文章编号：2095-2163(2024)01-0085-11 中图分类号：TP241 文献标志码：A

# 基于多任务学习与注意力机制的多层次音频特征情感识别研究

李 磊[1,2]，朱永同[1,2]，杨 琦[1,2,3]，赵金葳[4]，马 柯[1]
（1 上海理工大学 健康科学与工程学院，上海 200093；2 上海理工大学 机器智能研究院，上海 200093；
3 上海理工大学 机械工程学院，上海 200093；4 商丘学院 机械与电气信息学院，河南 商丘 476000）

**摘 要**：传统音频分类任务仅仅是从单层次音频提取特征向量进行分类，即便使用过大的模型，其过多的参数也会造成特征之间的耦合，不符合特征提取“高聚类，低耦合”的原则。由于注意到一些与情绪相关的协变量并没有得到充分利用，本文在模型中加入性别先验知识；将多层次音频特征分类问题转化为多任务问题进行处理，从而对多层次特征进行解耦再进行分类；针对特征分布的再优化方面设计了一个中心损失模块。通过在 IEMOCAP 数据集上的实验结果表明，本文提出模型的加权精度（WA）和未加权精度（UA）分别达到了 71.94%和 73.37%，与原本的多层次模型相比，WA 和 UA 分别提升了 1.38%和 2.35%。此外，还根据 Nlinear 和 Dlinear 算法设计了两个单层次音频特征提取器，在单层次音频特征分类实验中取得了较好的结果。

**关键词**：语音情感分类；MFCC；中心损失；多任务学习；先验信息；Dlinear

# Multilevel emotion recognition of audio features based on multitask learning and attention mechanism

LI Lei[1,2], ZHU Yongtong[1,2], YANG Qi[1,2,3], ZHAO Jinwei[4], MA Ke[1]
(1 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai 200093, China;
3 School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
4 School of Mechanical and Electrical Information, Shangqiu University, Shangqiu Henan 476000, China)

**Abstract**: The traditional audio classification tasks involve extracting feature vectors from single-level audio, which can result in coupling between features even when using large models due to excessive parameters, violating the principle of high cohesion and low coupling in feature extraction. We observe that some emotion-related covariates are not fully utilized. Therefore, we incorporate gender-prior knowledge into the model. Furthermore, we transform the multilevel audio feature classification problem into a multi-task problem, thereby decoupling and classifying multilevel features separately. Finally, we introduce center loss for further optimization of feature distribution. Experimental results on the IEMOCAP dataset demonstrate that the proposed model achieves a weighted accuracy (WA) of 71.94% and an unweighted accuracy (UA) of 73.37%, which are improved by 1.38% and 2.35% respectively compared to the original multilevel model. In addition, we have designed two single-level audio feature extractors based on the Nlinear and Dlinear algorithms, which have yielded promising results in single-level audio feature classification experiments.

**Key words**: Emotion Recognition; MFCC; Center loss; Multi-task Learning; Prior Information; Dlinear

## 0 引 言

情感是一种综合了人类行为、思想和感觉的现象，而情感识别作为自然语言处理领域的重要子领域，旨在通过机器学习分析和理解人类的各种情绪。情感识别的研究意义广泛，涵盖了从商业应用到医疗保健、教育和社会科学等多个领域。通过深入了解和分析人类情感，可以更好地理解人类行为、需求和社会动态，并开发出更具人性化和智能化的技术和服务。在具体应用上，直接对大量相关数据进行

作者简介：李 磊（1997-），男，硕士研究生，主要研究方向：语音情绪识别；朱永同（1998-），男，博士研究生，主要研究方向：点云配准、深度学习；杨 琦（1998-），男，硕士研究生，主要研究方向：语音情绪识别；赵金葳（2003-），男，本科生，主要研究方向：机器学习。
通讯作者：马 柯（1975-），男，博士，主任医师，主要研究方向：疼痛学。Email：3142272293@qq.com

收稿日期：2023-10-08

情感分析可以实现社会舆情分析、情感健康和心理疾病诊断等功能。特别是在人机交互领域,通过交互时的情感分析,可以设计出更加智能的交互系统[1]。

由于语音信息相对于面部表情、身体姿势等媒介具有获取简单方便且信息丰富等特点,所以语音情感识别是重要的情感识别方法之一。语音情感识别流程一般由 3 部分组成,分别是音频信号采集、音频特征提取和情感分类。情感识别技术的关键在于音频特征提取和特征分类两部分[2]。

原始的音频信号包含着丰富的特征,通过傅里叶变换、滤波等一系列操作,可以从时域信号中获取 MFCC、功率谱等一系列时域和频域上的音频特征。为了满足情感识别任务的需求,需要设计更加有效的特征提取器来提取特征。特征提取器主要有经典的卷积神经网络(CNN)[3],Alexnet[4-5] 等。鉴于语音信息的时序性,多用循环神经网络(RNN)[6] 的一系列衍生结构(如:长短期记忆网络[7-8]、门控循环单元[9] 等)和基于 transformer[10] 的神经网络(如:Bert[11]、对比预测编码[12] 等)进行特征提取。当前解决时序问题最常用的模型是基于 attention 机制的 transformer 系列算法,但是也出现一些新的无 attention 机制的算法(如:Dlinear[13] 和 TiDEs[14]),其在 Traffic 和 Electricity 等时序数据集[15] 上取得了更好的预测结果,本文设计了基于 Dlinear 和 Nlinear 算法的特征提取器,提取单层次语音信息并做情感分类。

相对于单层次语音信号,多层次语音信息可以更方便准确地提取出多样化的特征,从而增强模型的泛化能力。频谱图包含大部分语音频域信息,MFCC 主要包含可被人类感知的音频频域信息[16],而多层次语音情绪识别则一般会从频谱图、MFCC 和语音时域信号中提取特征再进行情感分类[17]。

鉴于声道等生理构造的性别差异,本文改进了原有的多层次音频表征形式。通过加入性别先验知识,将原有模型演化为多任务学习[18] 的类型,从而降低原有模型之间的耦合性。此外,在特征分类方面,设计了一个中心损失模块[19],以优化原有特征在类内的距离,并增加类间的可区分性,从而提高了情感识别的鲁棒性和准确率。使用多层次音频特征融合往往需要更深层的神经网络,这将会增大模型的梯度弥散,因此本文提出将多层次特征分类任务转化为多任务学习架构[18],通过对不同层次特征进行优化,防止多层次特征之间存在耦合带来的梯度弥散。

