计算机科学
COMPUTER SCIENCE

改进MFCC和并行混合模型的语音情感识别

崔琳，崔晨露，刘政伟，薛凯

引用本文

崔琳，崔晨露，刘政伟，薛凯.改进MFCC和并行混合模型的语音情感识别[J]．计算机科学, 2023, 50(6A): 220800211-7.

CUI Lin, CUI Chenlu, LIU Zhengwei, XUE Kai. Speech Emotion Recognition Based on Improved MFCC and Parallel Hybrid Model [J]. Computer Science, 2023, 50(6A): 220800211-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800211

改进 MFCC 和并行混合模型的语音情感识别

崔　琳[1,2]　崔晨露[1]　刘政伟[1]　薛　凯[1]
1 西安工程大学电子信息学院　西安 710600
2 西北工业大学航海学院　西安 710072
(cuilin789@163.com)

摘　要　传统 MFCC 不仅忽略了浊音信号中基音频率的影响,还不能表征语音的动态特征,因此提出利用滑动平均滤波器滤除浊音信号的基音频率,并在提取完静态 MFCC 特征后再通过提取其一阶差分与二阶差分来获取动态特征。将得到的特征送入模型中进行训练,为了构建更高效的语音情感识别模型,搭建了一种融合多头注意力机制的并行混合模型。多头注意力机制不仅可以有效防止梯度消失现象,构建更深层的网络,各个注意力头还可以执行不同的任务来提高准确率。最后进行情感特征分类,传统 softmax 在进行分类时类内距离可能会变大导致模型的置信度差,因此引入了中心损失函数,将两者联合来进行分类。实验结果表明,所提方法在 RAVDESS 数据集和 EMO-DB 数据集上的准确率可以分别达到 98.15%和 96.26%。

关键词:语音情感识别;MFCC;多头注意力机制;滑动平均滤波器;softmax
中图法分类号　TP183

Speech Emotion Recognition Based on Improved MFCC and Parallel Hybrid Model

CUI Lin[1,2],CUI Chenlu[1],LIU Zhengwei[1] and XUE Kai[1]
1 School of Electronic Information,Xi'an Polytechnic University,Xi'an 710600,China
2 School of Navigation,Northwestern Polytechnical University,Xi'an 710072,China

Abstract　The traditional MFCC not only ignores the influence of the pitch frequency in the voiced signal,but also cannot characterize the dynamic characteristics of the speech. Therefore,a moving average filter is proposed to filter out the pitch frequency of the voiced signal. After extracting the static MFCC features,the dynamic features are obtained by extracting their first-order difference and second-order difference. The obtained features are sent to the model for training. To construct a more efficient speech emotion recognition model,a parallel hybrid model integrating a multi-head attention mechanism is built. The multi-head attention mechanism can not only effectively prevent the gradient disappearance phenomenon from constructing a deeper network,but also perform different tasks to improve the accuracy. Finally,when classifying emotional features,the traditional softmax may increase the intra-class distance during classification,resulting in poor confidence in the model. Therefore,the center loss function is introduced to combine the two for classification. Experimental results show that the accuracy of the proposed method can reach 98.15 % and 96.26 % on the RAVDESS dataset and EMO-DB dataset,respectively.

Keywords　Speech emotion recognition,MFCC,Multi-head attention mechanism,Moving average Filter,softmax

随着科技的不断发展,人机交互技术也变得越来越成熟,但是现在的机器还不能很好地感知到人类的情感[1]。因此,语音情感识别成为了现在的研究热点。但目前该技术还存在两大难点:1)有效的语音情感特征的寻找;2)适当的情感识别模型的建立。

