本文引用格式：钟善机,张学习,陈楚嘉,等.基于多尺度卷积和多头自注意力的语音情感识别模型[J].自动化与信息工程,2024,45(4):36-41;49.

ZHONG Shanji, ZHANG Xuexi, CHEN Chujia, et al. Speech emotion recognition model based on multi-scale convolution and multi-head self-attention[J]. Automation & Information Engineering, 2024,45(4):36-41;49.

基于多尺度卷积和多头自注意力的语音情感识别模型*

钟善机　张学习　陈楚嘉　高学秋　陶杰

（广东工业大学，广东 广州 510006）

摘要：针对传统的卷积神经网络在语音情感识别中无法充分捕捉时域和频域细节信息的问题，提出一种基于多尺度卷积和多头自注意力（MCNN-MHA）的语音情感识别模型。首先，通过多尺度卷积神经网络在不同尺度下对输入进行卷积操作，获得不同时域和频域上的特征；然后，引入多头自注意力机制自动学习语音信号中相关和重要的特征，并关注不同特征的子空间，增强重要特征的感知能力；最后，利用 SpecAugment 中的频域掩码和时域掩码来增强数据样本，提高模型的泛化性和鲁棒性。实验结果表明，MCNN-MHA 模型在 RAVDESS 数据集上取得了 90.35%的准确率。

关键词：语音情感识别；多尺度卷积神经网络；多头自注意力机制；SpecAugment

中图分类号：TN912.3　　**文献标志码：**A　　**文章编号：**1674-2605(2024)04-0006-07

DOI：10.3969/j.issn.1674-2605.2024.04.006　　开放获取

Speech Emotion Recognition Model Based on Multi-scale Convolution and Multi-head Self-attention

ZHONG Shanji　ZHANG Xuexi　CHEN Chujia　GAO Xueqiu　TAO Jie

(Guangdong University of Technology, Guangzhou 510006, China)

Abstract: A speech emotion recognition model based on multi-scale convolution and multi head self attention (MCNN-MHA) is proposed to address the problem of traditional convolutional neural networks being unable to fully capture temporal and frequency domain details in speech emotion recognition. Firstly, a multi-scale convolutional neural network is used to convolve the input at different scales, obtaining features in different time and frequency domains; Then, a multi head self attention mechanism is introduced to automatically learn relevant and important features in speech signals, and to focus on the subspaces of different features to enhance the perception ability of important features; Utilize the frequency domain mask and time domain mask in SpecAugment to enhance data samples and improve the generalization and robustness of the model. The experimental results showed that the MCNN-MHA model achieved an accuracy of 90.35% on the RAVDESS dataset.

Keywords: speech emotion recognition; multi-scale convolution neural network; multi-head self-attention mechanism; SpecAugment

0　引言

语音情感识别（speech emotion recognition, SER）用于理解和识别人类语音中的情绪信息[1-2]，是人机交互、人工智能等领域的重要研究课题。

早期的 SER 主要采用支持向量机（support vector machine, SVM）、K 最近邻（K-nearest neighbor, KNN）算法和高斯混合模型等传统机器学习方法，人工提取特征来进行分类[3-4]。陈闯等[5]提出一种改进的灰狼优化算法来优化 SVM 的分类模型，可有效提高模型的识别准确率。BHOYAR 等[6]将融合的音调、过零率和梅尔频率倒谱系数等特征输入到 KNN 分类模型中，

* 基金项目：国家自然科学基金项目（62276069）

在 EMO-DB 数据集上，模型的识别准确率达到 86.02%。SUN 等[7]提出一种基于决策树 SVM 的 SER 算法，不仅可以挖掘语音信号的深度情感信息，还可以从易混淆的情感信息中提取更具区别性的特征。但这些传统算法依赖人工提取特征，需要相关人员具有丰富的专业知识和经验，且主观性较强。

