第 42 卷 第 5 期
2023 年 9 月

应用声学
Journal of Applied Acoustics

Vol. 42, No. 5
September, 2023

◇ 研究报告 ◇

注意力机制融合前端网络中间层的语声情感识别

朱应俊 周文君 朱 川 马建敏†

(复旦大学航空航天系 上海 200433)

摘要: 为了使机器能够更好地理解人的情感并改善人机交互体验,可对语声特征及分类网络进行融合以提升情感识别性能。该文从网络融合的角度,把基于梅尔倒谱系数和逆梅尔倒谱系数的二维卷积神经网络和基于散射卷积网络系数的长短期记忆网络作为前端网络,提取前端网络的中间层作为话语级的特征表示,利用压缩-激励(SE)通道注意力机制对前端网络的中间层的权重进行调整并融合,然后由深度神经网络后端分类器输出情感分类结果。在汉语情感数据集中进行五折交叉验证的对比实验,实验结果表明,基于SE通道注意力机制的网络融合方式可以有效地利用不同前端网络在语声情感识别任务中的优势,提高语声情感识别的准确率。

关键词: 注意力机制;语声特征;网络融合

中图法分类号: TN912.3 **文献标识码:** A **文章编号:** 1000-310X(2023)05-1090-09

DOI: 10.11684/j.issn.1000-310X.2023.05.023

Speech emotion recognition using the attention mechanism to fuse the intermediate layer of front-end networks

ZHU Yingjun ZHOU Wenjun ZHU Chuan MA Jianmin

(*Department of Aeronautics and Astronautics, Fudan University, Shanghai* 200433, *China*)

Abstract: In order to enable machines to better understand human emotions and improve human-computer interaction experience, speech features and classification networks can be fused to improve emotion recognition performance. From the perspective of network fusion, this paper builds front-end networks including two dimensional convolutional neural network (2D-CNN) based on Mel-frequency cepstral coefficients, 2D-CNN based on inverted Mel-frequency cepstral coefficients, long short-term memory based on scattering convolution network coefficients. The intermediate layer of the front-end networks are then extracted as the feature representation of the discourse level, and the squeeze-and-excitation (SE) channel attention mechanism is introduced to adjust and fuse the weights of the intermediate layer. Eventually the sentiment classification results are output by the back-end network based on the deep neural network. A comparison experiment of five-fold cross-validation was carried out on the Chinese speech emotion data set. The experimental result showed that the network fusion based on the SE channel attention mechanism can effectively utilize the advantages of different front-end networks in speech emotion recognition tasks, and improve the accuracy of speech emotion recognition.

Keywords: Attention mechanism; Speech feature; Network fusion

2022-06-04 收稿; 2023-01-12 定稿

作者简介: 朱应俊 (1998–), 男, 山东济南人, 硕士研究生, 研究方向: 语声情感识别。

† 通信作者 E-mail: jmma@fudan.edu.cn

0　引言

语声情感识别(Speech emotion recognition, SER)已在娱乐产品的情感交互、远程教育的情感反馈、智能座舱的情绪监测中得到广泛应用。在应用中，通过建立语声信号的声学特征与情感的映射关系，对语声的情感进行分类。基于单一特征的SER模型因受到特征信息量不足的制约而影响识别准确率。随着对语声情感特征研究的逐步深入，通过对多种语声特征进行融合以消除特征中的冗余信息并提升识别准确率的方法受到越来越多的关注，已形成了特征级、中间层级、决策级等融合方式。

对语声情感特征进行特征级的融合可以在增加信息量并提高识别准确率的同时有效减小特征维度。Liu等[1]使用基于相关性分析和Fisher准则的特征选择方法，去除来自同一声源且具有较高相关性的冗余特征。Cao等[2]也提出了基于Spearman相关性分析和随机森林特征选择的方法提取相关性最弱的特征以进行融合。基于网络中间层进行的融合则利用神经网络将原始特征转化为高维特征表达，以获取不同模态数据在高维空间的融合表示。Cao等[3]在话语级别的情感识别中使用门控记忆单元(Gated memory unit, GMU)来获取语声信号的静态与动态特征融合后的情感中间表示。Zhang等[4]提出了基于块的时间池化策略用于融合多个预训练的卷积神经网络(Convolutional neural network, CNN)模型学习到的片段级情感特征，得到固定长度的话语级情感特征。语声特征的融合还可基于多个模型在其输出阶段进行决策级融合以集成其情感分类结果[5]。Noh等[6]使用基于验证准确度的指数加权平均法则组成了分级投票决策器对多个CNN模型的决策结果进行融合。Yao等[7]使用基于置信度的决策级融合整合了在多任务学习中获得的循环神经网络(Recurrent neural network, RNN)、CNN和深度神经网络(Deep neural network, DNN)。

注意力机制可用于自动计算输入数据对输出数据的贡献大小，近年来也在语声识别相关领域得到了较多运用。Bahdanau等[8]将注意力机制应用于RNN和n-gram语言模型，建立了端到端的序列模型。Mirsamadi等[9]将基于局部注意力机制的加权时间池化策略用于RNN模型，以学习与情感相关的短时帧级特征。Kwon[10]使用特殊的扩张CNN从输入的过渡语声情感特征中提取空间信息并生成空间注意力图以对特征进行加权。