本文综合分析考虑情绪识别任务的模型复杂度,在使用多任务学习架构的基础上进行训练,并针对难分类情绪类别的耦合设计了一个中心损失模块,以此来增加类间间隔。此外,添加了现有模型中所忽略的性别信息,充分指导模型学习样本数据的特征信息,且通过实验表明了本文方法的有效性。

## 1 相关工作

语音情感识别的关键在于音频特征提取和情感分类。对于音频特征方面,本文对多层次音频特征进行了特征融合;在情感分类方面,为了增强模型泛化能力,引入了多任务学习[18] 和中心损失模块[19],对特征分布进行调整。

### 1.1 多层次音频特征提取

音频信号的描述形式主要分为时域和频域两种。原始音频时域信号包含信息全面,但是过量冗余的信息增大了处理难度。一般的音频信号采样频率为 48 kHz,但是基于 transformer 的 bert 系列算法只能在序列长度为 128 的时序信号取得较好的效果[11]。因此,一般在音频特征处理中,常使用卷积神经网络等方式对原始音频信号进行降采样,再对得到的较短序列采用基于 BERT 系列的 wav2vec[20] 算法进行关键特征的提取。wav2vec 是一种自监督学习算法,其结合了 transformer 和随机掩码的方法来捕捉音频时域信号的特征,并通过对比学习[21] 来区分潜在的有效特征和无关特征。该方法在音频处理领域具有较高的效果和应用价值。因此,本文选择 wav2vec 作为特征提取器的基础模型。

频域特征是通过对时域信号进行一系列变换得到的。在音频处理中,适当地进行分帧,可以在保留关键信息的同时缩短序列长度以及获得音频的频谱图。频谱图也可以作为图像进行一系列卷积处理。Alexnet 是一个经典且有效的深度卷积神经网络框架,在图像分类任务中有广泛的应用和较高的准确度[4],用 Alexnet 可以提取频谱图特征。MFCC 是进行加窗滤波处理后的特殊频谱图,MFCC 指定的窗口使得其更加符合人的听觉特征[22]。本研究希望在提取特征的同时包含时序信息,所以使用 LSTM 系列算法。BiLSTM[23] 是基于长短期神经网络的一种算法,主要由前向 LSTM 和后向 LSTM 构成,BiLSTM 不仅可以像 LSTM 一样捕捉长时序信息,同时也关注了上下文信息。因此,本文采用 BiLSTM 提取 MFCC 特征。

在音频情感的特征表征中，除了时域特征和频域特征之外，从生理学角度来看，不同性别在表现相同情感时，会因声带等生理结构的差异而产生不同的基音频率、音素、音节等音频特征[24]，这些生理差异也包含在情感分类的相关信息中，并对情感的表达产生影响。在情感分类中，当情感较为激烈时（如：愤怒），基音频率往往会较高，而女性的基音频率通常也较高，因此增大了情感分类的难度。为了解决这个问题，本文在多层次特征表示中，将性别信息进行编码并融合到语音特征中。通过对实验结果的分析，这种策略显著提高了情感分类的精度。

本文所使用的特征提取器中，基于 transformer 的 wav2vec 提取的单层次信息可以获得比其它两个提取器更好的结果。当前出现了一系列 LTSF－Linear 的模型（包括 Dlinear、Nlinear 和 TiDE）在一些时序问题上取得了比 transformer 系列算法更好的效果。其基本原理是通过更深和更宽的线性层以及残差网络建立对样本进行编码然后解码预测。本文分别使用 Dlinear、Nlinear 编码器作为特征提取器进行单层次语音情感分类任务，并与其它提取器和多层次语音分类形成对比。

### 1.2　多层次音频特征融合

多层次信息融合是为了更充分地利用音频信号中不同层次的特征信息，从而提高分类性能。通过将不同层次的特征进行组合，可以获得更加丰富和多样化的表示，有助于更准确地区分不同情感类别。在融合方法方面，可以使用简单的求和或拼接操作来整合多个层次的特征，也可以采用更复杂的交叉注意力机制，使模型能够自动学习不同特征之间的关联性。

通过求和（add）融合特征，可以使描述语音图谱特征下的信息量增多，但是图谱的维度本身并没有增加，只是每一维下的信息量在增加。而通过拼接（concatenate）融合特征，则是通道数的合并，其描述的图谱本身特征数（通道数）增加了，而每一特征下的信息量并没有增加。因此可以认为，特征求和是拼接的一种特殊形式[25]。虽然前者计算量较少，但为了描述多层次特征之间的复杂关系而选择用拼接的方式融合多层次特征。

在时序问题中，特征的重要性可能随着时间的推移而变化。注意力机制能够自适应地学习每个时刻输入特征的权重，使得模型能够更加关注对预测结果有较大贡献的时刻信息。通过对关键信息进行加权，注意力机制提供了更强大的建模能力，使得模型能够更好地捕捉序列数据中的长期依赖和重要特征。在图像识别和 NLP 的机器翻译领域注意力机制获得了广泛应用[26]。

因此，本文通过拼接的方式将 MFCC 特征和频谱图特征进行融合后，通过注意力机制计算拼接特征和时域特征序列之间的相似度，根据相似度对时域特征进行加权，从而将多层次特征融合。

### 1.3　特征分布调整

由于模型参数过大等原因，本文提取的单层次特征以及融合的多层次特征出现了类间分布不规律和特征间耦合等现象。针对于此，采用多任务学习和中心损失模块进行了特征调整。

多任务学习是相对于单任务学习的一种概念。在多任务学习中，将多个相关任务共享一个模型，通过共享信息，相互补充，以提升彼此的表现[18]。本文在多层次语音分类任务基础上，增加了每个层次信息特征向量作为单独的分类任务。同时，将单层次分类任务和多层次分类任务共享特征提取模块，其目的是让单层次信息特征更多地与分类任务相关联，从而提高整体结构的分类准确度。

中心损失（center loss）是一种在分类任务中的损失函数，其基本假设是每个类别的特征在高维空间中都分布在一个点附近。该损失函数旨在通过优化特征与对应特征中心之间的距离，使得同类特征更加接近其所对应的中心，从而增加分类的准确性[19]。本文中，为每一种情绪学习一个特征中心，即为每个情绪类别学习一个特定的中心点；然后通过优化融合特征与其对应情绪类别中心的欧氏距离，使得同种情绪的特征向量聚集在彼此附近，而异种情绪的特征向量则被推开。中心损失的优化过程有助于减少同类特征的类内差距，即让同一情绪类别的特征更加紧密地聚集在其对应的中心周围。同时，其还可以增大类间差距，即让不同情绪类别的特征向量之间的距离变得更远，从而使得分类任务更加清晰和可分，从而在情绪分类任务中取得更好的学习效果和性能提升。

