第 41 卷 第 5 期
2022 年 9 月

应用声学
Journal of Applied Acoustics

Vol. 41, No. 5
September, 2022

◇ 研究报告 ◇

基于STA-CRNN模型的语声情感识别*

张志浩[1,2] 王坤侠[1,2†]

(1 安徽建筑大学电子与信息工程学院 合肥 230601)

(2 安徽建筑大学安徽省建筑声环境重点实验室 (安徽建筑大学) 合肥 230601)

摘要: 语声情感识别对人机交互和情感计算研究领域具有重要作用,各类研究方法层出不穷。近期研究学者应用卷积神经网络和长短期记忆网络方法提取对数Mel谱图空间特征和时间特征,取得了一定的成果。然而不论是卷积神经网络还是长短期记忆网络提取特征时,都会产生特征冗余,导致语声情感识别效果下降。针对这一问题,该文提出了一种基于时空注意力机制的卷积-递归神经网络模型,采用对数Mel谱图和其一阶差分、二阶差分作为特征输入,在使用卷积神经网络提取空间特征和长短期记忆网络提取时间特征时,加入空间注意力和时间注意力机制,从而使上述网络能够更好地提取到对数Mel谱图中有效表征情感的空间特征和时间特征。该模型在Emo-DB和IEMOCAP语声数据集上的加权准确率分别达到86.8%、69.4%,未加权准确率分别达到84.7%、65.5%,优于当前大多数先进方法。

关键词: 语声情感识别;对数Mel频谱图;时空注意力;时间特征;空间特征

中图法分类号: TN912.34 **文献标识码:** A **文章编号:** 1000-310X(2022)05-0843-08

DOI: 10.11684/j.issn.1000-310X.2022.05.021

Speech emotion recognition based on STA-CRNN model

ZHANG Zhihao[1,2] WANG Kunxia[1,2]

(1 *College of Electronic and Information Engineering, Anhui Jianzhu University, Hefei* 230601, *China*)

(2 *Key Laboratory of Architectural Acoustic Environment of Anhui Higher Education Institutes* (*Anhui Jianzhu University*), *Hefei* 230601, *China*)

Abstract: Speech emotion recognition (SER) plays an important role in the research fields of human-computer interaction and affective computing. Many new research methods have emerged. Recently, researchers applied convolutional neural network (CNN) and long short-term memory (LSTM) to extract spatial and temporal features from Log-Mel spectrum, and achieved better performance. However, when CNN and LSTM networks extract features, they will lead to feature redundancy and reduce the performance of speech emotion recognition. In this paper, we propose a convolution recursive neural network model based on spatiotemporal attention mechanism (STA-CRNN). The Log-Mel spectrum, its first-order difference and second-order difference are used as feature input. We extract spatial features by CNN and temporal features by LSTM, and adopt spatial attention and temporal attention mechanism to further decrease the redundancy of features. The experiment

2022-03-15 收稿; 2022-05-06 定稿

*国家自然科学基金项目 (62001004), 安徽省高校学科 (专业) 拔尖人才学术资助项目 (gxbjZD2021067), 安徽建筑大学科研发展基金项目 (JZ202118), 安徽省高校自然科学研究重点项目 (KJ2020A0470), 安徽建筑大学安徽省建筑声环境重点实验室开放课题 (AAE2021ZR02)

作者简介: 张志浩 (1998–), 男, 安徽滁州人, 硕士研究生, 研究方向: 语声情感识别。

†通信作者 E-mail: kxwang@ahjzu.edu.cn

results show that the weighted accuracy (WA) of the model on Emo-DB and IEMOCAP Speech database are 86.8% and 69.4% respectively, and the unweighted accuracy (UA) are 84.7% and 65.5% respectively. The proposed model STA-CRNN achieves better performance than most advanced methods for SER.

