PE加权的基础上，融合基于MFCC相关特征输入SVM2分类器得到的后验概率。

由SVM2基于MFCC的相关特征**StaMC**得到的各类别后验概率，用$\boldsymbol{p}_m=(p_{m1},p_{m2},\cdots,p_{m\mathrm{C}})$表示，最终情感类别由$\boldsymbol{p}_w^{\mathrm{norm}}$和$\boldsymbol{p}_m$按如下规则进行判定。

规则1：当$\boldsymbol{p}_w^{\mathrm{norm}}\mapsto\mathrm{C}_1$，且$\boldsymbol{p}_m\mapsto\mathrm{C}_1$时，将输入划分为$\mathrm{C}_1$类。

规则2：当$\boldsymbol{p}_w^{\mathrm{norm}}\mapsto\mathrm{C}_2$，且$\boldsymbol{p}_m\mapsto\mathrm{C}_2$时，将输入划分为$\mathrm{C}_2$类。

规则3：当 $\boldsymbol{p}_w^{\text{norm}} \mapsto \mathrm{C}_1$，而 $\boldsymbol{p}_m \mapsto \mathrm{C}_2$ 时，若 $p\left(\boldsymbol{p}_{\mathrm{w}}^{\text{norm}} \mapsto \mathrm{C}_1\right) \geqslant p\left(\boldsymbol{p}_{\mathrm{m}} \mapsto \mathrm{C}_2\right) - \mathrm{th}$，将输入划分为 C_1 类，否则，划分为 C_2 类。

规则4：当 $\boldsymbol{p}_w^{\text{norm}} \mapsto \mathrm{C}_2$，而 $\boldsymbol{p}_m \mapsto \mathrm{C}_1$ 时，若 $p\left(\boldsymbol{p}_{\mathrm{w}}^{\text{norm}} \mapsto \mathrm{C}_2\right) \geqslant p\left(\boldsymbol{p}_{\mathrm{m}} \mapsto \mathrm{C}_1\right) - \mathrm{th}$，将输入划分为 C_2 类，否则，划分为 C_1 类。

其中，符号“$\mapsto$”表示按最大概率值分类，C_1 和 C_2 为预测所属类别，th为设定的偏差调整系数。根据不同数据库的多次试验，本文th值分别在EMODB中设置为0.37，在RAVDESS中设置为0.34，在IEMOCAP中设置为0.1，在eNTERFACE05中设置为0.09。

3 实验与结果

3.1 数据集

本文在4种公开[12, 25]的语音情感数据集上进行了实验，它们分别是EMODB、RAVDESS、IEMOCAP和eNTERFACE05。EMODB是一个德语情感数据集，由535条表达7种不同情感的录音样本组成，样本以16 kHz的采样率、16 bit的采样精度保存为wav格式。RAVDESS是一个多模态数据集，涵盖8种不同的情感类别，录音样本以48 kHz的采样率、16 bit的采样精度保存为wav格式。IEMOCAO是一个多模态对话数据集，样本以16 kHz的采样率、16 bit的采样精度保存为wav格式。eNTERFACE05也是一个多模态数据库，包含了6种不同情绪的录音样本，以48 kHz的音频采样频率保存在avi格式文件中。为了更好地与现有的SER文献进行比较，本文使用的SER数据集见表1。

3.2 评估方法

为便于与相关文献结果进行对比，本文实验采用遗漏一个说话人（leave one speaker out，LOSO）的交叉验证策略，同时，使用准确率（accuracy，ACC）和未加权平均召回率（unweighted average recall，UAR）两个指标来评估模型性能，准确率通过正确分类的语音数量与测试集中的语音总数之间的比值得到，而UAR度量[25]由式（16）给出：

$$\mathrm{UAR} = \frac{1}{C}\sum_{i=1}^{C}\frac{\boldsymbol{A}_{ii}}{\sum_{j=1}^{C}\boldsymbol{A}_{ij}} \tag{16}$$

其中，$\boldsymbol{A}$ 为混淆矩阵，$\boldsymbol{A}_{ij}$ 指类别为 i 的样本分类为类别 j 的数量。

3.3 结果与分析

首先，在EMODB、RAVDESS、IEMOCAP和eNTERFACE05数据集上，以MFCC和小波散射网络（scattering network，ScatNet）作为基线[16]模型，与本文提出的改进算法进行对比，MFCC、ScatNet和PEW-BAR在不同数据集的性能比较见表2。