## 2　关键实现技术

如图 1 所示，一个完整的情感分类任务包含音频预处理、特征提取、特征分类 3 部分。首先对原始音频信号做不同预处理，获取不同层次的语音信息；然后通过深度学习进行特征提取获得不同的特征向量；最后使用分类算法对这些特征向量进行分类，从而得出情感分类的结果。

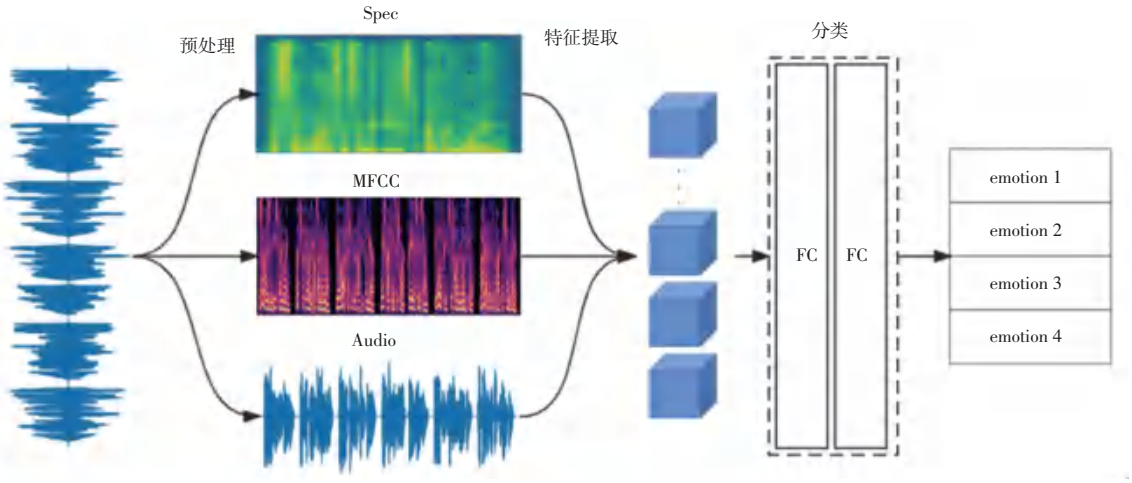



**图 1 多层次音频特征情感识别流程**

**Fig. 1 Multilevel audio features emotion recognition process**

## 2.1 信号时频域变换

直接获取的语音信号(如:.wav 和.mp3 等类型文件)都是时域信号,需要经过时频转换获取频域信息,如图 2 所示。其通过傅里叶变换和滤波等流程,从时域信息中获取 MFCC、频谱图等频域信息[17]。实现过程如下:

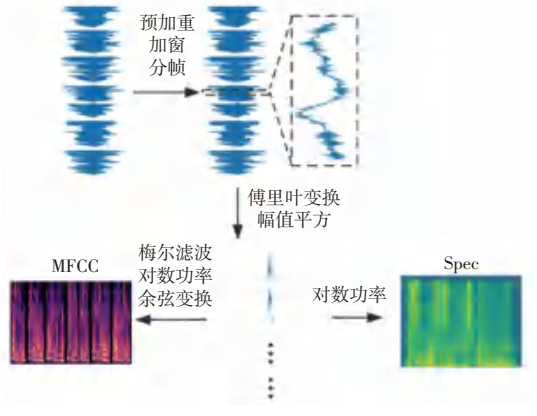



**图 2 时域信息转频域信息**

**Fig. 2 Time domain information to frequency domain information**

(1)首先对原始语音信号进行预加重(Pre-emphasis)。通常使用一个一阶滤波器以减小信号中的低频成分,增强与情绪分类相关的高频成分,如式(1)所示:

$$x_{pre}[n] = x_{original}[n] - \alpha x_{original}[n-1] \quad \alpha \in [0.9,1] \tag{1}$$

其中,$x_{original}[n]$ 为原始信号;$xpre[n]$ 为加重之后的信号;$\alpha$ 是预加重系数。

(2)预加重后的信号要进行分帧和加窗处理,将较长的连续信号按照一定的帧长和帧移分为便于处理的短帧。对短帧信号加窗进行平滑处理,以减小分帧引入的频谱泄漏(Spectral Leakage),如式(2)所示:

$$x_{win}[n] = x_{pre}[n] * w[n] \tag{2}$$

其中,$x[n]$ 和 $x_{win}[n]$ 分别是加窗前后的信号,$w[n]$ 则是窗函数。

(3)对获取的每一帧加窗信号 $x_{win}[n]$ 进行傅里叶变换,即可获取频域信息 $X[k,i]$,如式(3)所示:

$$X[k,i] = \sum_{n=0}^{N-1} x_{win}[n] * e^{-j2\pi\frac{kn}{N}} \tag{3}$$

其中,连续信号每一帧对应的各频率通道的强度就组成了一段语音信号的频谱图。将频域信号通过梅尔滤波器,便获得了 MFCC 信息。

## 2.2 特征提取器

将语音信号时频域规范化和预处理之后,可以获取时域信号 $x_w$、频谱图 $x_s$ 和梅尔倒谱系数 $x_m$,通过图 3 所示流程可以对其进行特征提取和分类。

首先,分别使用提取器 wav2vec、Alexnet 和 BiLSTM 提取三者的特征向量 $\boldsymbol{x}'_w$、$\boldsymbol{x}'_s$ 和 $\boldsymbol{x}'_m$,如式(4)所示:

$$\begin{cases} \boldsymbol{x}'_w = wav2vec2(x_w) \\ \boldsymbol{x}'_s = f_s(AlexNet(x_s)) \\ \boldsymbol{x}'_m = f_m(\mathrm{BiLTSM}(x_m)) \end{cases} \tag{4}$$