近年来，随着深度学习技术的发展，其在 SER 领域逐渐取代传统算法。深度学习利用神经网络的多层结构，自动识别学习高级特征，无需人工处理[8]。XU 等[9]在深度卷积神经网络中利用多尺度区域注意力来关注不同粒度的情感特征，并利用声道长度扰动进行数据增强，从而提高分类器的泛化能力。ZHAO 等[10]结合并行卷积网络、压缩与激励网络，联合连接主义时序分类损失函数提出的模型在IEMOCAP数据集上的 SER 准确率达 73.1%。ATILA 等[11]提出一种基于 3D CNN+LSTM 的模型，并使用注意力机制进行指导，在 SAVEE 数据集上模型的识别准确率达 87.5%。戴妍妍等[12]构建基于高效通道注意力（efficient channel attention, ECA）的 CNN_ECA 模型，有选择地强调通道图，改进特定情感的表达。刘方如等[13]设计一种基于 Resnet 和 Transformer 结构的 Res-Transformer 模型，利用 Resnet 处理空间信息，Transformer 处理时间信息，在 RAVDESS 数据集上模型的识别准确率为 84.89%。KAKUBA 等[14]提出基于注意力的多任务模型，考虑了语音特征之间的空间线索，结合自注意力机制，关注相关特征并学习其对情感识别任务的重要性，在 RAVDESS 数据集上模型的识别准确率达 85.34%。综上所述，深度学习与注意力机制相结合的方法在 SER 方面取得了一定的进展，但仍然存在一些问题，如传统卷积神经网络可能存在特征提取不足，以及小样本数据集导致的过拟合等问题。

本文结合多尺度卷积神经网络（multi-scale convolutional neural network, MCNN）与多头自注意力机制（multi-head attention, MHA），提出 MCNN-MHA 模型，进一步提升了 SER 性能。MCNN 能够从频谱图的不同维度提取特征，捕捉丰富的时域和频域信息；MHA 使 MCNN-MHA 模型关注语音信号中相关和重要的特征。为避免小样本导致的过拟合问题，采用 SpecAugment 对 Log-mel 谱图进行数据增强。

1　模型设计

本文提出的 MCNN-MHA 模型由多尺度卷积层、深度卷积层、多头自注意力层和全连接层组成，结构如图 1 所示。

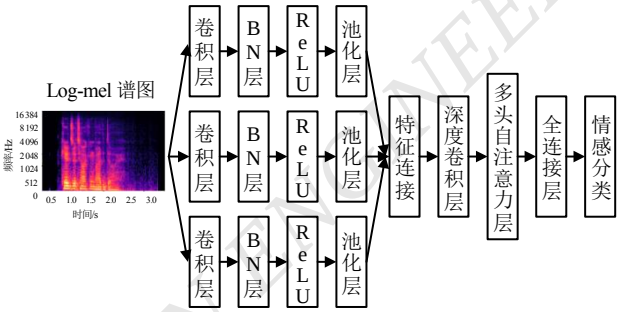

图 1　MCNN-MHA 模型结构

首先，将从原始音频中提取的 Log-mel 谱图特征输入到 MCNN-MHA 模型，经过由 3 个并行卷积组成的多尺度卷积层分别提取 Log-mel 谱图的频域、时域和局部特征。

然后，将频域、时域和局部特征连接成一个特征，输入到深度卷积层。深度卷积层通过堆叠多个卷积层来进一步提取高级特征，使 MCNN-MHA 模型更好地区分不同的语音情感类别。

接着，将高级特征输入到多头自注意力层。多头自注意力层引入了自注意力机制，并采用多个注意力头，使 MCNN-MHA 模型可以同时考虑不同的特征子空间，并学习到更具区分性的特征表示。