在已有对语声特征融合及注意力机制在SER任务中应用研究的基础上，通过对语声信号进行预加重和分帧加窗等处理，得到基于谱特征和时序特征的前端网络，利用压缩-激励(Squeeze-and-excitation, SE)通道注意力机制对前端网络中间层进行融合，有效利用不同前端网络在SER任务中的优势提高情感识别准确率。通过在汉语情感数据集中的对比实验，对前端网络选择的合理性和SE通道注意力机制用于对前端网络中间层进行融合的有效性进行验证。

1　SER模型

本文判断语声信号情感类别的SER模型如图1所示，该模型由3个模块组成：前端网络模块、注意力机制融合模块和后端网络分类模块。前端网络模块对输入的语声信号进行预加重和分帧加窗等处理后，提取梅尔倒谱系数(Mel-frequency cepstral coefficients, MFCC)和逆梅尔倒谱系数(Inverted MFCC, IMFCC)作为谱特征，把谱特征输入到二维卷积神经网络(Two dimensional CNN, 2D-CNN)得到MFCC 2D-CNN和IMFCC 2D-CNN；提取散射卷积网络系数(Scattering convolution network coefficients, SCNC)作为时序特征，把时序特征输入到长短期记忆网络(Long-short term memory, LSTM)中得到SCNC LSTM。注意力机制融合模块引入SE通道注意力机制，将MFCC 2D-CNN、IMFCC 2D-CNN和SCNC LSTM前端网络中提取的中间层进行加权融合得到融合深度特征(Fusion deep feature, FDF)。后端分类模块基于DNN构建分类器，依据输入的FDF映射输出情感分类结果。

1.1　基于MFCC和IMFCC特征的2D-CNN前端网络

MFCC和IMFCC谱特征中不同频谱区间的频谱能量分布体现着不同情感状态下的声道形状和发声状态[11]，其中计算MFCC特征时使用的Mel三角滤波器模拟了人耳听觉的非线性机制，更加关注于语声信号的低频部分而对中高频的变化不够敏感[12]；IMFCC特征则通过IMel滤波器在高频区域分布更加密集来获取更多高频信息[13]。Hz 频率

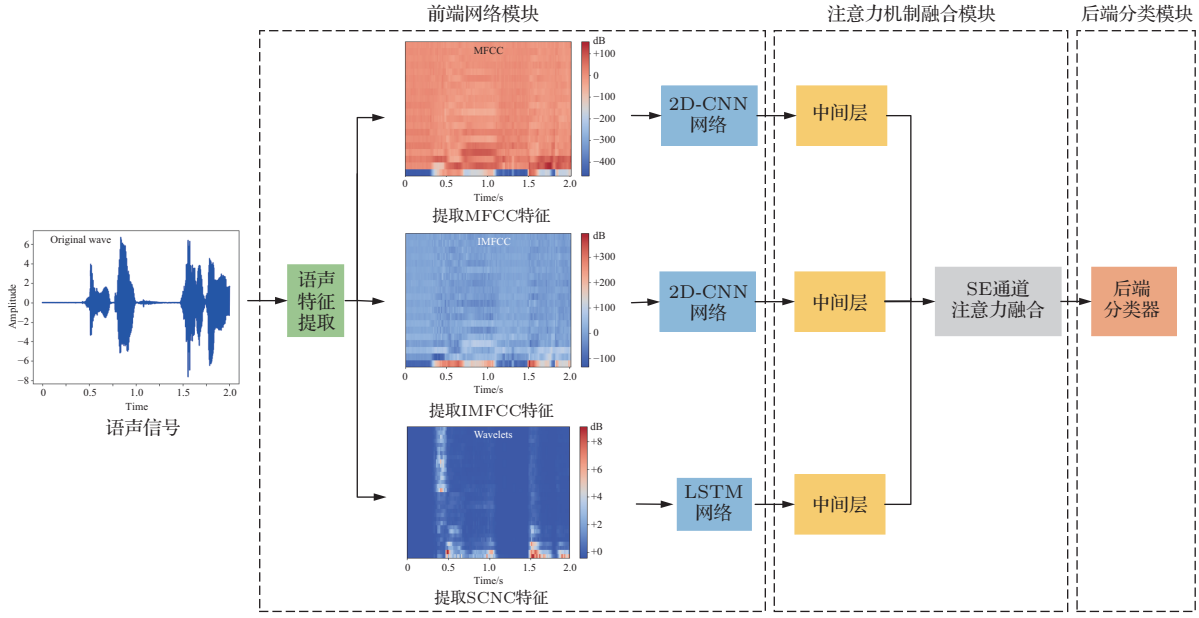


图 1 SER 模型结构

Fig. 1 Structure of SER model

与 Mel 频率及 IMel 频率之间的定量关系可分别表示为 [14]

$$f_{\mathrm{Mel}} = 2595 \times \lg\left(1 + \frac{f}{700}\right), \tag{1}$$

$$f_{\mathrm{IMel}} = 2146.1 - 1127 \times \lg\left(1 + \frac{4000 - f}{700}\right), \tag{2}$$