多层次特征通过注意力机制以及拼接的方式进行特征融合获取融合特征 $x$,如式(5)所示,通过 dropout 方法[27]对所得特征 $x$ 的各元素 $x_i$ 进行泛化获得泛化特征 $x'$,然后基于此泛化特征,使用多任务学习和中心损失模块进行特征调整,如式(6)所示:

$$x = \mathrm{softmax}(\boldsymbol{x}'_m \oplus \boldsymbol{x}'_s) * \boldsymbol{x}'_w \tag{5}$$

$$x'_i = \begin{cases} 0 & \text{with probablity p} \\ \dfrac{x_i}{1-p} & \text{otherwise} \end{cases} \tag{6}$$

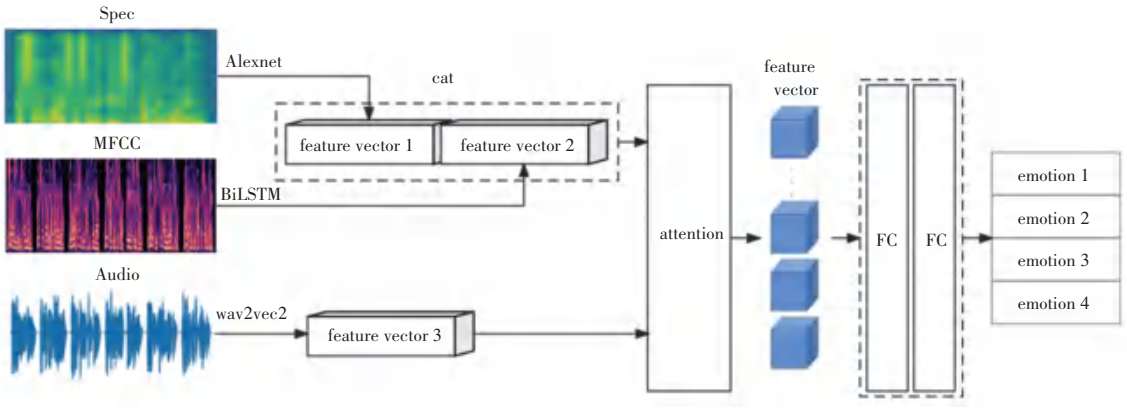



**图 3　多层次信息分类流程**

**Fig. 3　Multilevel information classification process**

本文调整特征分布的方式是加入其它相关特征、解耦融合前特征和加入中心损失对融合后特征进行约束处理。性别特征可以被认为是情感的协变量之一，采用拼接的方式加入此特征，以丰富特征包含的信息。

除了以上 3 种特征提取器，还尝试了 LTSF-Linear 系列[28]的模型（包括 Dlinear、Nlinear），用来提取音频特征。梅尔倒谱系数在标准化后可以认为是时序长度（对应总帧数）为 300，元素数为 40（谱图的 40 种频率）的时序特征。梅尔倒谱系数可用式（7）表示：

$$X=\begin{bmatrix} X_1^1 & \cdots & X_1^L \\ \vdots & \ddots & \vdots \\ X_C^1 & \cdots & X_C^L \end{bmatrix} \tag{7}$$

其中，$X_i^t$ 表示元素（频率）$i$ 在 $t$ 时间步（帧率）的值。

分类任务流程如图 4 所示，先对其进行归一化，将元素 $i$ 全序列 $\{X_i^1,\cdots,X_i^L\}$ 通过全连接层映射为 $n$ 维向量 $\{\hat{\boldsymbol{X}}_i^1,\cdots,\hat{\boldsymbol{X}}_i^n\}$，获取各元素（频率）特征向量之后，沿零维展开便可得到频谱图通过 Nlinear 提取器提取的特征向量。

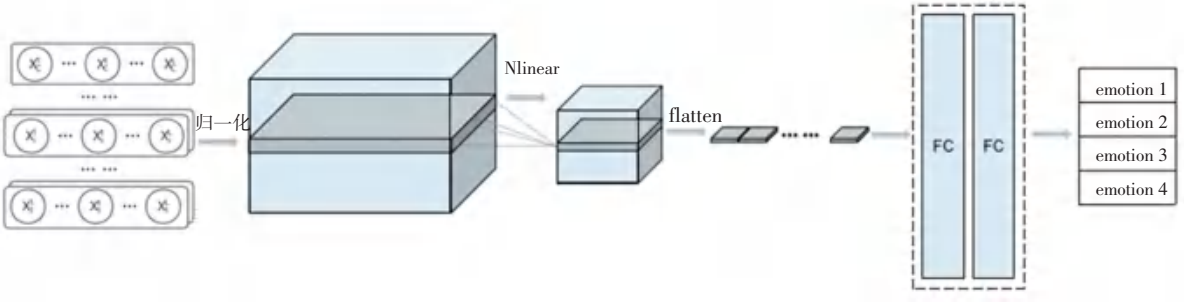



**图 4　Nlinear 对 MFCC 特征进行分类**

**Fig. 4　Nlinear classifies MFCC features**

Dlinear 原理如图 5 所示，Dlinear 认为时序数据可以进行分解为两部分，分别是趋势（Trend）分量和剩余（Remainder）分量。

即元素 $i$ 在所有时间步中，其变化由较为宏观的缓慢变化和较为微小剧烈的周期性变化构成。将元素 $i$ 序列用首尾 $w$ 个元素填充为 $\{X_i^1,\cdots,X_i^w,X_i^1,\cdots,X_i^L,X_i^{L-w},\cdots,X_i^L\}$，然后通过长度为 $w$ 的窗口作平均池化，此时得到元素 $i$ 趋势分量 $\{T_i^1,\cdots,T_i^L\}$。趋势分量与时间序列 $X$ 之差便是剩余分量 $R$，如式（8）所示。

$$X=T+R=\begin{bmatrix} X_1^1 & \cdots & X_1^L \\ \vdots & \ddots & \vdots \\ X_C^1 & \cdots & X_C^L \end{bmatrix}=\begin{bmatrix} T_1^1 & \cdots & T_1^L \\ \vdots & \ddots & \vdots \\ T_C^1 & \cdots & T_C^L \end{bmatrix}+\begin{bmatrix} R_1^1 & \cdots & R_1^L \\ \vdots & \ddots & \vdots \\ R_C^1 & \cdots & R_C^L \end{bmatrix} \tag{8}$$

与 Nlinear 类似，将元素 $i$ 趋势分量 $T_i$ 和剩余分量 $R_i$ 映射为 $n$ 维向量 $\{\hat{\boldsymbol{T}}_i^1,\cdots,\hat{\boldsymbol{T}}_i^n\}$ 与 $\{\hat{R}_i^1,\cdots,\hat{R}_i^n\}$，然后对其进行求和并沿着零维展开，便是 Dlinear 提