最后，通过全连接层，完成语音情感分类识别。

1.1　多尺度卷积层

文献[15]研究表明，将 CNN 应用于分类任务的特征提取时，可通过增大感受野来提高分类精度。为此，MCNN-MHA 模型采用 3 个并行的 CNN 来提取 Log-mel 谱图的多维特征。其中，卷积核为 9×1 的 Conv1_a 用于捕捉 Log-mel 谱图的频域特征；卷积核为 1×13 的 Conv1_b 用于捕捉 Log-mel 谱图的时域特征；卷积核为 3×3 的 Conv1_c 用于捕捉 Log-mel 谱图的局部

特征。这有助于 MCNN-MHA 模型更好地理解语音信号的结构和特征，提高语音情感分类识别的准确率。多尺度卷积层的结构及参数如表 1 所示。

表 1　多尺度卷积层的结构及参数

卷积层名称	网络结构及参数
Conv1_a	Conv2D: 9×1, 8 BN+ReLU MaxPooling: 2×2 Dropout: 0.3
Conv1_b	Conv2D: 1×13, 8 BN+ReLU MaxPooling: 2×2 Dropout: 0.3
Conv1_c	Conv2D: 3×3, 8 BN+ReLU MaxPooling: 2×2 Dropout: 0.3

1.2　深度卷积层

深度卷积层由 4 个卷积块串联组成。每个卷积块由二维卷积层（Conv2D）、批量归一化（batch normalization, BN）层、ReLU 激活函数、最大池化（Max Pooling）层和 Dropout 层组成，其结构及参数如表 2 所示。

表 2　深度卷积层的结构及参数

卷积层名称	网络结构及参数
Conv2	Conv2D: 3×3, 32 BN+ReLU MaxPooling: 2×2 Dropout: 0.3
Conv3	Conv2D: 3×3, 64 BN+ReLU MaxPooling: 2×2 Dropout: 0.3
Conv4	Conv2D: 3×3, 128 BN+ReLU MaxPooling: 2×2 Dropout: 0.3
Conv5	Conv2D: 3×3, 128 BN+ReLU MaxPooling: 2×2 Dropout: 0.3

深度卷积层通过堆叠多个卷积块来提取 Log-mel 谱图的高级特征。其中，BN 层对 Conv2D 的输出进行批量归一化处理，通过调整数据分布来加速网络训练，提高 MCNN-MHA 模型的泛化能力；ReLU 激活函数对 Conv2D 的输出进行非线性变换，引入非线性特征可增加 MCNN-MHA 模型的表达能力；Max-Pooling 层可降低特征图的空间维度，减少参数的数量和计算量，且具有一定的平移不变性；Dropout 层以指定的概率随机丢弃部分神经元，以减小 MCNN-MHA 模型过拟合风险，提高模型的泛化能力。

1.3　多头自注意力层

自注意力机制是一种通过线性变换将输入序列 $X=\left(x_1,x_2,\ldots,x_m\right)$ 的元素投影到查询（Query）、键（Key）和值（Value）3 个不同表示上的方法[16]。定义查询表示为 $Q\in\mathbb{R}^{m\times d_m}$，键表示为 $K\in\mathbb{R}^{m\times d_m}$，值表示为 $V\in\mathbb{R}^{m\times d_m}$。自注意力通过对查询、键和值进行缩放点积，得到

$$A_{\text{Attention}}\left(Q,K,V\right)=S_{\text{softMax}}\left(\frac{QK^{\text{T}}}{\sqrt{d_m}}\right)V \tag{1}$$

多头自注意力与一次使用 d_m 维查询、键和值进行自注意力计算不同，其可以并行 h 次自注意力计算，其中每次查询、键、值的维度为 $d_h=d_m/h$。这样可以同时关注来自不同表示子空间的信息，将每个注意力头的计算结果拼接，得到最后的输出。