Keywords: Speech emotion recognition; Log-Mel; Spatiotemporal attention; Time features; Spatial features

0 引言

语声情感识别(Speech emotion recognition, SER)是“情感计算”研究领域的一个重要分支[1]。SER在人机智能辅助[2]、人机交互[3−4]、行为识别[5]等应用中发挥着重要作用。在人机交互中,通过输入的语声信号识别说话人的情感状态,可以起到监管、协助和指引的作用。因此SER的研究是一项关键并富有挑战性的任务[6−7]。

近年来,SER研究者们通过对多种特征和分类器的深入研究,使得SER的性能逐渐提高[8]。SER最显著的特征是从整条语声中计算出的一维的低级描述符(LLDs),例如能量、基频(F0)和Mel频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)等[9]。这些特征可以全面地捕捉语声的情感信息,进而有效地改善SER的识别率。然而这些手工特征对于表征语声中的情感信息并不是最有效的,这可能导致性能不佳[10]。而卷积神经网络(Convolutional neural network, CNN)和长短期记忆(Long short-term memory, LSTM)网络在SER特征提取方面表现出卓越的性能[11]。这两种网络能够从大量训练样本中提取关键信息特征进而提高SER的识别率[12]。Mao等[13]提出利用CNN提取LLDs有效的语声情感信息,并在多个公开数据集上表现出优异的性能。Senthilkumar等[14]使用LSTM学习语声信号之间帧与帧的特征信息,获得了较好的SER结果。但是,传统的一维LLDs特征存在着频域信息缺失问题[15],因此,研究者们纷纷将语声信号转换成二维时频特征如频谱图、对数Mel频谱图(Log-Mel)等作为SER模型的输入,用来提取语声高级情感特征,与传统声学特征相比表现出更好的性能[16−17]。如Trigeorgis等[18]利用CNN和LSTM网络提取语声频谱图的时空特征,其实验结果要优于在一维特征上的实验结果。Zhang等[19]则将一维的语声信号转换为具有RGB三通道的频谱图,并将其作为CNN模型的输入。其实验结果表明,在三通道语声频谱图上进行SER的性能比一维特征优越。尽管上述研究者们利用频谱图作为特征并取得了较为不错的效果,但是在使用CNN或LSTM网络进行特征提取时,忽略了在情感识别方面语声频谱图的不同片段区域存在较大差异性。而模型对于图中有效的空间特征和时间特征的提取能力有限,导致大量的有效特征和无效特征冗余,从而限制了SER模型的性能。而注意力机制(Attention mechanism)则可以利用其加权机制来过滤掉冗余特征[20−22],捕获频谱图中的关键情感信息,有利于关键特征的提取与学习,进而提高SER识别率。

为了解决有效的情感特征提取问题,本文基于以上研究提出时空注意力-卷积递归神经网络(Spatiotemporal attention-Convolution recursive neural network, STA-CRNN)模型,即在CRNN模型中引入空间注意力(Spatial attention)[23]机制和时间注意力(Temporal attention)机制。在CNN进行空间特征提取时,空间注意力机制可以聚焦空间关键信息,使网络能够关注情感显著区域。在LSTM网络进行时间特征提取时,时间注意力机制可以对不同时间序列片段特征给予权重,提高有效特征的提取能力。

本文贡献如下:(1) 提出了一种基于时空注意机制的CRNN网络模型,包括CNN和LSTM两个模块;(2) 经过实验确定了空间注意力机制在CNN网络层中的最佳层间位置;(3) 验证了时空注意力机制的CRNN网络模型能够明显提高SER识别率。

1 STA-CRNN模型结构

本文提出了一种基于时空注意力机制的CRNN模型。模型分为两大部分:基于空间注意力机制的CNN网络和基于时间注意力机制的LSTM网络。模型结构如图1所示,首先将Log-Mel谱图和其一阶、二阶差分组成三维Log-Mel谱图,输入到基于空间注意力机制的CNN网络中,充分提取其空间特征;其次将输出结果输入到基于时间注意力机制的LSTM网络中,再将得到的向量输入到全连接层中,进行Softmax分类,最终得到情感类别。