取器提取的特征向量。

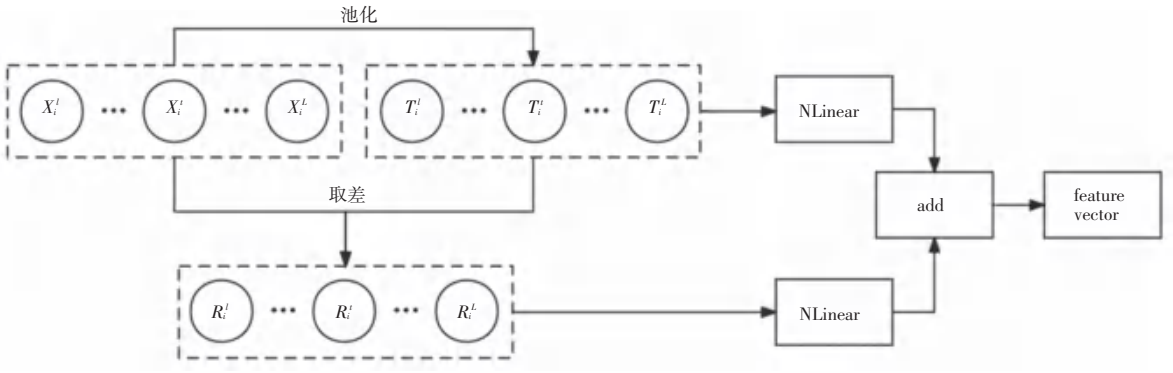



**图 5 Dlinear 提取 MFCC 特征**

**Fig. 5 Dlinear extracting MFCC features**

### 2.3 目标函数

将提取到的特征 $x$ 输入多层感知机并激活，最终得到分类结果。将全连接层的分类结果和标签的交叉熵 $Ls$ 作为损失函数是基本的优化方式。在此基础上加入多任务损失 $L_{MTL}$ 和中心损失 $L_C$ 作为联合损失 $L$ 进行优化，如式(9) 所示，总体流程如图 6 所示。

$$L = L_s + L_{MTL} + L_c \tag{9}$$

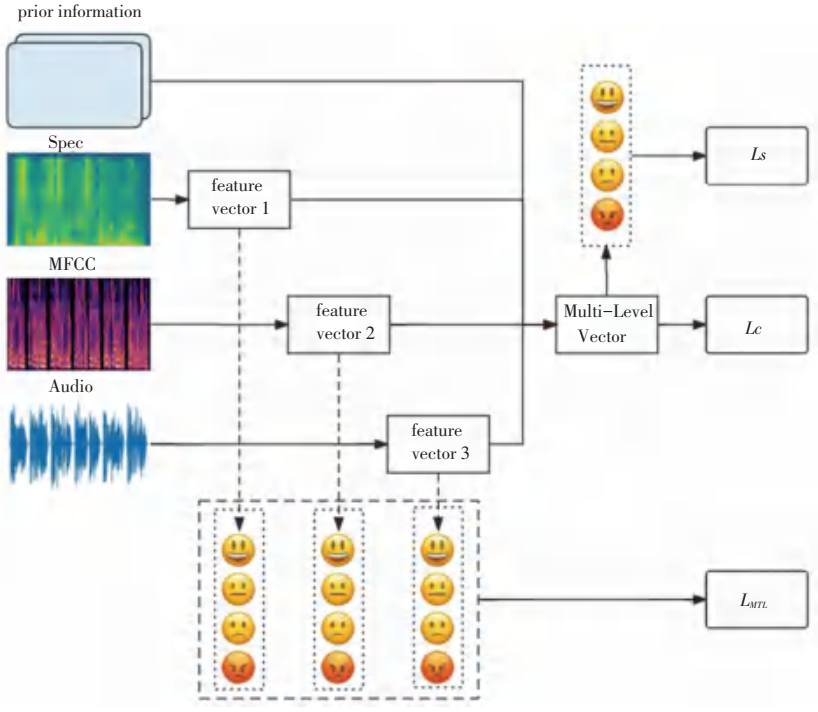



**图 6 加入先验信息、多任务学习和中心损失的流程**

**Fig. 6 The process of incorporating prior information, multitasking learning and central loss**

#### 2.3.1 多任务学习机制优化特征分布

多任务学习中，任务 $i$ 的损失为 $L_i$，根据各任务的重要性对其赋予权重系数 $w_i$。多任务学习的总损失 $L_i$ 如式(10) 所示：

$$L_{MTL} = \sum_i w_i \cdot L_i \tag{10}$$

本文在做融合特征分类的基础上，增加了对多层次特征的约束，为每个层次特征增加了分类任务，特征提取器参数为多个任务共享，共享参数 $W_{sh}$ 在梯度下降优化时的原理如式(11) 所示：

$$W_{sh} = W_{sh} - \gamma \sum_i w_i \cdot \frac{\partial L_i}{\partial W_{sh}} \tag{11}$$

#### 2.3.2 中心损失模块优化特征分布

分类任务中特征 $x_i$ 的类别是 $y_i$，假设同类特征在空间分布上聚集于同一特征中心，本文设计了一个中心损失模块，用于调整特征分布，其原理如图 7 所示。

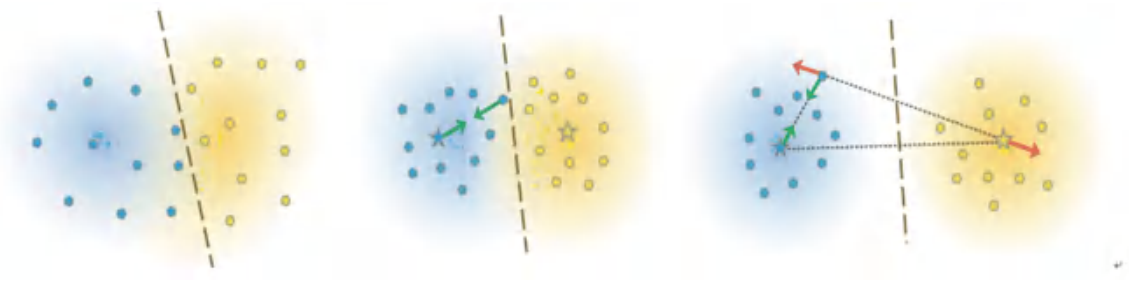


图 7　中心损失优化

Fig. 7　Center loss optimization

$y_i$ 类特征均是在特征中心 $c_{yi}$ 周围分布，中心损失 $Lc$ 通过式(12) 定义。

$$Lc = \frac{1}{2}\sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2 \tag{12}$$