通过引入多头自注意力机制，MCNN-MHA 模型可以关注不同类别的语音信号，如情感表达的声调、语速、重音等，从而提高语音情感识别任务的性能和准确率。

2　数据集与特征提取

2.1　实验数据集

本实验采用的语音情感数据集为 RAVDESS[17]。该数据集由 24 名北美口音的演员（12 男，12 女）录制，共有 1 440 个语音情感样本，包括中性、快乐、悲伤、愤怒、恐惧、厌恶、惊讶和平静 8 类情感。其中，除了中性情感类别样本有 96 个外，其他类别样本均为 192 个，分布较为均匀。

2.2 预处理与特征提取

本实验采用的特征是 Log-mel 谱图，与原始的频谱图相比，其可以准确地提取感兴趣的情感特征，减少特征空间的维度[18]；能够减少可能发生在频带中的干扰，并模拟人耳对声音的非线性感知[19]；可以提高分类模型的训练速度和识别速度。Log-mel 谱图的提取流程如图 2 所示。

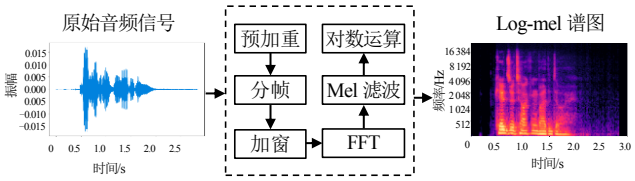

图 2 Log-mel 谱图提取流程

Log-mel 谱图的提取流程如下：

1） 将原始的音频信号进行预加重处理，通过高通滤波器来增强高频分量，以补偿声音在传播过程中损失的高频信息，提高高频分辨率；

2） 采用固定长度的窗口（如 25 ms）将预加重的语音信号分成短时帧，为减少信号失真并确保频谱平滑覆盖，一般采用汉明窗函数作为窗口函数；

3） 利用快速傅里叶变换（FFT），将每个分帧和加窗后的语音信号从时域转为频域，得到每个帧的频谱表示；

4） 计算每帧频谱信号的功率谱，即频谱的振幅平方，并将功率谱通过一组 Mel 滤波器映射到 Mel 尺度上，用于测量不同频率范围内的能量；

5） 对 Mel 滤波器组的输出进行对数转换，将线性刻度的滤波器输出转换为对数刻度，以符合人类对声音的感知，使特征表示更具有可解释性和鲁棒性。

2.3 数据增强

在训练深度学习模型时，数据量较小可能会导致训练模型过拟合。通过数据增强来增加现有数据的有效性是避免过拟合的可行方法。在语音情感识别领域，传统的数据增强方法主要包括对原有样本的声波加噪音、改变音调、改变语速等，但这些方法会增加计算时间。PARK 等[20]提出一种在线数据增强方法 SpecAugment，其直接对频谱图进行在线增强，方法简单，且计算代价较小。

本文采用 SpecAugment 两种常用的策略对数据进行增强：1） 时域掩码，在频谱图的时间轴上随机将一段连续的时域区域置零；2） 频域掩码，在频谱图的频域轴上随机将一段连续的频域区域置零。时域掩码和频域掩码后的 Log-mel 谱图如图 3 所示。

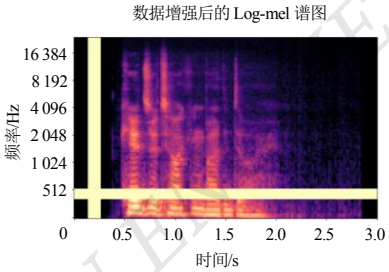

图 3 时域掩码和频域掩码后的 Log-mel 谱图

3 实验过程与分析

3.1 实验设置

在 RAVDESS 数据集上采用随机划分的 10-fold 交叉验证，训练集与测试集的比例为 9∶1。在 MCNN-MHA 模型训练过程中，选择 Adam 优化器，学习率设置为 0.001，权重衰减设置为 0.000 1。迭代次数设置为 400 次，多头自注意力层的头数为 8 头，损失函数采用交叉熵函数。为防止过拟合，每次卷积后的 Dropout 设置为 0.3。此外，批量化大小设置为 64 组，用于控制每次迭代的样本数量。