其中，f 表示 Hz 频率，f_{Mel} 和 f_{IMel} 分别为 Mel 频率及 IMel 频率。

将语声信号的功率谱通过 Mel 及 IMel 三角滤波器，并将对数能量带入离散余弦变换 (Discrete cosine transform, DCT) 以消除相关性，可计算得到语声信号的 MFCC 系数及 IMFCC 系数。还引入其一阶二阶差分项作为动态特征以体现语声情感的时域连续性 [15]。特征差分项 d_t 的实现如下：

$$d_t = \frac{\sum_{st=1}^{ST} st \times (c_{t+st} - c_{t-st})}{2\sum_{st=1}^{ST} st^2}, \tag{3}$$

其中，c_t 表示 MFCC 或 IMFCC 倒谱系数，st 表示一阶导数的时间差。将一阶差分结果重复带入即可得到二阶差分，最终可计算得到带有差分项的 MFCC 及 IMFCC 特征。

为了利用 CNN 在提取特征矩阵的局部空间相关性信息方面的优势 [16]，本文搭建了改进 Alexnet 的 2D-CNN，网络结构简图如图 2 所示，网络卷积部分的结构参数如表 1 所示。卷积层使用了 ReLU 激活函数，并进行了 L2 正则化，正则化参数为 0.02。在完成卷积运算后，使用扁平化层 (Flatten) 对卷积特征进行降维，输入到节点数分别为 2048 和 512 的两层全连接层对特征进行整合，并由 6 个节点的 Softmax 分类层得到情感分类结果。将 MFCC 和 IMFCC 特征分别输入 2D-CNN 训练得到 MFCC 2D-CNN 前端网络和 IMFCC 2D-CNN 前端网络。

表 1 2D-CNN 前端网络卷积层参数

Table 1 Parameters of convolutional layers in 2D-CNN front-end network

网络层数	核尺寸	核数量	步长
2D Conv1	5×5	32	2
Maxpooling1	3×3		2
2D Conv2	5×5	128	2
Maxpooling2	3×3		2
2D Conv3	3×3	256	1
2D Conv4	3×3	256	1
2D Conv5	3×3	128	1

在反向传播过程中，为了应对由样本量过少及训练数据分布不均衡导致的网络性能下降的问题，本文引入了 Focal loss 损失函数 [17]，通过给难分类样本 (Hard example) 较大的权重，给易分类样本 (Easy example) 较小的权重，来放大难分类样本的损失并抑制易分类样本的损失，从而使网络聚焦于难分类样本的学习，提高分类准确率。Focal loss 损失函数 L_{fl} 的计算如下：

$$L_{\mathrm{fl}} = -\alpha_t (1 - p_t)^{\gamma} \lg(p_t), \tag{4}$$

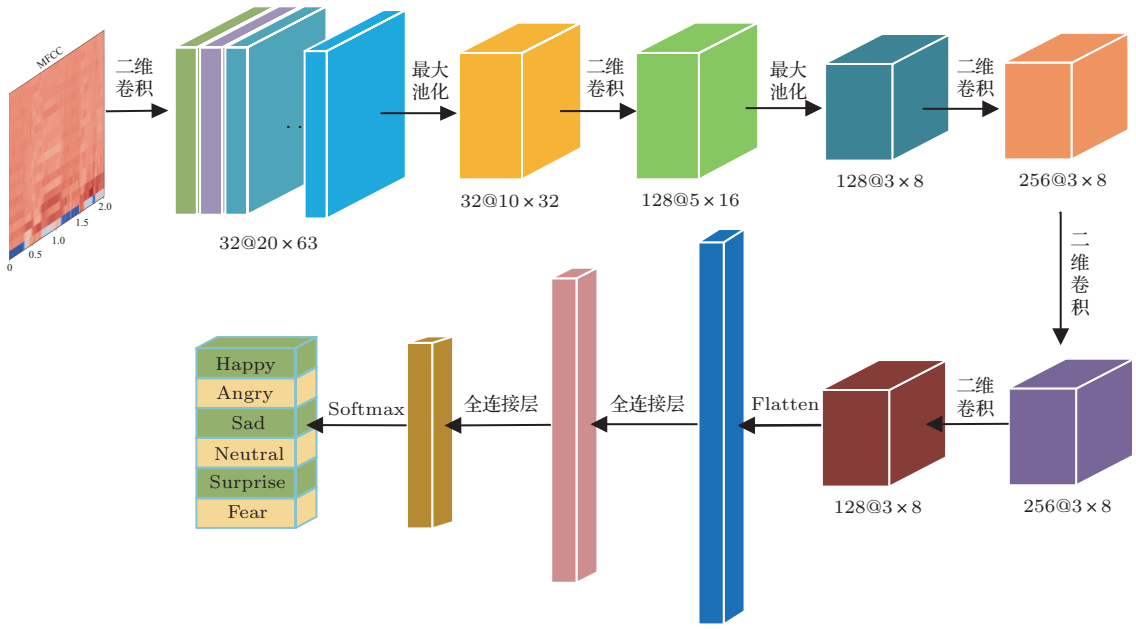


图 2　基于 MFCC 与 IMFCC 的 2D-CNN 前端网络结构

Fig. 2　2D-CNN front-end network structure based on MFCC and IMFCC

其中，p_t 表示分类器预测的概率值，γ 为权重放大因子，α_t 是类别权重。为了增大 2D-CNN 前端网络对难分类样本的权重，将 γ 取为 4，因为数据集中各类情感样本数目相同，将 α_t 设置为 1。