如公式(13)，中心损失模块是和特征 $x_i$ 与本身特征中心 $y_i$ 与其它特征中心对比得到的。

$$Lc = \frac{1}{2}\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\frac{\exp\| x_i - c_{y_i} \|_2^2}{\exp\| x_i - c_j \|_2^2} - \frac{\exp\| x_i - c_{y_i} \|_2^2}{\exp\| x_i - c_j \|_2^2}\right) = \frac{1}{2}\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\exp(\| x_i - c_{y_i} \|_2^2 - \| x_i - c_j \|_2^2) - 1\right) \tag{13}$$

通过优化特征到特征中心的距离，可以增强模型的泛化能力。$Lc$ 的梯度通过式(14)计算。

$$\frac{\partial L_C}{\partial x_i} = \sum_{i=1}^{m}\left[\sum_{j=1}^{n}(-c_{y_i} + c_j)\exp(\| x_i - c_{y_i} \|_2^2 - \| x_i - c_j \|_2^2)\right] \tag{14}$$

$c_{yi}$ 的更新方式如式(15) 所示：

$$\Delta c_j = \frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot(c_j - x_i)}{1 + \sum_{i=1}^{m}\delta(y_i = j)} \tag{15}$$

加入中心损失 $Lc$ 和多任务损失之后，原损失函数 $L$ 变化为式(16)：

$$L = L_s + L_{MTL} + L_c = L_s + \sum_i w_i \cdot L_i + \frac{\lambda}{2}\sum_{i=1}^{m}\cdot \left(\sum_{j=1}^{n}\exp(\| x_i - c_{y_i} \|_2^2 - \| x_i - c_j \|_2^2) - 1\right) \tag{16}$$

## 3　验　证

为了验证本文提取器和音频分类模型的有效性，本文使用控制变量法，基于 Interactive Emotional Dyadic Motion Capture(IEMOCAP)[29] 数据集，设置了对比试验。

### 3.1　数据集

研究使用 IEMOCAP 数据集，该数据集中包含了十个人(其中：男性和女性各 5 名)的视频、文字和语音等多种与情感相关的信息。数据集分为十份，与每个人的数据对应，数据编号分别为：1M、1F、2M、2F、3M、3F、4M、4F、5M、5F(其中：F 为男性、M 为女性)。通过将本数据集多种情绪进行合并，将情绪划分为愤怒、悲伤、快乐和中性 4 类[20]。

为了评估本文模型性能，采用了十折交叉验证的评价方式。具体来说，是将数据集分为十份，其中九份用于模型训练，而一份用于模型测试。例如：将 1M 作为测试集时，其余九份就作为训练集评估模型。这样的交叉验证过程会进行十次，每次都轮流选择不同的测试集。最终，将所有折的评价指标进行加权平均，得到加权精度(*WA*) 和未加权精度(*UA*) 的平均值作为模型的最终评估结果[30]。在实验数据中，$N$ 表示类别数，$TP_i$ 表示 $i$ 类被预测正确样本个数，$FP_i$ 表示被错误预测为 $i$ 类的样本数，计算公式如下：

$$WA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i} \times 100\% \tag{17}$$

$$UA = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \times 100\% \tag{18}$$

未加权精度计算简单直观，但是忽略了样本不平衡因素。加权精度考虑了每个类别的样本数量，其是每个类别的分类准确率按照类别的样本比例加权平均得到的，更适用于样本不平衡的情况。通过这样的评估方式，可以全面地衡量模型在不同测试集上的性能，从而更客观地评估模型的泛化能力和鲁棒性。

### 3.2　实验设置

原始音频信号的采样频率为 16 kHZ，将原始音

频每 3 s 拆分成片段。除音频信息外,还用到其它层次语音信息,即梅尔倒谱系数和频谱图。

预处理信息主要分别做单层次语音特征提取和多层次语音特征提取。首先分别使用 wav2vec、BiLSTM、Alexnet、Nlinear 和 Dlinear 提取器提取语音特征,以进行单层次音频特征分类;然后是多层次语音特征提取实验,对于这 3 个层次的信息,初始模型分别使用 wav2vec、BiLSTM 和 Alexnet 进行特征提取,然后使用交叉注意力机制进行特征融合并进行特征分布调整,最后通过全连接层分类器进行分类。

提取特征和特征融合之后进行分类,分类模型使用 AdamW 优化器,单次参数更新样本数量设为 64,学习率设为 1e-5,每一折的验证集精度在连续训练 8 次仍没有提升,则结束本折训练。

### 3.3 结果分析

单层次语音特征提取与分类,分别使用以下的特征提取方法和对象:

(1) MFCC + BiLSTM:使用 BiLSTM 提取器从 MFCC 提取特征;

(2) MFCC + NLinear:使用 NLinear 提取器从 MFCC 提取特征;

(3) MFCC + DLinear:使用 DLinear 提取器从 MFCC 提取特征;

(4) Spec+Alexnet:使用 Alexnet 提取器从频谱图提取特征;

(5) Spec+BiLstm:使用 BiLstm 提取器从频谱图提取特征;

(6) Audio+W2E:使用 wav2vec 提取器从时域信息提取特征。

单层次语音信息分类的加权精度(*WA*)和未加权精度(*UA*)见表 1。

**表 1 单层次特征分类结果**

**Table 1 Single level feature classification result**

| | *WA*/% | *UA*/% |
|---|---|---|
| MFCC+BiLSTM | 57.6 | 58.9 |
| MFCC+NLinear | 55.56 | 54.94 |
| MFCC+DLinear | 56.91 | 58.28 |
| Spec+Alexnet | 64.81 | **66.82** |
| Spec+BiLstm | 61.40 | 62.50 |
| Audio+W2E | **65.89** | 66.39 |

对于 MFCC 特征的提取和分类,Dlinear 和 Nlinear 提取器效果相比较为成熟的 BiLSTM 提取器效果欠佳。当使用 BiLSTM 提取器分别对 MFCC 和频谱图进行单层次特征提取分类时,频谱图的分类结果较好,而时域信息提取器 wav2vec 则表现出最高的未加权精度。