本文采用准确率（Accuracy）、精确率（Precision）和召回率（Recall）作为实验的评价指标。

实验的软硬件环境如表 3 所示。

表 3 软硬件实验环境

硬件	软件
CPU: E5-2620 v4 @ 2.10GHz	Pytorch 1.10.1
GPU: NVIDIA GeForce GTX 1080 Ti	CUDA 10.2
内存: 32 GB	Librosa 0.10.1
显存: 12 GB	Python 3.7.11
OS: Ubuntu 18.04	Anaconda3

3.2 消融实验

3.2.1 数据增强对实验结果的影响

采用时域掩码和频域掩码对原始数据集的 Log--mel 谱图进行数据增强。为验证数据增强的有效性，将原始 Log-mel 谱图和数据增强后的 Log-mel 谱图分别输入到 MCNN-MHA 模型，结果如表 4 所示。

表 4 数据增强对实验结果的影响

输入特征	准确率/%
原始 Log-mel 谱图	88.68
数据增强后的 Log-mel 谱图	90.35

由表 4 可知，相较于原始的 Log-mel 谱图，数据增强后的 Log-mel 谱图，准确率提升了 1.67%，说明数据增强能够减少过拟合，提升了模型性能。

3.2.2 多头自注意力对实验结果的影响

为验证多头自注意力的有效性，将数据增强后的 Log-mel 谱图分别输入到 MCNN 和 MCNN-MHA 模型，结果如表 5 所示。

表 5 多头自注意力对实验结果的影响

模型	准确率/%
MCNN	86.74
MCNN-MHA	90.35

由表 5 可知，相较于 MCNN 模型，MCNN-MHA 模型的准确率提升了 3.61%，说明多头自注意力能够学习不同的关注权重，提高了模型的表达能力，可以更好地理解和表示输入。

3.3 模型性能评估

为进一步分析 MCNN-MHA 模型的性能，将其在 RAVDESS 数据集上进行 10-fold 交叉验证，实验结果如表 6 所示。

由表 6 可知，MCNN-MHA 模型在 RAVDESS 数据集的平均准确率达 90.35%，平均精确率达 90.45%，平均召回率达 90.43%，表明模型可以很好地识别和分类音频中的情感。

表 6 10-fold 实验结果 %

实验次数	召回率	精确率	准确率
1	91.78	93.36	92.36
2	93.92	93.27	93.06
3	89.29	89.13	89.58
4	89.42	90.31	89.58
5	86.92	87.41	86.81
6	93.82	93.81	93.06
7	90.39	88.31	90.28
8	90.32	90.28	89.58
9	92.40	93.09	93.06
10	86.02	85.57	86.11
均值	90.43	90.45	90.35

第 7 次实验（最接近均值）的 Acc 曲线图、Loss 曲线图分别如图 4、5 所示。

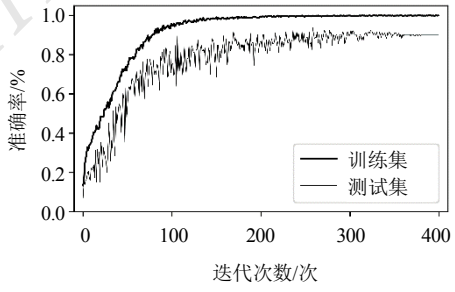

图 4 实验次数为 7 时的 Acc 曲线图

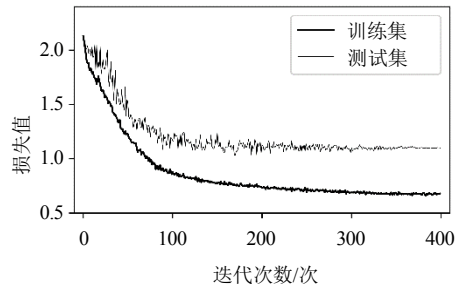

图 5 实验次数为 7 时的 Loss 曲线图

由图 4、5 可知，随着迭代次数的增加，准确率和损失值的变化幅度逐渐减小，并最终趋于稳定。在训练集上，损失值约为 0.6；在测试集上，损失值约为 1.1。同时 MCNN-MHA 模型表现出较好的拟合能力，无严重的过拟合现象，能够在未使用的数据上保持良