从表2可以看出，在不同的数据集上基于WST特征的模型都明显优于基于MFCC特征的模型，这体现了WST特征对语音情感的强表达能力。此外，本文所提算法在各个实验数据集上都

表1 使用的SER数据集

数据集	说话人	情感类别	样本数/个	语种
EMODB	5女 5男	愤怒、悲伤、无聊、恐惧、快乐、厌恶和中性	535	德语
RAVDESS	12女 12男	平静、快乐、悲伤、愤怒、中性、恐惧、惊讶和厌恶	1 440	英语
IEMOCAP	5女 5男	快乐、悲伤、愤怒和中立	4 936	英语
eNTERFACE05	8女 36男	快乐、悲伤、愤怒、恐惧、惊讶和厌恶	1 293	英语

表2　MFCC、ScatNet和PEW-BAR在不同数据集的性能比较

数据集	MFCC		ScatNet		PEW-BAR	
	ACC	UAR	ACC	UAR	ACC	UAR
EMODB	58.39	54.03	74.40	71.30	77.22	74.70
RAVDESS	36.74	34.77	50.00	48.50	52.85	51.37
IEMOCAP	55.54	47.19	60.41	50.40	59.72	57.29
eNTERFACE05	42.31	42.20	52.66	52.73	58.58	58.53

取得了最佳UAR，且改进效果明显；而在ACC指标上的表现，在IEMOCAP数据集上比ScatNet网络低0.69，这是ScatNet通过实验调整信号长度以及WST参数而得到的结果。因此，表2的结果证明了本文算法的有效性。

为了进一步分析本文算法在各情感类别上的特异性，首先给出了在EMODB和RAVDESS数据集上本文模型所得到的混淆矩阵。这里定义愤怒（Ang）、无聊（Bor）、厌恶（Dis）、恐惧（Fea）、快乐（Hap）、中性（Neu）、悲伤（Sad）和惊讶（Sur）。

本文算法在真实情感类别与预测情感类别的混淆矩阵如图4所示，展示了每个情感类别中有多少被准确或错误地预测出来。正确预测的情绪在混淆矩阵中对角表示，错误预测的情绪在每个类的相应行中表示。从图4可以观察到愤怒类的准确率最高，为93.70%，而快乐类最低，只有54.29%；且愤怒类易与快乐类相互混淆，无聊类易与中性类相互混淆。错误分类最多的是快乐、愤怒情绪，因为它们具有相似的唤醒但相反的效价特征。这一观察结果与相关文献[25]是一致的。本文算法在RAVDESS数据集上的混淆矩阵如图5所示。由图5可以看出，愤怒类的准确率最高，为72.40%，而中性类的准确率最低，为27.08%。可以观察到中性类易被误分类为悲伤类，平静类与悲伤类相互混淆。此外，快乐类的准确率也相对较低，且其易误分类为愤怒、恐惧和惊讶等高唤醒类。