关于多层次语音信息分类,初始模型是指三层次语音信息通过注意力机制融合并分类的模型。以此为基础,分别在初始模型基础上加入了性别先验信息、多任务机制和中心损失模块,为验证模型有效性,基于原始模型设置了以下实验,详细数据见表 2。

**表 2 多层次特征分类结果**

**Table 2 Multilevel feature classification results**

| | *WA*/% | *UA*/% |
|---|---|---|
| SER | 70.56 | 71.02 |
| SER+MF | **71.94** | 73.06 |
| SER+MTL | 71.88 | 72.26 |
| SER+CR | 70.68 | 72.35 |
| SER+MF+MTL | 71.75 | **73.37** |
| SER+MF+CR | 71.44 | 72.23 |
| SER+MTL+CR | 71.90 | 72.78 |
| SER+MF+MTL+CR | 70.31 | 71.64 |

(1)SER:初始模型;

(2)SER+MF:初始模型加入性别先验信息;

(3)SER+MTL:初始模型加入多任务机制;

(4)SER+CR:初始模型加入中心损失机制;

(5)SER+MF+MTL:初始模型加入性别先验信息和多任务机制;

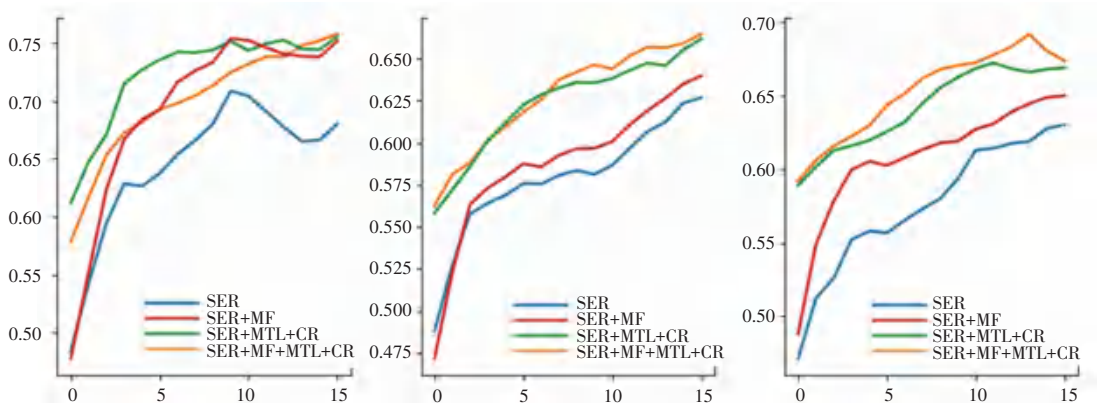
(6)SER+MF+CR:初始模型加入性别先验信息和中心损失机制;

(7)SER+MTL+CR:初始模型加入多任务机制和中心损失机制;

(8)SER+MF+MTL+CR:初始模型加入性别先验信息、多任务机制和中心损失机制。

通过表 1 和表 2 中多层次信息分类和单层次信息分类的结果(*WA* 和 *UA*)对比,多层次语音信息分类的效果明显优于单层次语音信息分类。这说明多个层次的语音特征通过注意力机制融合,对于情感分类任务是有效的。图 8 是部分测试集(1M,1F,2M)的精度曲线,横轴表示训练次数,纵轴表示未加权精度(*UA*)。可以看出,中心损失、先验知识和多任务学习机制提升了多层次语音分类效果。

由表 2 中的分类结果可以看出,多层次特征分类模型单独加入中心损失、多任务机制和先验信息之后,*WA* 和 *UA* 均得到了提升,验证了本文改进模型的有效性。其中,先验信息对性能提升最为明显。SER+MF+MTL 出现了最高的未加权精度(*UA*),且加权精度也较高,但相对于 SER+MF 略低。而 SER+MF+CR、SER+MTL+CR、SER+MF+MTL+CR 多个改进模块混合,相对于单个模块的效果似乎有所下降,这说明多个特征分布调整的改进未能形成合力。