1.2　基于 SCNC 特征的 LSTM 前端网络

本文引入了由不变散射卷积网络 (Invariant scattering convolution network, ISCN) 自动提取的 SCNC 特征 [18] 作为时序特征。将语声帧视作短时平稳信号，输入由多层小波散射变换与取模算子级联得到的 ISCN 中，提取其散射系数作为 SCNC 特征，该特征能够最小化信号的平移和形变的影响，具有较强的变形稳定性，且保留用于分类的高频信息，故在网络中间层对特征进行融合时能够维持分类鲁棒性 [19]。

对语声信号进行的小波变换可表示为 $\{x \otimes \psi_\lambda\}_\lambda$，其中指数 $\lambda = 2^{-j}r$ 给出了带通滤波器 ψ_λ 的频率位置，$\otimes$ 表示卷积运算，对于语声信号仅计算 λ 在 $r \in [0, \pi)$ 范围内所对应的路径。沿路径 $\boldsymbol{p} = (\lambda_1, \lambda_2, \cdots, \lambda_m)$ 迭代进行小波变换和取模运算可求得小波变换系数：

$$\begin{aligned} U[\boldsymbol{p}]x &= U[\lambda_m]\cdots U[\lambda_2]U[\lambda_1]x \\ &= |||x \otimes \psi_{\lambda_1}| \otimes \psi_{\lambda_2}| \cdots | \otimes \psi_{\lambda_m}|. \end{aligned} \tag{5}$$

为得到具有更好变形稳定性的局部描述符，将空间窗函数 $\Phi_2^J(u) = 2^{-2J}\Phi\left(2^{-J}u\right)$ 与小波系数进行积分以得到路径 $\boldsymbol{p}$ 上的加窗散射系数：

$$\begin{aligned} S[\boldsymbol{p}]x(u) &= U[\boldsymbol{p}]x \otimes \phi_{2^J}(u) \\ &= \int U[\boldsymbol{p}]x(v)\,\phi_{2^J}(u-v)\,\mathrm{d}v, \end{aligned} \tag{6}$$

其中，对于每条路径 $\boldsymbol{p}$，$S[\boldsymbol{p}]x(u)$ 是窗口位置 u 的函数，将式 (5) 代入其中即可得到计算 m 阶加窗散射系数的公式如下：

$$S[\boldsymbol{p}]x(u) = |||x \otimes \psi_{\lambda_1}| \otimes \psi_{\lambda_2}| \cdots | \otimes \psi_{\lambda_5}| \otimes \phi_{2^J}(u). \tag{7}$$

为了提高特征的高频分辨率，将分帧加窗后的语声片段输入由 5 层小波变换和取模算子级联得到的 ISCN 中，以提取网络的加窗散射系数作为 SCNC 特征。

LSTM 相较于 CNN 可以更好地处理时间序列的任务，同时 LSTM 解决了 RNN 的长时依赖问题 [20]，并避免了反向传播过程中的梯度消失 [21]。本文搭建了基于 SCNC 特征的 LSTM 前端网络，网络由 LSTM 层和 3 层全连接层组成，为对应每帧语声提取到的 32 维的 SCNC 特征，LSTM 层设置了 32 个节点，每个节点通过 126 个时间步进行更新 [22]。单个节点的结构如图 3 所示。

在 LSTM 节点中，X_t 表示 SCNC 特征沿时间轴的输入，C_t 表示由当前输入产生的细胞待更新的状态，由输入门 i_t 和遗忘门 f_t 决定当前细胞状态要如何更新，细胞状态的迭代公式为

$$C_t = f_t \times C_{t-1} + i_t \times \hat{C}_t. \tag{8}$$

h_t 表示当前节点输出的隐藏状态，由输出门 o_t 和当前细胞状态计算得到，使用 tanh 函数作为激活函数，其计算如下：

$$h_t = o_t \times \tanh(C_t). \tag{9}$$

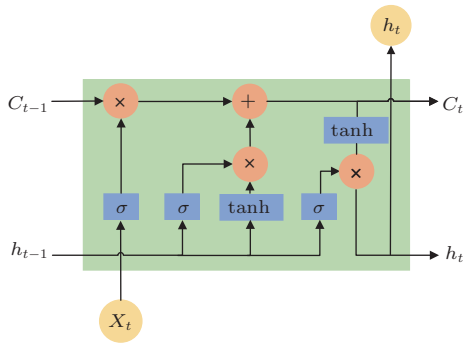


图 3　单个 LSTM 节点的内部结构

Fig. 3　Internal structure of LSTM node

将 LSTM 网络层输出的全部隐藏状态 $\boldsymbol{H}$ 使用 Flatten 层降维后输入到节点数分别为 1024 和 256 的全连接层进行特征整合，激活函数为 ReLU 函数，全连接层后使用了 Dropout 函数以抑制过拟合，Dropout 率为 0.3，并由 6 个节点的 Softmax 分类层得到情感分类结果。将 SCNC 特征输入 LSTM 以训练得到 SCNC LSTM 前端网络。

1.3　基于 SE 通道注意力机制的网络中间层融合

在 SER 任务中，MFCC 2D-CNN 和 IMFCC 2D-CNN 前端网络更加关注谱特征中的语声能量信息，而 SCNC LSTM 前端网络则侧重于语声的时序性信息。为了发挥两类网络的优势，本文将前端网络模型视作特征提取器，分别提取了 MFCC 2D-CNN 与 IMFCC 2D-CNN 前端网络最后一层卷积层的输出，提取了 SCNC LSTM 前端网络的隐藏状态 $\boldsymbol{H}$。前端网络的中间层深度特征作为话语级的特征表示，由于不同网络中的深度特征对情感分类的贡献程度不同，本文引入 SE 通道注意力机制，利用 SE Block 对各前端网络中间层权重进行调整 [23]，融合过程如图 4 所示。