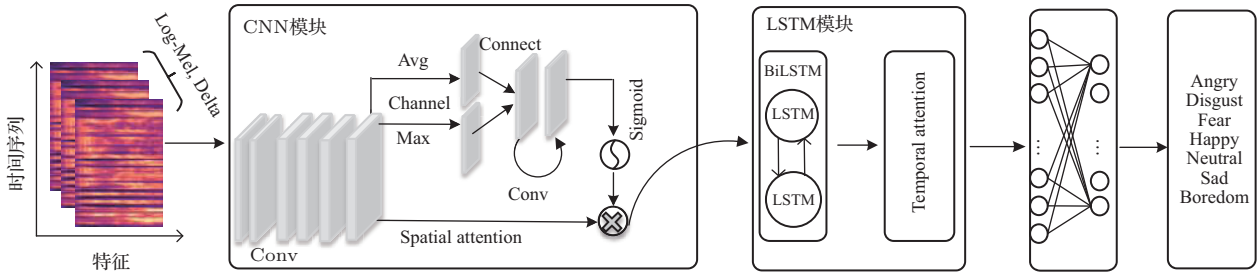


图 1　STA-CRNN 模型结构

Fig. 1　Structure of STA-CRNN

1.1　基于空间注意力机制的 CNN 网络

在 CNN 网络中，卷积模块包括 6 个卷积层、2 个最大池化层和 1 个空间注意力层。卷积层的第一层和第二层有 128 个输出通道，池化层的大小为 2×4。其他卷积层的输出通道为 256，卷积核的大小为 5×3，为了防止过拟合，在每个卷积后添加 Relu、Drop 和 BN 模块，并在最后一次卷积后，加入空间注意力层。

空间注意力的计算过程如下：首先将卷积模块输出的特征图作为空间注意力层的输入特征图，再对其做一个基于通道的最大池化 (Maxpool) 和平均池化 (Avgpool) 操作，并将它们连接起来生成一个有效的特征图，如公式 (1) 所示：

$$M_s(F)=\sigma(f^{7\times7}([F^S_{\text{avg}};F^S_{\max}])), \tag{1}$$

其中，$f^{7\times7}$ 表示 7×7 的卷积核尺寸，$F^S_{\text{avg}}\in R^{1\times H\times W}$ 和 $F^S_{\max}\in R^{1\times H\times W}$ 分别表示通道上的平均池化特征和最大池化特征。

接着将这两个结果做融合和卷积操作，降维至一个通道。再经过 Sigmoid 生成空间注意力特征图。最后将该特征图和该模块的输入特征图做乘法，得到最终生成的特征。

1.2　基于时间注意力机制的 BiLSTM 网络

为了获取 Log-Mel 谱图的时间特征，采用基于注意力机制的双向长短时记忆网络 (BiLSTM) 对卷积模块的输出结果进行时间序列上的特征提取。LSTM 主要由遗忘门、输入门、输出门以及隐藏状态所构成，具体的计算过程为

$$f(t)=\sigma(\boldsymbol{W}^f_h h_{t-1}+\boldsymbol{W}^f_x x_t+b_f), \tag{2}$$

$$i(t)=\sigma(\boldsymbol{W}^i_h h_{t-1}+\boldsymbol{W}^i_x x_t+b_i), \tag{3}$$

$$\alpha(t)=\tanh(\boldsymbol{W}^\alpha_h h_{t-1}+\boldsymbol{W}^\alpha_x x_t+b_\alpha), \tag{4}$$

$$o(t)=\sigma(\boldsymbol{W}^o_h h_{t-1}+\boldsymbol{W}^o_x x_t+b_o), \tag{5}$$

$$c(t)=c(t-1)*f(t)+i(t)*\alpha(t), \tag{6}$$

$$h(t)=o(t)*\tanh(c(t)), \tag{7}$$

其中，$f(t)$、$i(t)$、$o(t)$、$c(t)$ 分别表示 t 时刻遗忘门、输入门、输出门的值，$\alpha(t)$ 表示 t 时刻对 h_{t-1} 和 x_t 的初步特征提取。x_t 表示 t 时刻的输入，h_{t-1} 表示 $t-1$ 时刻的隐层状态值。$\boldsymbol{W}$ 表示权重矩阵，b 为偏置值；tanh 表示正切双曲函数，σ 表示激活函数 Sigmoid。

BiLSTM 由两个 LSTM 层组成，并通过方向相反的两个 LSTM 层来提取信息，包括将来和过去的隐藏信息，最后拼接并输出，见公式 (8)：

$$\boldsymbol{z}_t=\boldsymbol{w}_{\overrightarrow{z}x}\overrightarrow{\boldsymbol{h}}_t+\boldsymbol{w}_{\overleftarrow{z}x}\overleftarrow{\boldsymbol{h}}_t+b_x. \tag{8}$$