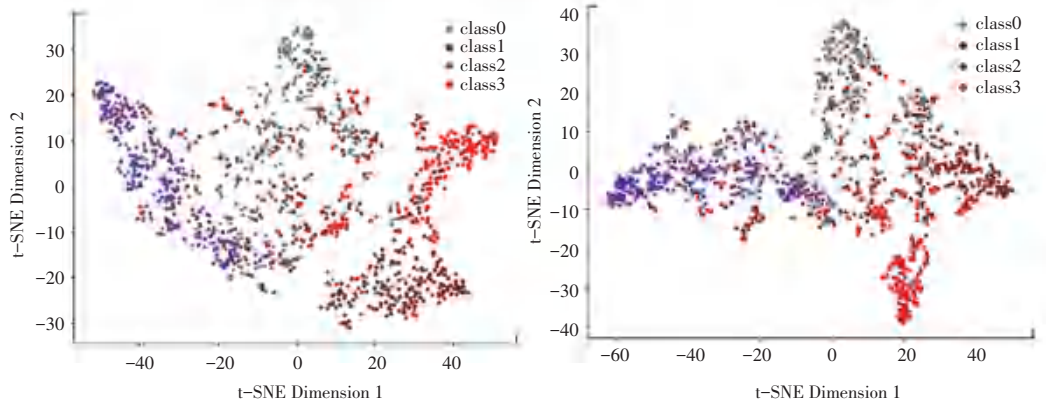



(a) 测试集为 1M　(b) 测试集为 1F　(c) 测试集为 2M

**图 8　不同测试集的精度曲线**

**Fig. 8　Accuracy curves for different test sets**

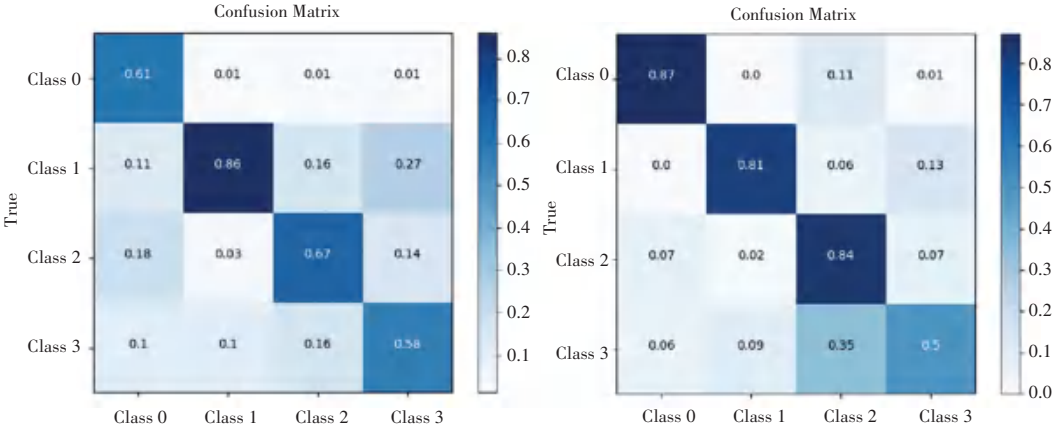
SER 和 SER+MF+MTL+CR 实验的特征可视化结果和混淆矩阵如图 9、图 10 所示。图 9 表示两个实验中的特征向量通过 t-sne 降维到二维时的分布图。图 10 中每一行代表了数据的真实归属类别，每一列代表了预测类别。通过对比可见，加入先验信息、多任务学习机制和中心损失模块后，融合后的特征向量分布得到了明显改善，同类特征更为聚集，异类特征之间的交叉部分更少。在混淆矩阵中，class0（愤怒）和 class2（快乐）的检测结果得到明显改善。整体的 *WA* 和 *UA* 证明了改进的有效性。




(a) SER Experiment　(b) SER+MF+MTL+CR Experiment

**图 9　通过 t-SNE 可视化提取的特征向量**

**Fig. 9　Visualizing the Features Extracted from the SER Experiment (a) and the SER+MF+MTL+CR Experiment (b) using t-SNE**




(a) SER Experiment　(b) SER+MF+MTL+CR Experiment

**图 10　SER 与 SER+MF+MTL+CR 实验的混淆矩阵**

**Fig. 10　Confusion matrices for the SER Experiment (a) and the SER+MF+MTL+CR Experiment (b)**

## 4 结束语

本文针对语音情感分类问题,在单层次音频特征分类任务中,本文设计的基于 Nlinear 和 Dlinear 提取器虽然也可以用于音频特征提取,但是其效果有待提升。在多层次音频特征的基础上,提出了一种加入性别先验信息和特征解耦的分类模型。将音频三特征(频谱、MFCC、先验性息)分别通过多任务学习机制作四分类处理进行解耦。最终将音频三特征与性别特征融合,并输入到分类器作四分类和计算中心损失。基于 IEMOCAP 数据集的消融试验表明,任务分类精度的加权精度和未加权精度均获得了提升,验证了模型改进的有效性。但是,改进模型的精度仍然有提升空间,当前也有许多新的语音特征提取器提出,未来将基于此进一步改进模型。

## 参考文献


[1] DEMSZKY D, MOVSHOVITZ-ATTIAS D, KO J, et al. GoEmotions: A dataset of fine-grained emotions[J]. arXiv preprint arXiv:2005.00547, 2020.

[2] FENG T, NARAYANAN S. PEFT-SER: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models[J]. arXiv preprint arXiv:2306.05350, 2023.

[3] AKHAND M A H, ROY S, SIDDIQUE N, et al. Facial emotion recognition using transfer learning in the deep CNN[J]. Electronics, 2021, 10(9): 1036.

[4] MARTIN C H, MAHONEY M W. Traditional and heavy-tailed self regularization in neural network models[J]. arXiv preprint arXiv:1901.08276, 2019.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017,60(6):84-90.

[6] LIPTON Z C, BERKOWITZ J, ELKAN C. A critical review of recurrent neural networks for sequence learning[J]. arXiv preprint arXiv:1506.00019, 2015.

[7] ZENG A, CHEN M, ZHANG L, et al. Are transformers effective for time series forecasting? [C]// Proceedings of the AAAI conference on artificial intelligence.2023, 37(9): 11121-11128.

[8] CLARK K, LUONG M T, MANNING C D, et al. Semi-supervised sequence modeling with cross-view training[J]. arXiv preprint arXiv:1809.08370, 2018.

[9] CHO K, VAN MERRIëNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv: 1406.1078, 2014.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017:6000-6010.

[11] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[12] OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv: 1807.03748, 2018.

[13] ZENG A, CHEN M, ZHANG L, et al. Are transformers effective for time series forecasting? [C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(9): 11121-11128.

[14] DAS A, KONG W, LEACH A, et al. Long-term forecasting with TiDE: time-series dense encoder[J]. arXiv preprint arXiv: 2304.08424, 2023.

[15] WOO G, LIU C, SAHOO D, et al. Learning deep time-index models for time series forecasting[C]// International Conference on Machine Learning. PMLR, 2023:37217-37237.

[16] XU H, ZHANG H, HAN K, et al. Learning alignment for multimodal emotion recognition from speech[J]. arXiv preprint arXiv:1909.05645, 2019.

[17] ZOU H, SI Y, CHEN C, et al. Speech emotion recognition with co-attention based multi-level acoustic information[C]// Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7367-7371.

[18] RUDER S. An overview of multi-task learning in deep neural networks[J]. arXiv preprint arXiv:1706.05098, 2017.

[19] KANG L, ZHANG L, JIANG D. Learning robust self-attention features for speech emotion recognition with label-adaptive mixup [C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[20] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations [C]// Advances in Neural Information Processing Systems, 2020: 12449-12460.

[21] HU D, BAO Y, WEI L, et al. Supervised adversarial contrastive learning for emotion recognition in conversations[J]. arXiv preprint arXiv:2306.01505, 2023.

[22] ALDENEH Z, PROVOST E M. Using regional saliency for speech emotion recognition[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 2741-2745.

[23] TATSUNAMI Y, TAKI M. Sequencer: Deep lstm for image classification[C]// Advances in Neural Information Processing Systems, 2022: 38204-38217.

[24] ZUPAN B, BABBAGE D, NEUMANN D, et al. Sex differences in emotion recognition and emotional inferencing following severe traumatic brain injury[J]. Brain Impairment, 2017, 18(1): 36-48.

[25] PLEISS G, CHEN D, HUANG G, et al. Memory-efficient implementation of densenets[J]. arXiv preprint arXiv: 1707.06990, 2017.

[26] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1290-1299.

[27] HEIDARI M, MOATTAR M H, GHAFFARI H. Forward propagation dropout in deep neural networks using Jensen-Shannon and random forest feature importance ranking[J]. Neural Networks, Neural Networks: the Official Journal of the Internatinal Neural Network Socienty, 2023,165:238-247.


(下转第 101 页)