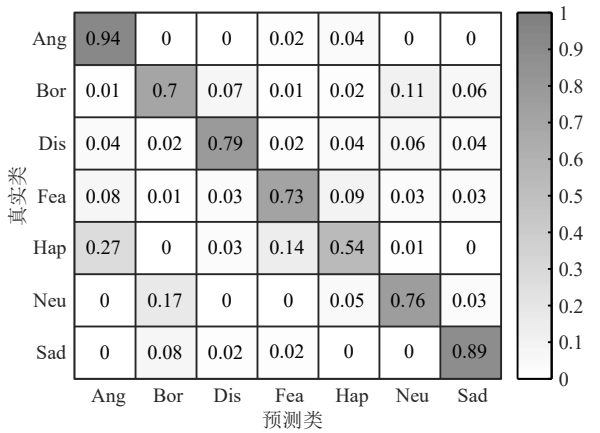

图4　本文算法在EMODB数据集上的混淆矩阵

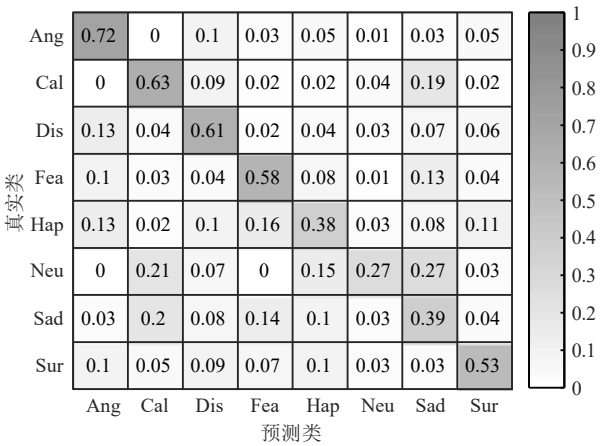

图5　本文算法在RAVDESS数据集上的混淆矩阵

将本文算法获得的结果与基于深度学习模型的SER相关工作进行了比较，EMODB和RAVDESS相关文献对比见表3。在SER文献中，评估系统性能的策略不同，使得SER算法的直接比较变得困难。因此，本文主要选用文献[25]选择的算法结果作对比，这些结果也是采用LOSO实验方案以及与本文相同的性能评估方法得到的。

表3　EMODB和RAVDESS相关文献对比

对比文献	特征	结构	EMODB		RAVDESS	
			ACC	UAR	ACC	UAR
文献[26]（2021年）	基于调制谱	DNN	55.40%	46.83%	37.77%	29.10%
文献[9]（2021年）	基于梅尔谱图	ResNet	76.34%	71.07%	52.56%	49.46%
文献[10]（2022年）	MFCC	并行CNN	73.37%	67.21%	48.68%	44.51%
文献[27]（2022年）	Interspeech 2009	Bi-LSTM+注意力	53.45%	47.49%	24.58%	21.79%
文献[12]（2022年）	基于CQT	Conv 2D-LSTM	65.69%	60.45%	46.94%	44.15%
文献[25]（2023年）	CQT-MSF	DNN-SVM	79.86%	76.17%	52.24%	48.83%
本文算法	WST、MFCC	PEW-BAR	77.22%	74.70%	52.78%	51.30%

由表3可以看出，本文算法在EMODB数据集和RAVDESS数据集上都可以得到较高的准确率。其中，在RAVDESS数据集上的实验取得了最佳结果，文献[25]在EMODB数据集取得略高于本文方法的结果，主要得益于使用数据增强技术增加了用于处理的数据量。可见相比深度学习方法，本文提出的SER算法依然有竞争力。

本文算法在IEMOCAP数据集上所得到的混淆矩阵如图6所示。由图6可以看出，在IEMOCAP中，中性类的准确率最高，为70.84%，而激动类的准确率最低，为27.95%，且其易误分类为中性类，而中性类和与悲伤类相互混淆。此外，愤怒类也易误分类为中性类。本文算法在eNTERFACE05数据集上所得到的混淆矩阵如图7所示。由图7可以看出，在eNTERFACE05中，愤怒类的准确率最高，为74.54%，而恐惧类和惊讶类的准确率最低，分别为44.44%和44.91%。厌恶类的准确率也相对较低，且其易误分类为快乐类和悲伤类。此外，可以观察到恐惧类与悲伤类相互混淆，快乐类与厌恶类相互混淆。

IEMOCAP和eNTERFACE05数据集上的相关模型对比见表4。对比文献采用与本文相同的实验策略。

分析表4可知，本文算法在IEMOCAP和eNTERFACE05两个数据集上的UAR指标，都显著高于对比文献，分别比最好的结果高出7.35%和3.73%。此外，本文算法在IEMOCAP上的ACC指标比最好结果高出2.41%，在eNTERFACE05数据集上的ACC稍弱于对比结果，这是由于各情绪类别数量不平衡。