SE 通道注意力机制的实现通过两步完成。第一步为 Squeeze 操作，对应于图 4 中的全局平均池化，其实现如下：

$$\boldsymbol{z}_c = F_{sq}(\boldsymbol{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{u}_c(i,j), \tag{10}$$

其中，压缩函数 F_{sq} 在特征维度上对中间层矩阵 $\boldsymbol{u}_c$ 进行压缩降维，将 $H \times W \times C$ 的多通道特征降为 $1 \times 1 \times C$ 的 C 维向量，以表征网络中间层的全局信息。第二步的 Excitation 操作对全局平均池化后生成的 $\boldsymbol{z_c}$ 依次进行了全连接、ReLU 激活、全连接、Sigmoid 激活，得到代表各通道重要性的权重矩阵，其表达式为

$$\begin{aligned} \boldsymbol{s} &= F_{ex}(\boldsymbol{z_c}, W) = \sigma(g(\boldsymbol{z_c}, W)) \\ &= \sigma(W_2 \delta(W_1 \boldsymbol{z_c})), \end{aligned} \tag{11}$$

其中，δ 为线性激活函数，W_1 与 W_2 为两个全连接层，σ 为 Sigmoid 激活函数。

将 Excitation 操作后求得的权重矩阵 $\boldsymbol{s}$ 与前端网络中间层矩阵相乘可得到 FDF 矩阵，从而实现由多通道的联合深度特征 (Joint deep feature, JDF) 向 FDF 的转变。

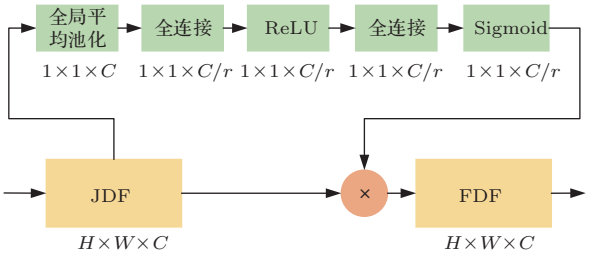


图 4　SE 通道注意力机制融合过程

Fig. 4　SE channel attention mechanism workflow

1.4　DNN 后端分类器

利用 SE 通道注意力机制融合前端网络中间层得到了 FDF 矩阵作为话语级的情感特征，输入基于 DNN 的后端网络分类器进行 SER，网络共有 5 层全连接层，节点数分别为 2048、512、256、64，激活函数均为 ReLU 函数，最后由 Softmax 分类层输出得到多分类预测矩阵，取概率最大的一类作为最终的情感预测结果。在网络中使用了 Dropout 来抑制过拟合，其中 Dropout 率为 0.2。为了研究基于 SE 通道注意力机制的网络中间层融合方式对每一类情感的识别效果，将 DNN 后端网络的分类结果基于混淆矩阵进行输出表示。

2　实验与结果分析

实验部分首先通过消融实验对语声特征的维度选择及前端网络设计的合理性进行了验证，其次通过与前端融合和中间层非计权融合的对比实验验证了 SE 通道注意力机制用于网络中间层融合的有效性，最后通过与参考文献中融合方式的对比

实验对基于SE通道注意力机制的网络融合方式在SER任务中的准确率与时间复杂度进行了分析。

2.1 实验平台与数据集

实验选用的CPU型号为11th Gen Intel Core i5-11400，搭配4666 MHz频率的双通道DDR4内存，容量共32 GB，用于深度学习加速的GPU型号为NVIDIA GeForce RTX3060，显存容量为12 GB，开发使用的语言版本为Python 3.8.3，使用的深度学习框架为Tensorflow 2.4.0。

本文实验基于中国科学院自动化研究所录制的汉语情感语料库的部分数据进行，该数据子集包含了来自4位说话者的1200条语声，其情感倾向包括生气(Anger)、悲伤(Sad)、害怕(Fear)、开心(Happy)、中性(Neutral)、惊讶(Surprise)，语声的采样率为16000 Hz。实验中，将语声片段的时长统一为2 s共32000个采样点，对其进行加窗分帧操作后可得到126个语声帧。求得各语声特征维度如表2所示。

表 2　语声特征及维度

Table 2　Speech features and its dimension

语声特征	特征维度
1D-MFCC	1×1×126
2D-MFCC	**1×39×126**
3D-MFCC	3×13×126
1D-IMFCC	1×1×126
2D-IMFCC	**1×39×126**
3D-IMFCC	3×13×126
16 SCNC	1×16×126
32 SCNC	**1×32×126**
64 SCNC	1×64×126

2.2 实验设置

为消除数据集划分方式对网络性能的影响，将中国科学院自动化研究所语声情感数据集进行随机排序，并按照80%、10%、10%的比例划分为训练集、验证集和测试集。取五折交叉验证后的各情感平均分类准确率(Average ACC)和宏F1得分(Macro-F1 Score)作为网络性能的评价指标。