好的性能。

将 MCNN-MHA 模型与文献[13]、[14]、[21]的模型进行对比实验，结果如表 7 所示。

表 7　本文模型与其他研究成果的性能对比

模型	准确率/%	划分比例
文献[13]	84.89	8∶1∶1
文献[14]	85.34	10-fold
文献[21]	85.82	10-fold
MCNN-MHA	90.35	10-fold

由表 7 可知，MCNN-MHA 模型在 RAVDESS 数据集上相比文献[13]、[14]、[21]3 种模型的准确率有所提升，相较于文献[13]、[21]的模型分别提升了 5.46%、4.53%，证明了本文提出的 MCNN-MHA 模型的有效性。

4　结论

本文提出一种基于多尺度卷积和多头自注意力的（MCNN-MHA）模型，相较于传统的单个卷积神经网络，多尺度卷积神经网络可以同时提取 Log-mel 谱图上多个维度的特征，充分利用时域和频域的细节信息。同时，引入多头自注意力机制使模型能够更准确地捕捉语音信号中的情感特征。此外，还利用频域掩码和时域掩码来增强数据样本，提高模型的泛化性和鲁棒性。在 RAVDESS 数据集上进行 10-fold 实验，验证了本文提出的 MCNN-MHA 模型的有效性。

参考文献

[1] AYADI M E, KAMEL M S, KARRAY F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011,44(3):572-587.

[2] SWAIN M, ROUTRAY A, KABISATPATHY P. Databases, features and classifiers for speech emotion recognition: A review[J]. International Journal of Speech Technology, 2018,21(1):93-120.

[3] RAMESH S, GOMATHI S, SASIKALA S, et al. Automatic speech emotion detection using hybrid of gray wolf optimizer and Naïve Bayes[J]. International Journal of Speech Technology, 2023,26(3):571-578.

[4] AL DUJAILI M J, EBRAHIMI-MOGHADAM A, FATLAWI A. Speech emotion recognition based on SVM and KNN classifications fusion[J]. International Journal of Electrical and Computer Engineering, 2021,11(2):1259.

[5] 陈闯, CHELLALI R,邢尹.改进 GWO 优化 SVM 的语音情感识别研究[J].计算机工程与应用,2018,54(16):113-118.

[6] BOMBATKAR A, BHOYAR G, MORJANI K, et al. Emotion recognition using speech processing using K-nearest neighbor algorithm[J]. International Journal of Engineering Research and Applications, 2014,4:68-71.

[7] SUN L, ZOU B, FU S, et al. Speech emotion recognition based on DNN-decision tree SVM model[J]. Speech Communication, 2019,115:29-37.

[8] DEY A, CHATTOPADHYAY S, SINGH P K, et al. A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition[J]. IEEE Access, 2020,8:200953-200970.

[9] XU M, ZHANG F, CUI X, et al. Speech emotion recognition with multiscale area attention and data augmentation[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6319-6323.

[10] 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.

[11] ATILA O, ŞENGÜR A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition[J]. Applied Acoustics, 2021,182:108260.

[12] 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.

[13] 刘方如,王亮.基于 Res-Transformer 模型的语音情感识别方法研究[J].物联网技术,2023,13(6):36-39;44.

[14] KAKUBA S, POULOSE A, HAN D S. Attention-based multi-learning approach for speech emotion recognition with dilated convolution[J]. IEEE Access, 2022,10:122302-122313.

[15] ARAUJO A, NORRIS W, SIM J. Computing receptive fields of convolutional neural networks[J]. Distill, 2019,4(11):e21.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017,30.

（下转第 49 页）