较早的语音情感识别研究通常采用传统的声学特征进行实验,声学特征大致分为 3 类:韵律学特征[2]、基于谱的相关特征[3]和声音质量特征[4]。Ververidis 等[5]从基音、能量以及语音频谱的动态行为中提取出 87 个静态特征,并对性别和情感进行了层次分类。Schuller 等[6]使用韵律特征的统计特征进行情感识别,取得了 86.8%的识别率。除了使用传统的声学特征外,研究人员还不断推陈出新,发掘了更具表现力的新型特征。2019 年,Liu 等[7]在 Gammatone 频率倒谱系数(Gammatone Frequency Cepstral Coefficients,GFCC)的基础上改进,提出了 VGFCC 特征,并在 EMO-DB 和 TYUT2.0 两个数据库上取得了 91.10%和 72.15%的识别率。Mel 倒谱系数(Mel Frequency Cepstral Coefficients,MFCC)是一种基于谱的语音情感特征值,但是传统的 MFCC 一方面只能反应语音信号的静态特征,另一方面,它在浊音信号时混入了基音频率进而影响了准确率。为此,本文提出了一种改进 MFCC。

使用传统方法进行语音情感识别不仅较为复杂且准确率较低,因此近些年越来越多的研究人员将深度学习应用到语音

基金项目:国家自然科学基金青年项目(61901347)
This work was supported by the Young Fund of the National Natural Science Foundation of China(61901347).
通信作者:崔晨露(1017071182@qq.com)

情感识别中。最常用的方法是卷积神经网络(Convolutional Neural Network,CNN)和循环神经网络(Recurrent Neural Network,RNN)。Mao 等[8]利用 CNN 从整个帧序列中提取最显著的帧来表示话语,但是 CNN 存在不能提取时序信息的问题,因此 Lee 等[9]采用 RNN 学习语音的时序特征。为了避免在长序列训练过程中产生梯度消失现象,Verkholyak 等[10]利用长短时记忆网络(Long and Short Time Memory network,LSTM)获得长期声学特征,为进一步优化 LSTM,门控循环单元(GRU)[11]网络被提出,其结构更加简单,效果也很好。近些年,越来越多的新型模型被提出。Li 等[12]提出了 BLSTM 和 CNN 堆栈架构来增强情感识别能力,为解决过拟合问题,利用 K-fold 交叉验证将 BLSTM 和 CNN 对概率量的估计合并为新数据。Chen 等[13]提出了双注意力的双向长短期记忆网络来识别语音情感。针对目前模型识别率较低的问题,本文提出了一种融合多头注意力机制的并行混合模型。该模型引入了 transformer 的多头注意力机制部分,不仅可以有效防止梯度消失现象,构建更深层的网络,各个注意力头还可以执行不同的任务。构建并行混合网络不仅可以提高训练速度还可以使得到的特征更加丰富。

1 改进的 MFCC 方法

MFCC 是基于谱的语音情感特征值[14],描述的是能够表现说话人声道特征的谱包络。在浊音信息方面,其不但包括谱包络还包括基音频率,基音频率会影响对声道特征的描述,从而影响识别率。人在不同情感下的语言信号非平稳性非常明显,但是传统的 MFCC 方法只是表述了语言信息的静态特性,而无法表达语言信息的动态特性。针对这两个问题,本文通过加入滑动平均滤波器和对 MFCC 进行动态特征提取来实现对传统 MFCC 的改进。改进 MFCC 的流程框图如图 1 所示。

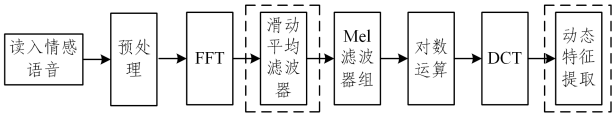


图 1 改进 MFCC 的流程框图

Fig. 1 Flow chart of improved MFCC

1.1 改进的谱包络提取

语音信号为浊音信号时,MFCC 的幅度谱中包含了谱包络和基音频率[15]。在对声道特性进行描述时基因频率会有所影响,进而影响识别率。因此,本文提出了一种在 MFCC 提取过程中谱包络提取的改进,即在信号通过 Mel 滤波器前先经过一个滑动平均滤波器,滑动平均滤波器先对浊音信号幅度谱的各谐波峰值点进行内插,得到各谐波点之间频点的新幅度谱值,从而得到一个与基音频率无关的近似谱包络,再将其输入到 Mel 滤波器组对频谱进行平滑化,并消除谐波,最后再进行对数运算和 DCT 变换得到改进后的 MFCC。