在 BiLSTM 的基础上，添加一个时间注意力层，该注意力机制会根据式 (9) 计算出在不同时间序列，BiLSTM 输出序列的权重参数，然后采用式 (10) 根据权重大小将编码序列中的每一个向量进行加权求 μ 和，最终得到 attention 数值，并通过 Softmax 分类器预测情感类别。

$$\mu_t=\frac{\exp(\boldsymbol{W}h_t)}{\sum_{t=1}^{T}\exp(\boldsymbol{W}h_t)}, \tag{9}$$

$$\mu=\sum_{t=1}^{T}\mu_t h_t. \tag{10}$$

2　实验与分析

为了检验本文所提出模型的性能，选择柏林语声情感数据集 (Emo-DB)[24] 和交互式情绪二元运动捕捉 (IEMOCAP) 数据集 [25] 进行相关实验。

Emo-DB：该数据集由来自 7 种情绪状态 (anger、boredom、disgust、fear、happy、neural、sad) 共 535 个话语组成。库中的话语由讲德语的 10

名专业演员 (5 名男性，5 名女性) 记录。Emo-DB 语料库的平均重新编码时间约为 2 ∼ 3 s，采样率为 16 kHz。Emo-DB 作为一种标准数据集，已被广泛用于情绪研究。

IEMOCAP：该数据集由 10 名专业演员在 5 个不同的会话中记录，每个会话有 2 名演员 (1 名男性，1 名女性)。演员以脚本和即兴版本记录不同情绪的对话，由 3 位专家进行注释，并由至少 2 位专家同意选择。最终形成包括视频、语声和文本以及 9 种情感 (anger、happy、excitement、sadness、frustration、fear、surprise、other、neural) 的离散标签，其中语声平均话语时长为 3.5 s。由于该数据库存在样本不均衡的情况，本文和诸多国内外研究者们 [11,20] 相同，选择 (anger、happy、sad、neural) 这 4 类人类基本情感进行实验，并且该数据集所有对比实验也均是上述 4 类情感。

2.1 数据预处理

实验采用 Log-Mel 谱图作为输入。Log-Mel 谱图提取过程如下：首先将语声信号进行分帧、加窗，随后进行离散傅里叶变换计算每一帧的功率谱，并通过 Mel 滤波得到 Mel 频谱图，最后进行对数运算，即可得到 Log-Mel 谱图 [26]。但由于语声数据的时长参差不齐，帧长不一致，故在提取 Log-Mel 谱图时，统一将帧长归为 300 帧，不足 300 帧的用 0 填充，超过 300 帧的进行分割处理。此外，为了得到语声的动态特征，提取 Log-Mel 谱图的一阶、二阶差分并和 Log-Mel 谱图组成 3D Log-Mel 特征集。

2.2 评估策略及参数设置

本文将 Emo-DB 数据库和 IEMOCAP 数据库划分为 10 份数据，采用十折交叉验证法，轮流将其中 9 份数据作为训练集，一份作为测试集，最后将 10 轮实验结果的准确率取平均值作为最终识别准确率。实验使用两种常见的评估指标：加权准确率 (Weighted accuracy, WA) 和未加权准确率 (Unweighted accuracy, UA)[27]，来衡量模型的性能。WA 是测试集中所有样本的准确率，UA 是全部情绪准确率的平均值。

本文在实验中使用了自适应矩估计 (Adam) 优化器，并将其初始学习率设置为 0.001，权值衰减为 5×10^{-4}，批量大小等于 32。该模型的参数通过最小化交叉熵目标函数进行优化，而最大迭代次数设置为 150。

2.3 实验结果分析

2.3.1 空间注意力机制在 CNN 层间位置的实验结果分析

为了研究 CNN 网络中空间注意力层和卷积层不同层间位置关系对于识别率的影响，依次把空间注意力层放入第一层卷积至最后一层卷积之后，记为 CNN(AC1)-CNN(AC7)，所有的实验均在 Emo-DB 库和 IEMOCAP 上进行了 150 次迭代，实验结果如表 1 所示。

表 1 空间注意力机制在 CNN 不同层间位置实验结果

Table 1 Experimental results of spatial attention mechanism in different layers of CNN

(单位: %)

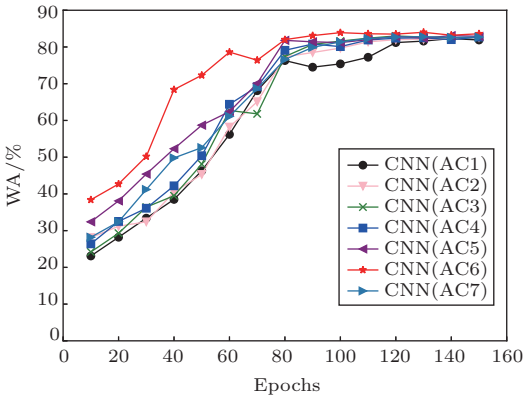
方法	Emo-DB(WA/UA)	IEMOCAP(WA/UA)
CNN(AC1)	81.9/80.1	63.4/59.1
CNN(AC2)	82.3/80.5	63.6/59.3
CNN(AC3)	82.7/80.9	64.6/60.7
CNN(AC4)	82.5/81.3	64.9/60.8
CNN(AC5)	82.9/81.3	65.4/61.5
CNN(AC6)	83.6/82.4	66.2/62.4
CNN(AC7)	82.7/81.0	65.3/61.4