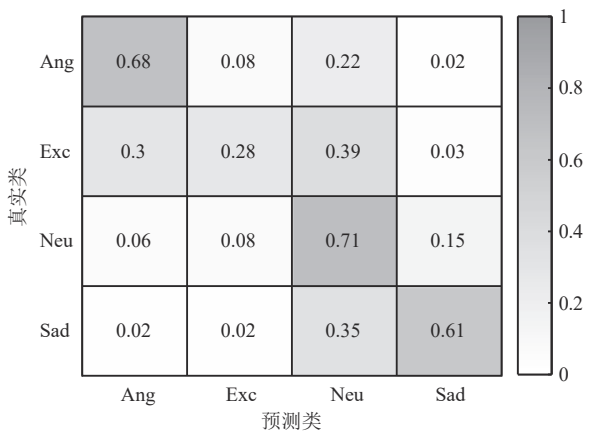

图6　本文算法在IEMOCAP数据集上的混淆矩阵

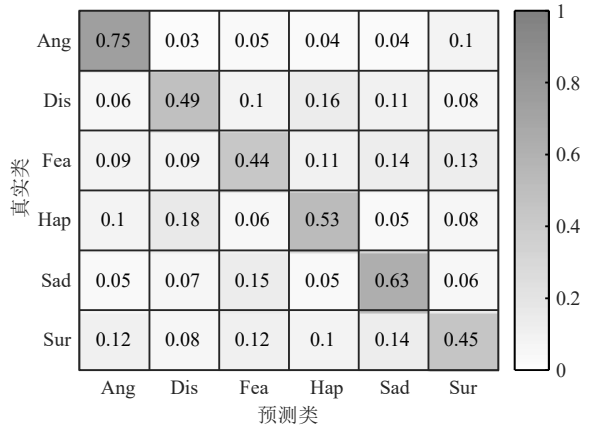

图7　本文算法在eNTERFACE05数据集上的混淆矩阵

表4　IEMOCAP和eNTERFACE05数据集上的相关模型对比

对比模型	特征	结构	IEMOCAP		eNTERFACE05	
			ACC	UAR	ACC	UAR
文献[12]（2022年）	CQT	Conv2D	53.08%	45.04%	46.77%	41.31%
文献[12]（2022年）	MFSC	LSTM+注意力	45.68%	40.41%	40.02%	39.16%
文献[12]（2022年）	CWT	Conv2D-LSTM	57.31%	49.94%	59.05%	54.80%
本文算法	WST、MFCC	PEW-BAR	59.72%	57.29%	58.58%	58.53%

4　结束语

本文提出了一种基于小波散射和MFCC的双特征语音情感识别融合算法，该算法通过排列熵加权和偏差调整规则，有效地解决了WST尺度特征差异性和单一时间尺度局限性的问题，提高了情感识别的准确率。本文在EMODB、RAVDESS、IEMOCAP和eNTERFACE05 4个公开数据集上进行了实验验证，结果表明，本文算法能够充分利用WST尺度特征的情感信息不平衡性，同时发挥了WST特征和MFCC特征在表达语音情感信息方面的互补性，提升了语音情感识别性能。

参考文献：

[1] FAHAD M S, RANJAN A, YADAV J, et al. A survey of speech emotion recognition in natural environment[J]. Digital Signal Processing, 2021(110): 102951.

[2] SUN C, LI H, MA L. Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network[J]. Frontiers in Psychology, 2023(13): 1075624.

[3] 王海坤, 潘嘉, 刘聪. 语音识别技术的研究进展与展望[J]. 电信科学, 2018, 34(2): 1-11.
WANG H K, PAN J, LIU C. Research development and forecast of automatic speech recognition technologies[J]. Telecommunications Science, 2018, 34(2): 1-11.

[4] 杨震, 王天朗, 郭海燕, 等. 跨域注意力特征融合的说话人确认方法[J]. 通信学报, 2023, 44(8): 89-98.
YANG Z, WANG T L, GUO H Y, et al. Speaker verification method based on cross-domain attentive feature fusion[J]. Journal on Communications, 2023, 44(8): 89-98.

[5] FALAHZADEH M R, FARSA E Z, HARIMI A, et al. 3D convolutional neural network for speech emotion recognition with its realization on intel CPU and NVIDIA GPU[J]. IEEE Access, 2022(10): 112460-112471.

[6] FALAHZADEH M R, FAROKHI F, HARIMI A, et al. A 3D tensor representation of speech and 3D convolutional neural network for emotion recognition[J]. Circuits, Systems, and Signal Processing, 2023, 42(7): 4271-4291.

[7] FALAHZADEH M R, FAROKHI F, HARIMI A, et al. Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition[J]. Circuits, Systems, and Signal Processing, 2023, 42(1): 449-492.

[8] 王金华, 应娜, 朱辰都, 等. 基于语谱图提取深度空间注意特征的语音情感识别算法[J]. 电信科学, 2019, 35(7): 100-108.
WANG J H, YING N, ZHU C D, et al. Speech emotion recognition algorithm based on spectrogram feature extraction of deep space attention feature[J]. Telecommunications Science, 2019, 35(7): 100-108.