1.2 动态特征提取

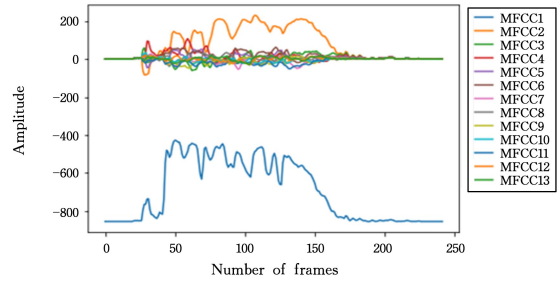
由于人耳对声信号的动态特性更为敏感,而 MFCC 只反映了语音信号的静态特征,因此在提取完 MFCC 特征后再提取其一、二阶差分。其中一阶差分表示当前语音帧与前一帧之间的关系,二阶差分表示前一阶差分与后一阶差分之间的关系,通过提取的一、二阶差分与静态 MFCC 进行拼接能够更好地反映不同情感状态的语音信号特征,得到更全面的语音信号情感特征。MFCC 的一阶差分系数(ΔMFCC)如式(1)所示:

$$D(n)=\frac{1}{\sqrt{\sum_{i=-k}^{i=k} i^2}}\sum_{i=-k}^{i=k} i\cdot C(n+i) \tag{1}$$

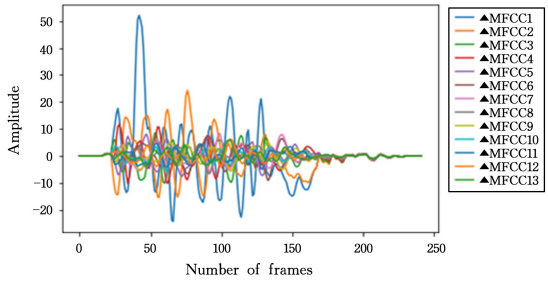
其中,$C(n+i)$是一帧 MFCC 参数,k 的值设为 2。将结果 $D(n)$代入式(2),即第二个差值,得到 MFCC 参数(Δ2 MFCC)。

$$D_2(n)=\frac{1}{\sqrt{2\sum_{i=-k}^{i=k} i^2}}\sum_{i=-k}^{i=k} i\cdot D(n+i) \tag{2}$$

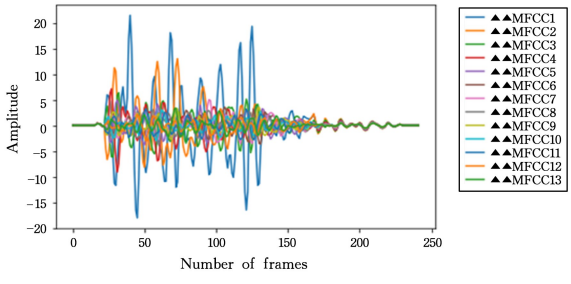
其中,$D_2(n)$是 MFCC 系数的二次差分。MFCC 的静态、一阶、二阶特征曲线图如图 2 所示。



(a)原始 MFCC 特征曲线图



(b)一阶 MFCC 特征曲线图



(c)二阶 MFCC 特征曲线图

图 2 不同阶 MFCC 特征曲线图

Fig. 2 Different-order MFCC feature curve map

2 并行混合模型框架

为提高模型的识别准确率,设计了一种融合多头注意力机制的并行混合模型,该模型大致分为 3 部分:首先将改进后的 MFCC 特征作为输入;然后送入融合多头注意力机制的并行混合模型中进一步提取更深层的特征,并行混合模型由 CNN-BiLSTM 与多头注意力机制组成,通过上下两个分支分别提取完特征后再进行拼接;最后利用带有联合损失函数的全连接层完成分类,得到最终的分类结果。该模型的整体框图如图 3 所示。