通过表 1 可知，随着卷积层的加深，网络提取的有效特征越来越多，此时可以通过注意力机制聚焦关键情感特征，从而提高模型的识别率。IEMOCAP 库中 AC1 至 AC6 识别率持续升高，Emo-DB 库虽然 AC4 变低，但总体也呈上升趋势。然而从 AC6 到 AC7 时识别率开始变低。此外，由于层数的增多，也会增加训练负担。

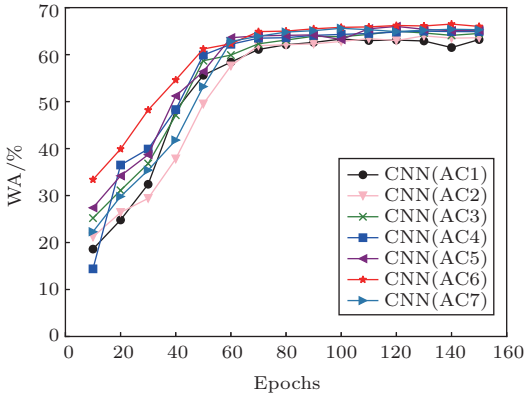
图 2 展示了在 150 次迭代中，AC1 至 AC7 的收敛曲线。

从图 2 中可以看出，在 Emo-DB 库中，经过 80 次迭代后，模型趋于稳定，其中 AC6 不仅在第 10 代准确率最高，而且收敛速度也最快。在 IEMOCAP 库中，经过 60 次迭代后，模型趋于稳定，AC6 也同样取得了最好的实验效果。

综上所述，本文选择在第 6 层卷积后加入空间注意力机制具有科学性和有效性。



(a) 在Emo-DB库中



(b) 在IEMOCAP库中

图 2　CNN(AC1-7) 在 Emo-DB 库和 IEMOCAP 库中的收敛曲线

Fig. 2　Convergence curves of CNN (AC1-7) in Emo-DB and IEMOCAP

2.3.2　STA-CRNN 实验结果分析

为了验证本文所提模型的有效性，创建了由 CNN+BiLSTM 模型以及在此基础上加入两种注意力机制所组成的基线模型。设置了具体实验如下：

(1) Base1(CNN+BiLSTM)：以 CNN+BiLSTM 作为基线模型。

(2) Base2(ACNN+BiLSTM)：在 Base1 的基础上，在 CNN 中加入时空注意力层。

(3) Base3(CNN+A-BiLSTM)：在 Base1 的基础上，在 BiLSTM 网络后加入时间注意力层。

(4) Base4(并行 STA-CRNN)：将带有空间注意力机制的 CNN 和时间注意力机制的 BiLSTM 采取并行方式，采用特征融合的方法，和本文串行的 STA-CRNN 模型进行对比。

(5) 本文模型 (STA-CRNN)：和上述实验进行对比，验证本文所提出网络的有效性。

表 2 为 4 种模型的实验结果。

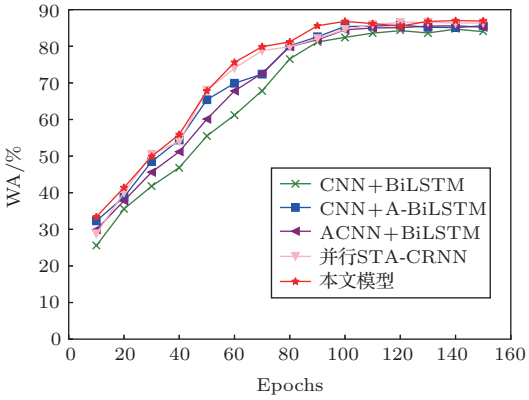
表 2　4 种模型实验结果

Table 2　Experimental results of four types of models

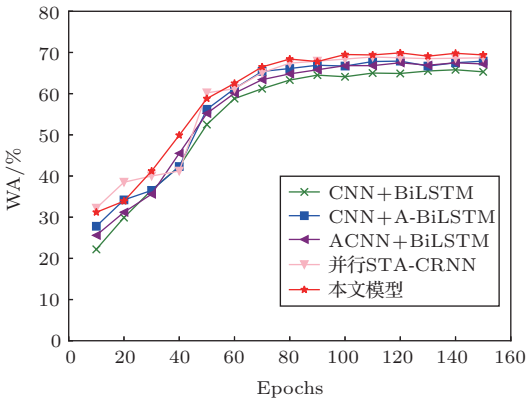
(单位：%)