[9] GERCZUK M, AMIRIPARIAN S, OTTL S, et al. EmoNet: A transfer learning framework for multi-corpus speech emotion recognition[J]. IEEE Transactions on Affective Computing, 2023,14(2): 1472-1487.

[10] AFTAB A, MORSALI A, GHAEMMAGHAMI S, et al. LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition[C]//Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6912-6916.

[11] 徐嘉, 简志华, 金宏辉, 等. 采用恒Q调制包络的合成语音伪装检测方法[J]. 电信科学, 2023, 39(11): 107-115.
XU J, JIAN Z H, JIN H H, et al. A method of synthetic speech spoofing detection using constant Q modulation envelope[J]. Telecommunications Science, 2023, 39(11): 107-115.

[12] SINGH P, WALDEKAR S, SAHIDULLAH M, et al. Analysis of constant-Q filterbank based representations for speech emotion recognition[J]. Digital Signal Processing, 2022(130): 103712.

[13] ZHANG S Q, TAO X, CHUANG Y L, et al. Learning deep multimodal affective features for spontaneous speech emotion recognition[J]. Speech Communication, 2021(127): 73-81.

[14] BRUNI V, CARDINALI M L, VITULANO D. An MDL-based wavelet scattering features selection for signal classification[J]. Axioms, 2022, 11(8): 376.

[15] MEI N, WANG H, ZHANG Y, et al. Classification of heart sounds based on quality assessment and wavelet scattering transform[J]. Computers in Biology and Medicine, 2021(137): 104814.

[16] SINGH P, SAHA G, SAHIDULLAH M. Deep scattering network for speech emotion recognition[C]//Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO). Piscataway: IEEE Press, 2021: 131-135.

[17] 孙聪珊, 马琳, 李海峰. 基于CM-OMEMD和小波散射网络的语音情感识别[J]. 信号处理, 2023, 39(4): 688-697.
SUN C S, MA L, LI H F. Speech emotion recognition based on CM-OMEMD and wavelet scattering network[J]. Journal of Signal Processing, 2023, 39(4): 688-697.

[18] CHIN C S, ZHANG J H. Wavelet scattering transform for multiclass support vector machines in audio devices classification system[C]//Proceedings of the 2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). Piscataway: IEEE Press, 2021: 735-740.

[19] 樊鑫, 赵晓光, 唐胜利, 等. WSD-SVM在工作面底板破坏深度微震事件自动识别中的应用[J]. 西安科技大学学报, 2023, 43(1): 160-166.
FAN X, ZHAO X G, TANG S L, et al. Application of WSD-SVM in micro-seismic events automatic recognition of the damage depth of working face floor[J]. Journal of Xi'an University of Science and Technology, 2023, 43(1): 160-166.

[20] LIU Z S, YAO G H, ZHANG Q, et al. Wavelet scattering transform for ECG beat classification[J]. Computational and Mathematical Methods in Medicine, 2020: 3215681.

[21] KEK X Y, CHIN C S, LI Y. An intelligent low-complexity computing interleaving wavelet scattering based mobile shuffling network for acoustic scene classification[J]. IEEE Access, 2022(10): 82185-82201.

[22] PHAM T D. Classification of motor-imagery tasks using a large EEG dataset by fusing classifiers learning on wavelet-scattering features[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023(31): 1097-1107.

[23] HAJIHASHEMI V, GHARAHBAGH A A, CRUZ P M, et al. Binaural acoustic scene classification using wavelet scattering, parallel ensemble classifiers and nonlinear fusion[J]. Sensors, 2022, 22(4): 1535.

[24] AL-TIMEMY A H, SERRESTOU Y, KHUSHABA R N, et al. Hand gesture recognition with acoustic myography and wavelet scattering transform[J]. IEEE Access, 2022(10): 107526-107535.

[25] SINGH P, SAHIDULLAH M, SAHA G. Modulation spectral features for speech emotion recognition using deep neural networks[J]. Speech Communication, 2023(146): 53-69.

[26] AVILA A R, AKHTAR Z, SANTOS J F, et al. Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild[J]. IEEE Transactions on Affective Computing, 2021, 12(1): 177-188.

[27] LIU Y, SUN H, GUAN W, et al. Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework[J]. Speech Communication, 2022(139): 1-9.

[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。

吴顺朋（1994– ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为语音信号处理。

杨萌（1980– ），男，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SAR图像目标检测与识别及大模型应用。

邹雨鉴（1998– ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为语音信号处理。