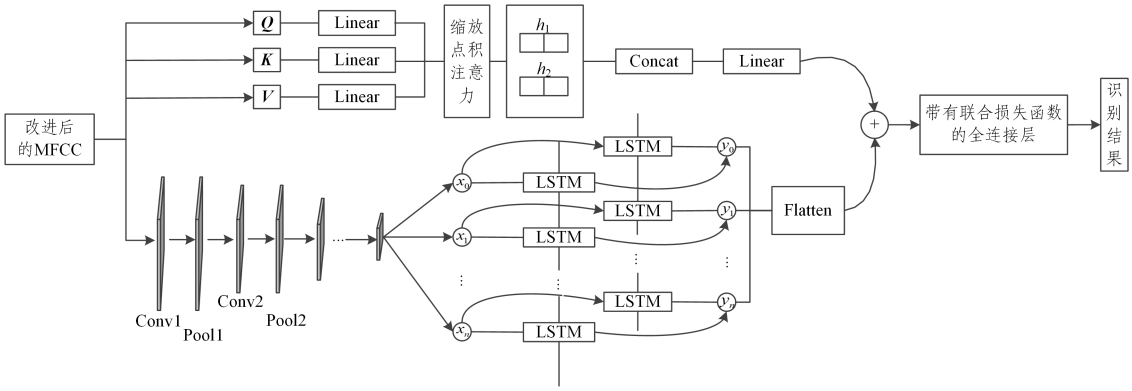


图 3　并行混合模型图

Fig. 3　Parallel mixed model graph

2.1　CNN-BiLSTM 网络

语音情感信息包括空间特征关系和时序特征关系,为此构建了 CNN-BiLSTM 模型,如图 3 并行混合模型图的下分支所示。CNN-BiLSTM 由 CNN 卷积、池化层、BiLSTM 和 Flatten 层构成。CNN 可以有效处理语音情感信息中的空间信息,通过 CNN 虽然提取了语音中的空间信息,但是无法挖掘语音的上下文信息。为使情感特征表征地更加充分,在提取完空间信息后,利用 BiLSTM 网络来获取时序信息。BiLSTM 结合前向 LSTM 与后向 LSTM,能够双向提取数据序列的特征信息,在 BiLSTM 的某个时刻,可以对当前时间节点的前后信息进行学习,相比 LSTM 可以获得更多的时序信息,得到相对于 LSTM 更细粒度的分类结果,此外 BiLSTM 还可以实现浅层特征与深层特征的融合。

2.2　多头注意力机制

BiLSTM 在长距离传播中会损失较多信息且对特征重要度不敏感,因此加入了多头注意力机制来对重要特征进行提取。多头注意力机制[16]基于自注意力模块,用来提取更具表现力的序列,同时联合多个注意力表征进行情感建模。其结构图如图 4 所示。

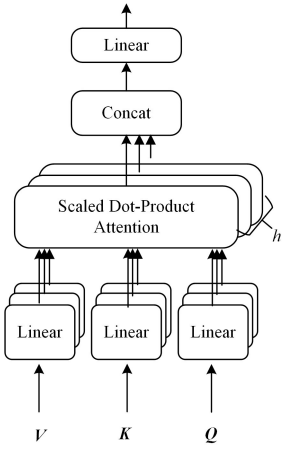


图 4　多头注意力机制的结构图

Fig. 4　Structure diagram of multi-head attention mechanism

如图 4 所示,多头注意力机制的计算是先将查询 $\boldsymbol{Q}$(query)、键 $\boldsymbol{K}$(key)、值 $\boldsymbol{V}$(value)3 个输入进行线性变换,然后计算 h 次放缩点注意力,最后将所有的输出拼接到另一个线性层获得最终结果。本文采用缩放点积注意力机制,计算式如式(3)所示:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