方法	Emo-DB (WA/UA)	IEMOCAP (WA/UA)
(Base1)CNN+BiLSTM	84.4/82.3	65.6/61.3
(Base2)ACNN+BiLSTM	85.1/83.4	67.2/63.5
(Base3)CNN+A-BiLSTM	85.4/83.8	67.6/63.9
(Base4) 并行 STA-CRNN	86.4/83.9	68.7/64.6
本文模型 (STA-CRNN)	86.8/84.7	69.4/65.5

图 3 为在 Emo-DB 和 IEMOCAP 库中 4 种模型的收敛曲线。



(a)　在Emo-DB库中



(b)　在IEMOCAP库中

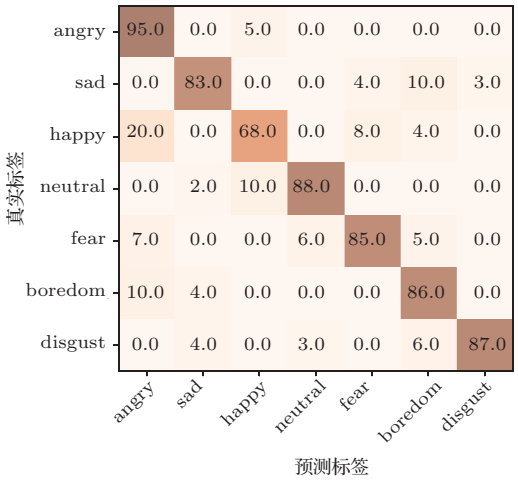
图 3　在 Emo-DB 和 IEMOCAP 库中 4 种模型收敛曲线

Fig. 3　Convergence curves of four model in Emo-DB and IEMOCAP

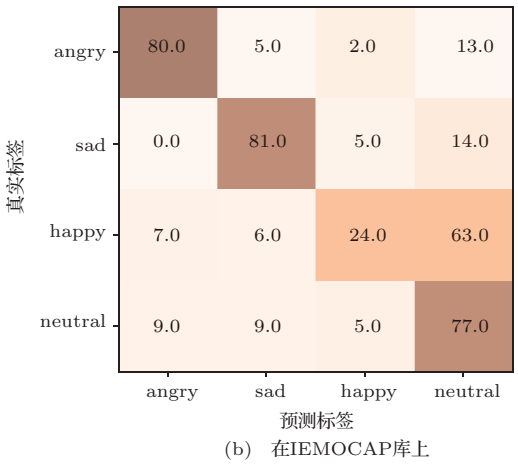
通过表 2 和图 3 可知，在 Base1 的基础上，不论是在 CNN 中 (Base2) 还是在 BiLSTM 中 (Base3) 加入注意力层，WA 和 UA 都有增长，但是都没有达到预期的效果。而本文的模型把两种注意力机制结合

形成了时空注意力机制，集合了两者的优点，无论是串行结构还是并行结构均使得模型的性能得到了显著改善。串行的STA-CRNN模型相较于并行的STA-CRNN模型，在两个数据库上均取得了最好的结果，在Emo-DB库中WA和UA分别高了0.2%、0.8%，在IEMOCAP中分别高了1.3%、1.1%，并且其收敛曲线也最为稳定。

图4展示了本文模型在Emo-DB和IEMOCAP库上实验的混淆矩阵，通过分析混淆矩阵可知，本文所提出的模型对于angry的识别率较好，在Emo-DB和IEMOCAP库上分别达到了95%和80%的准确率。



(a) 在Emo-DB库上



(b) 在IEMOCAP库上

图4 本文模型在Emo-DB和IEMOCAP库上实验的混淆矩阵

Fig. 4 The confusion matrix of the model is tested on Emo-DB and IEMOCAP

但是本模型对于happy的识别效果较差：在IEMOCAP库中仅有24%，因为大部分的happy被识别成neutral情感；在Emo-DB库中识别率为68%，有20%的happy被识别成愤怒。上述分析结果与之前的研究结果一致[28−29]，由于训练数据的有限性以及happy类别比其他类别更依赖语境，导致happy类别很难被识别。