为验证前端网络设置及对应特征维度选择的合理性，实验分别对比了：(1) 基于一维谱特征1D-MFCC与1D-IMFCC的1D CNN前端网络。(2) 基于三维谱特征3D-MFCC与3D-IMFCC的3D-CNN前端网络。(3) 使用平均池化(Ave-pool)层的2D-CNN前端网络。(4) 基于16维与64维SCNC特征的LSTM前端网络。(5) 基于32维SCNC特征的2D-CNN前端网络。为验证在网络中间层进行融合相较于特征级融合的优势，实验对比了两类前端融合方式：(1) 前端特征级注意力机制融合。(2) 前端特征级非计权融合。除此之外，还比较了对网络中间层进行非计权融合后的网络性能。

为了进一步验证SE通道注意力机制用于网络中间层融合的适用性，还和文献[2]中基于随机森林特征选择算法的前端融合、文献[3]中基于GMU的分层网络中间层融合和文献[7]中基于置信度的后端融合方式进行了比较分析，并取预测测试集的总耗时作为时间复杂度指标进行讨论。

2.3 实验结果与讨论

不同维度语声特征在对应前端网络中的分类结果如表3中所示。由表3可知基于二维MFCC特征的2D-CNN前端网络相较于基于一维及三维MFCC特征的前端网络取得了更高的平均准确率和宏F1得分；基于二维IMFCC特征的2D-CNN前端网络亦优于基于一维与三维IMFCC特征的前端网络；且最大池化在2D-CNN前端网络中的效果好于平均池化。对比16维与64维的SCNC特征可知，基于32维SCNC特征的LSTM前端网络性能更好，且优于基于SCNC特征的2D-CNN前端网络。

分析可知，对于二维MFCC和IMFCC特征，2D-CNN前端网络可有效利用特征矩阵中的频谱能量信息进行分类。而最大池化相较于平均池化，对特征矩阵中的纹理信息更加敏感，更有利于对区分性信息的提取。对于SCNC特征，LSTM前端网络能够更好地学习序列中的时间相关性，由5层ISCN提取的32维SCNC特征则可较好地保留用于分类的高频信息。

将本文所选的3类前端网络的分类结果表示为混淆矩阵，如图5所示，其中对角线数据表示网络对每类情感的识别准确率。观察混淆矩阵可知，3类前端网络对“中性(Neutral)”与“愤怒(Angry)”两类情感的识别准确率显著高于其余情感类别。

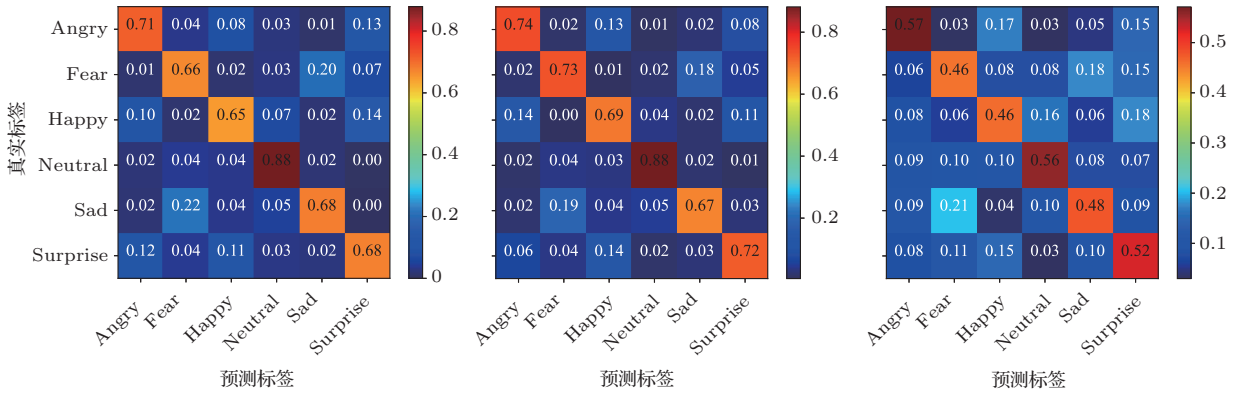
基于SE通道注意力机制的网络中间层融合方式对比前端融合方式与中间层非计权融合方式的情感分类结果如表4所示，观察可知，前端特征级的拼接融合或注意力机制融合相较于单一特征仅能使情感分类的平均准确率小幅提升，这证明了前端融合特征泛化能力有限，无法充分利用多种语声

特征的优势。而基于网络中间层进行非计权拼接融合后的准确率相较于特征级融合有了显著提高，但其表现依旧差于采用SE通道注意力机制的融合方式。这证明了基于网络中间层进行的融合优于特征级的融合，也进一步验证了基于SE通道注意力机制进行融合的有效性。不同融合方式取得的分类混淆矩阵分别如图6所示，观察可知后端分类网络均在“中性”情感上取得了最高的识别准确率，这也证明了前端网络在某一类情感识别中的优势在融合后可以得到保留。

表3　三类语声特征在不同前端网络中的分类结果

Table 3　Classification results of three SER features in different front-end networks