其中,将改进 MFCC 作为输入向量,d_k 表示维度。通过执行缩放操作,使得 softmax 函数的参数在使用较大尺寸的键时不会变得过大。

为了从不同的维度和表示子空间中学习到情感特征的相关信息,多次使用不同的参数对向量 $\boldsymbol{Q}$、矩阵 $\boldsymbol{K}$ 和矩阵 $\boldsymbol{V}$ 做一个线性变换,并将结果输入到 Attention 中,以获取不同的注意力输出。那么每个头的注意力输出 $\boldsymbol{Q}_i$ 可以通过式(4)求得。

$$\boldsymbol{Q}_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}) \tag{4}$$

其中,$\boldsymbol{W}_i^{Q}\in R^{d_{\mathrm{model}}\times d_q}$,$\boldsymbol{W}_i^{K}\in R^{d_{\mathrm{model}}\times d_k}$,$\boldsymbol{W}_i^{V}\in R^{d_{\mathrm{model}}\times d_v}$。

最后将各个头部进行合并,产生多头注意力输出,如式(5)所示:

$$MHAtt(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concatenate(\boldsymbol{O}_1,\boldsymbol{O}_2,\cdots,\boldsymbol{O}_N) \tag{5}$$

2.3　联合损失函数

语音情感识别实际上是一个多分类问题,深层特征提取完成后利用传统 softmax 进行情感特征分类时,类内的距离可能会变大,导致模型的置信度差。为此本文提出了一种联合损失函数,在传统 softmax 的基础上引入了中心损失函数,中心损失函数在学习每个类深层特征中心的同时,惩罚预期对应类中心特征,从而增大类间距离,减小类内差距,通过两者联合决策的方式,进一步提高了分类的准确度。softmax 与中心损失函数联合损失函数公式如式(6)所示:

$$\begin{aligned}L&=L_S+\lambda L_C\\&=-\frac{1}{m}\left[\sum_{i=1}^{m}\log\frac{\mathrm{e}^{\boldsymbol{W}_{y(i)}^{\mathrm{T}}x(i)}}{\sum_{i=1}^{k}\mathrm{e}^{\boldsymbol{W}_i^{\mathrm{T}}x(i)}}\right]+\frac{\lambda}{2}\sum_{i=1}^{m}\|x_i-c_{y_i}\|_2^2\end{aligned} \tag{6}$$

其中,λ 用来平衡两个损失,选择合适的 λ 有助于提高网络特征的识别能力,c_{y_i} 是 y_i 类的样本中心,当 $\lambda=0$ 时,表示仅有 softmax 损失。为验证 softmax 与中心损失函数的特性,使用 MNIST 数据集在卷积神经网络模型结构上进行实验,对本文所提出的联合损失模型进行研究,在实验环境相同的条件下,采用 softmax 交叉熵损失、联合损失函数两种不同的损失函数来对网络进行训练,得到两种不同的二维特征映射图,如图 5 所示。可以看出 softmax 可以很好地实现不同类的分类但是类内分类不够明显,本文提出的联合损失函数得到的特征映射如图 6 所示,其相比 softmax 特征之间的独立性更强,既保证了不同类别之间的特征可分离又使得类内的特征更加紧凑,表明了本文的联合损失函数可行性。