2.4 与现有SER实验结果对比

本文选择以下8种网络作为对比实验：基于三维注意的卷积递归神经网络(Convolutional recurrent neural networks with attention, ACRNN)[12]，并行卷积递归神经网络(Parallelized convolutional recurrent neural network, PCRN)[30]，多重卷积递归神经网络(Multiple convolution recurrent neural network, MCRN)[28]，1D和3D多特征融合网络[31]，基于卷积的胶囊神经网络(Capsule neural network based CNN, CapCNN)[32]，时间注意池网络(Attentive temporal pooling, ATP)[33]，上下文叠加拓展的卷积神经网络[34]，以及卷积循环神经网络(CNN-LSTM)[11]。结果如表3所示。

表3 与现有SER结果准确率比较

Table 3 Comparison with the accuracy of existing SER results

(单位: %)

数据集	方法	WA	UA
Emo-DB	ACRNN[12]	—	82.8
Emo-DB	PCRN[29]	86.4	84.5
Emo-DB	CapCNN[32]	—	82.
Emo-DB	ATP[33]	85.7	82.9
Emo-DB	本文模型(STA-CRNN)	86.8	84.7
IEMOCAP	CNN-LSTM[11]	64.7	56.6
IEMOCAP	MCRN[28]	67.3	62.0
IEMOCAP	1D+3D[31]	—	61.2
IEMOCAP	DiCCOSER-CS[34]	65.8	56.7
IEMOCAP	本文模型(STA-CRNN)	69.4	65.5

通过表3可以看出，在Emo-DB库中，本模型比ACRNN[12]的UA提高了1.9%；比PCRN[29]模型的WA和UA分别高了0.4%、0.2%；而与最近较为火热的CapCNN[32]、ATP[33]对比，也体现了一定的优势，两者UA皆是提高了1.8%。在IEMOCAP库中，本文模型比传统CNN-LSTM[11]的WA增长了4.7%，而UA的增幅高达8.9%；和MCRN[28]

相对比也表现出了更好的识别率，WA 和 UA 提升达 2.1%、3.5%；在和最新的一些的方法对比中可以得出比 1D+3D 网络 [30] 的 UA 提高了 4.3%；和 DiCCOSER-CS[34] 相比，本文模型的 WA 提高了 3.6%，而 UA 的提升幅度高达 8.8%。综上所述，本文所提出的模型优于大多数先进方法。

3 结论

本文提出了一种用于 SER 的 STA-CRNN 模型。该模型包含 CNN、LSTM 两大模块。分别在 CNN 和 LSTM 网络中加入了空间注意力机制和时间注意力机制，以便更好地提高模型性能，从而提高语声情感识别率。从两个情感数据集中的实验结果以及在和其他先进的方法对比中可以得出，本文的模型可以更好地提取语声频谱图中的有效特征信息，过滤掉无效特征信息，使得 SER 的识别率大幅度提高。由于本文所提取的特征是类似于图像的 RGB 三通道结构，而通道与通道之间的重要性不同，故也会影响卷积过程中特征的提取。因此，在未来的研究中，本文会在 CNN 中加入通道注意力机制以提高 SER 的效果。

参 考 文 献

[1] Mustaqeem, Kwon S. Att-Net: enhanced emotion recognition system using lightweight self-attention module[J]. Applied Soft Computing, 2021, 102(4): 107101.

[2] Abbaschian B J, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models[J]. Sensors, 2021, 21(4): 1249.

[3] Wani T M, Gunawan T S, Qadri S A A, et al. A comprehensive review of speech emotion recognition systems[J]. IEEE Access, 2021, 9: 47795–47814.

[4] Zhang J, Xing L, Tan Z, et al. Multi-head attention fusion networks for multi-modal speech emotion recognition[J]. Computers & Industrial Engineering, 2022, 168: 108078.

[5] Liu L, Wang S, Hu B, et al. Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition[J]. Pattern Recognition, 2018, 81: 545–561.

[6] Li S, Xing X, Fan W, et al. Spatiotemporal and frequential cascaded attention networks for speech emotion recognition[J]. Neurocomputing, 2021, 448(2): 238–248.