前端网络	准确率/%							宏F1/%
	Angry	Fear	Happy	Neutral	Sad	Surprise	Average	
1D-MFCC 1D-CNN	72.09	45.45	51.61	62.16	58.00	40.00	54.89	54.40
2D-MFCC 2D-CNN(Max-pool)	**71.10**	**66.47**	**64.83**	**87.90**	**67.86**	**67.53**	**70.95**	**70.84**
2D-MFCC 2D-CNN(Avg-pool)	83.33	67.65	56.41	81.82	65.52	60.46	69.20	69.04
3D-MFCC 3D- CNN	72.97	48.84	58.97	71.05	60.00	68.42	63.38	62.92
1D-IMFCC 1D-CNN	50.00	61.36	44.00	64.71	67.57	51.35	56.50	56.34
2D-IMFCC 2D-CNN(Max-pool)	**74.11**	**72.87**	**68.97**	**88.16**	**67.44**	**71.78**	**73.89**	**73.88**
2D-IMFCC 2D-CNN(Avg-pool)	72.74	69.05	64.71	90.75	70.23	66.96	72.41	72.36
3D-IMFCC 2D-CNN	71.79	60.00	52.17	77.78	50.00	76.19	64.67	63.95
32-SCNC 2D-CNN	64.09	47.73	34.94	42.59	56.36	37.99	47.28	47.25
16-SCNC LSTM	58.97	44.44	50.00	45.45	48.94	40.54	48.06	48.51
32-SCNC LSTM	**57.12**	**45.52**	**46.12**	**56.88**	**47.52**	**52.33**	**50.91**	**50.97**
64-SCNC LSTM	61.29	39.02	39.53	40.54	60.98	55.32	49.45	49.32



(a) MFCC 2D-CNN　(b) IMFCC 2D-CNN　(c) SCNC LSTM

图5　三类前端网络的分类混淆矩阵

Fig. 5　Confusion matrix for three front-end networks

表4　不同网络融合方式的对比实验结果

Table 4　Comparative test results of different network fusion methods

方法	准确率/%							宏F1/%
	Angry	Fear	Happy	Neutral	Sad	Surprise	Average	
前端拼接	66.14	60.33	65.17	87.69	79.04	68.29	71.11	70.94
前端融合	75.61	74.11	70.62	86.50	74.14	76.53	76.25	76.21
中间层拼接	92.74	81.85	83.84	96.88	89.21	80.75	87.55	87.55
中间层融合	**90.97**	**93.42**	**91.50**	**96.08**	**84.59**	**92.59**	**91.52**	**91.50**