[7] Xu M, Zhang F, Cui X, et al. Speech emotion recognition with multiscale area attention and data augmentation[C]. International Conference on Acoustics, Speech and Signal Processing, 2021.

[8] Mba A, Ko B. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers-ScienceDirect[J]. Speech Communication, 2020, 116: 56–76.

[9] Chiba Y, Nose T, Ito A. Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition[C]. Interspeech, 2020.

[10] Li D, Liu J, Yang Z, et al. Speech emotion recognition using recurrent neural networks with directional self-attention[J]. Expert Systems with Applications, 2021, 173: 114683.

[11] Tzirakis P, Zhang J, Schuller B. End-to-end speech emotion recognition using deep neural networks[C]. Speech and Signal Processing, 2018.

[12] Chen M, He X, Jing Y, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters, 2018, 25(10): 1440–1444.

[13] Mao Q, Dong M, Huang Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. IEEE Transactions on Multimedia, 2014, 16(8): 2203–2213.

[14] Senthilkumar N, Karpakam S, Devi M G, et al. Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks[J]. Materials Today: Proceedings, 2022, 57: 2180–2184.

[15] Khalil R A, Jones E, Babar M I, et al. Speech emotion recognition using deep learning techniques: a review[J]. IEEE Access, 2019, 7: 117327–117345.

[16] Bakhshi A, Harimi A, Chalup S. CyTex: transforming speech to textured images for speech emotion recognition[J]. Speech Communication, 2022, 139: 62–75.

[17] Chen Q, Huang G. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition[J]. Engineering Applications of Artificial Intelligence, 2021, 102: 104277.

[18] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]. IEEE International Conference on Acoustics, 2016.

[19] Zhang S, Zhang S, Huang T, et al. Learning affective features with a hybrid deep model for audio–visual emotion recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(10): 3030–3043.

[20] 徐华南, 周晓彦, 姜万, 等. 基于自身注意力时空特征的语音情感识别算法 [J]. 声学技术, 2021, 40(6): 807–814.
Xu Huanan, Zhou Xiaoyan, Jiang Wan, et al. Speech emotion recognition algorithm based on self-attention spatiotemporal features[J]. Technical Acoustics, 2021, 40(6): 807–814.

[21] Yu Y, Kim Y J. Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database[J]. Electronics(Basel), 2020, 9(5): 713.

[22] Li P, Yan S, Mcloughlin I, et al. An attention pooling based representation learning method for speech emotion

recognition[C]. Interspeech, 2018.

[23] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]. European Conference on Computer Vision, 2018.

[24] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]. Interspeech, 2005.

[25] Busso C, Bulut M, Lee C, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335–359.

[26] 戴妍妍, 金赟, 马勇, 等. 基于高效通道注意力机制的语音情感识别方法 [J]. 信号处理, 2021, 37(10): 1835–1842.
Dai Yanyan, Jin Yun, Ma Yong, et al. Speech emotion recognition based on efficient channel attention[J]. Journal of Signal Processing, 2021, 37(10): 1835–1842.

[27] Schuller B W, Batliner A, Bergler C, et al. Computational paralinguistics challenge: elderly emotion, breathing & masks[C]. Interspeech, 2020.

[28] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms[C]. Interspeech, 2017.

[29] Ma X, Wu Z, Jia J, et al. Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space[C]. Interspeech, 2017.

[30] Jiang P, Fu H, Tao H, et al. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition[J]. IEEE Access, 2019, 7: 90368–90377.

[31] 徐华南, 周晓彦, 姜万, 等. 基于 3D 和 1D 多特征融合的语音情感识别算法 [J]. 声学技术, 2021, 40(4): 496–502.
Xu Huanan, Zhou Xiaoyan, Jiang Wan, et al. Speech emotion recognition algorithm based on 3D and 1D multi-feature fusion[J]. Technical Acoustics, 2021, 40(4): 496–502.

[32] Wen X C, Liu K H, Zhang W M, et al. The application of capsule neural network based CNN for speech emotion recognition[C]. International Conference on Pattern Recognitio, 2021.

[33] Xia X, Jiang D, Sahli H. Learning salient segments for speech emotion recognition using attentive temporal pooling[J]. IEEE Access, 2020, 8: 151740–151752.

[34] Tang D, Kuppens P, Geurts L, et al. Adieu recurrence? End-to-end speech emotion recognition using a context stacking dilated convolutional network[C]. European Signal Processing Conference, 2021.