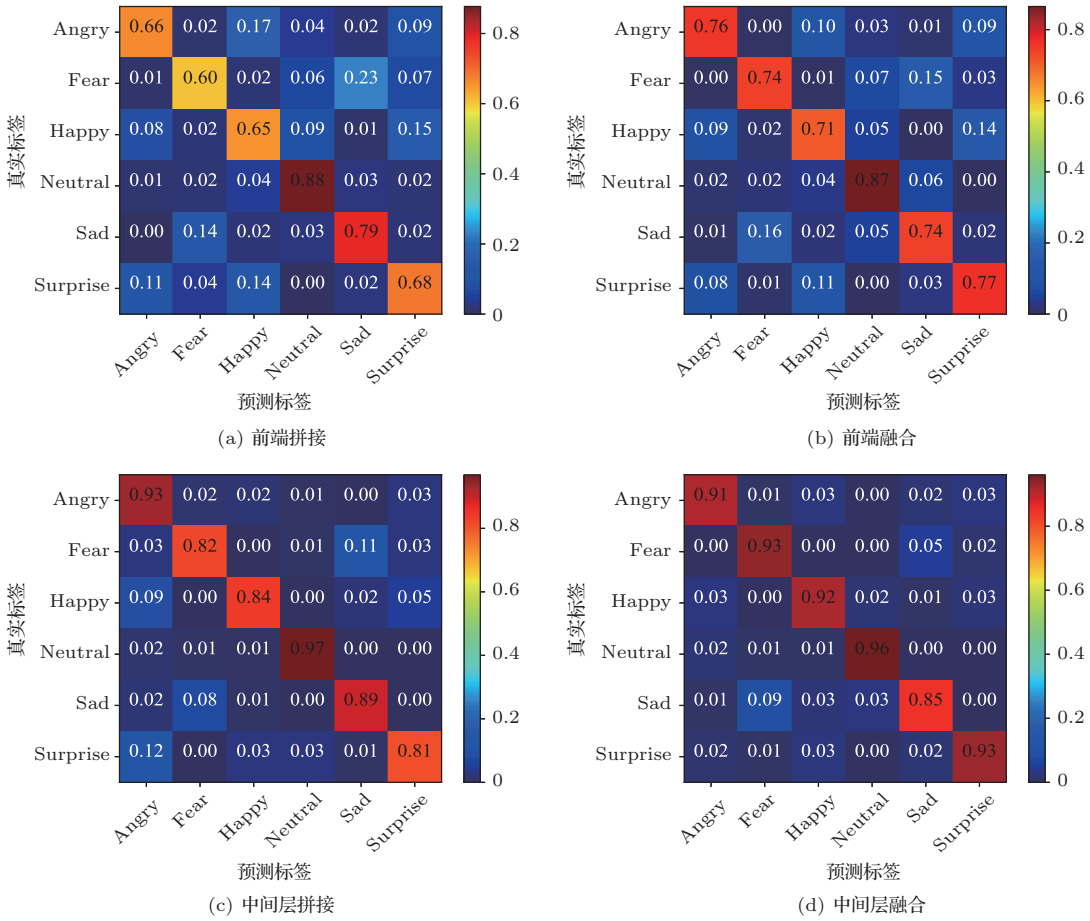


图 6　不同网络融合方式的分类混淆矩阵

Fig. 6　Confusion matrix for different network fusion methods

文献 [2–3,7] 中不同阶段的融合方式在测试集上的平均准确率和预测耗时如表 5 所示。观察数据可知，基于随机森林特征选择算法的特征融合方式 [2] 所用预测时间最短，这也体现了传统机器学习方法在预测效率上的优势。基于置信度的后端决策级融合方式 [7] 在使用多类语声特征获得较高的准确率的同时耗费了最长的预测时间。而基于 GMU 的网络中间层融合方式 [3] 对动静态谱特征进行融合则可兼顾识别效率与准确率。本文相较于融合方式 [3] 在谱特征的基础上增加了时序特征，使用 SE 通道注意力机制用于网络中间层融合，平均准确率提高了 5.39%，预测耗时则仅增加 0.015 s。对比实验证明了本文基于通道注意力机制的融合网络用于 SER 任务时，通过对多种语声特征和分类网络的有效利用，可以实现更高的平均识别准确率。

表 5　融合方式的准确率与复杂度对比

Table 5　Accuracy and complexity comparison

方法	文献 [2]	文献 [3]	文献 [7]	本文
平均准确率/%	78.61	86.13	87.05	**91.52**
时间复杂度/s	0.035	0.064	0.102	**0.079**

3　结论

本文把 SE 通道注意力机制用于对基于谱特征的和时序特征的前端网络的中间层融合，并进行了实验验证。实验结果表明，多特征分类相较于单一特征分类在情感识别准确率上具有明显的优势；中间层融合的多特征融合方式优于前端特征级的融合方式；利用 SE 通道注意力机制对前端网络中间层进行融合，能有效利用不同前端网络在 SER 任务中的优势提高情感识别准确率。

参　考　文　献

[1] Liu Z T, Wu M, Cao W H, et al. Speech emotion recognition based on feature selection and extreme learning machine decision tree[J]. Neurocomputing, 2018, 273: 271–280.

[2] Cao W H, Xu J P, Liu Z T. Speaker-independent speech emotion recognition based on random forest feature selection algorithm[C]. 2017 36th Chinese Control Conference, 2017: 10995–10998.

[3] Cao Q, Hou M, Chen B, et al. Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021: 6334–6338.

[4] Zhang S, Chen A, Guo W, et al. Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition[J]. IEEE Access, 2020, 8: 23496–23505.

[5] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423–443.

[6] Noh K, Lim J, Chung S, et al. Ensemble classifier based on decision-fusion of multiple models for speech emotion recognition[C]. 2018 International Conference on Information and Communication Technology Convergence, 2018: 1246–1248.

[7] Yao Z, Wang Z, Liu W, et al. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN[J]. Speech Communication, 2020, 120: 11–19.

[8] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition[C]. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 4945–4949.

[9] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, 2017: 2227–2231.

[10] Kwon S. Att-Net: enhanced emotion recognition system using lightweight self-attention module[J]. Applied Soft Computing, 2021, 102: 107101.

[11] Nwe T L, Foo S W, de Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech Communication, 2003, 41(4): 603–623.

[12] Kishore K V K, Satish P K. Emotion recognition in speech using MFCC and wavelet features[C]. 2013 3rd IEEE International Advance Computing Conference, 2013: 842–847.

[13] 胡峰松, 张璇. 基于梅尔频率倒谱系数与翻转梅尔频率倒谱系数的说话人识别方法 [J]. 计算机应用, 2012, 32(9): 2542–2544.
Hu Songfeng, Zhang Xuan. Speaker recognition method based on Mel frequency cepstrum coefficient and inverted Mel frequency cepstrum coefficient[J]. Journal of Computer Application, 2012, 32(9): 2542–2544.

[14] Tang Y Y, Lu Y, Yuan H. Hyperspectral image classification based on three-dimensional scattering wavelet transform[J]. IEEE Transactions on Geoscience and Remote sensing, 2014, 53(5): 2467–2480.

[15] 钟浩, 鲍鸿, 张晶. 一种改进的语音动态组合特征参数提取方法 [J]. 电脑与信息技术, 2017, 25(3): 4–7.
Zhong Hao, Bao Hong, Zhang Jing. An improved extraction method of speech dynamic combination characteristic parameters[J]. Computer and Information Technology, 2017, 25(3): 4–7.

[16] Stolar M N, Lech M, Bolia R S, et al. Real time speech emotion recognition using RGB image classification and transfer learning[C]. 2017 11th International Conference on Signal Processing and Communication Systems, 2017: 1–8.

[17] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980–2988.

[18] Stevens S S, Volkmann J. The relation of pitch to frequency: a revised scale[J]. The American Journal of Psychology, 1940, 53(3): 329–353.

[19] Bruna J, Mallat S. Invariant scattering convolution networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1872–1886.

[20] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks[C]. International Conference on Machine Learning, 2013: 1310–1318.

[21] Prabowo Y D, Warnars H L H S, Budiharto W, et al. LSTM and simple RNN comparison in the problem of sequence to sequence on conversation data using bahasa indonesia[C]. 2018 Indonesian Association for Pattern Recognition International Conference, 2018: 51–56.

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

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132–7141.