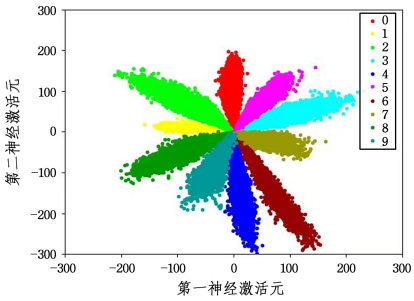


图 5 softmax 特征映射图

Fig. 5 Softmax feature map

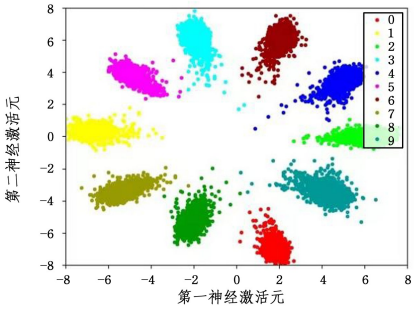


图 6 Center loss+softmax 特征映射图

Fig. 6 Center loss+softmax feature map

3 实验分析

3.1 实验设置

3.1.1 数据集

本文使用 RAVDESS[17] 和 EMO-DB[18] 数据集进行实验，其中 RAVDESS 数据集有 1 440 条语音文件，情感包括生气(angry)、中立(neutral)、平静(clam)、开心(happy)、伤心(sad)、害怕(fearful)、厌恶(disgust)、惊喜(surprise)，由 24 位演员(12 男 12 女)录制而成。EMO-DB 数据集是由柏林工业大学录制的德语情感语音库，情感包括中立(nertral)、生气(anger)、害怕(fear)、开心(happiness)、伤心(sadness)、厌恶(disgust)、无聊(boredom)，由 10 位演员(5 男 5 女)录制而成，共有 535 条语音。

3.1.2 数据扩增

要发挥神经网络的性能需要大量的数据，当数据过少时，会出现在训练集上精度很高但是验证集上精度较低的现象。有足够的数据不仅能发挥神经网络的性能而且能更好地刻画出模型在空间上的分布。为了从根本上提高模型的性能，防止神经网络学习到不相关的模式，本文在预处理阶段通过将高斯白噪声和波形位移这两种不同的传统数据增强方法相结合的方式来实现数据扩增，确保模型可以提取到更多有效信息。

3.1.3 参数设置和评估指标

将扩增后的数据集按 8∶2 的比例划分为训练集与测试集，设置训练 epoch 总数为 1000，优化器选择为 Adam。使用 F1_score、准确率和混淆矩阵对整体模型性能进行定量测量。F1 的计算公式如式(7)所示：

$$F1=\frac{2*precision*recall}{precision+recall} \tag{7}$$

3.2 实验结果与分析

3.2.1 λ 值的确定

为选择合适的 λ 值，以 CNN 模型为基准在 RAVDESS 数据集与 EMO-DB 数据集上进行测试，测试结果如图 7 所示。可以看出，本文提出的联合监督算法相比仅在 softmax 损失监督(当 $\lambda=0$ 时，训练网络仅有 softmax 损失监督)下情感识别的测试精度要高。从图 7 还可以看出，一个合适的 λ 值有助于提高情感识别精度，通过实验得出参数 λ 的最佳结果为 0.001。

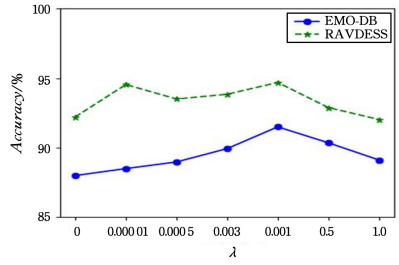


图 7 不同 λ 值下准确率曲线

Fig. 7 Accuracy curve with different λ values

3.2.2 消融实验

为证明本文模型的有效性，在 EMO-DB 和 RAVDESS 数据集上使用准确率与 F1_score 进行评估，深入评估了所提方法在不同评价指标下的预测性能。

以 CNN 作为基线模型，采用消融实验测试 CNN-BiLSTM 的实验结果，并与本文提出的模型进行对比，结果如表 1 所列。

表 1 消融实验

Table 1 Ablation experiment

(单位：%)

模型	数据集	
	RAVDESS	EMO-DB
CNN	94.68	91.48
CNN+BiLSTM	96.06	92.83
本文模型	97.34	94.39

由表 1 可以看出，本文所提模型在 RAVDESS 数据集上得到的准确率相比 CNN 和 CNN+BiLSTM 分别提升了 2.66%和 1.28%。在 EMO-DB 数据集上得到的准确率相比 CNN 和 CNN+BiLSTM 分别提升了 2.91%和 1.56%。RAVDESS 数据集上不同模型的准确率曲线对比如图 8 所示。可以看出本文模型的准确率最高，在 500 epoch 后趋于稳定；其次是 CNN+BiLSTM，该模型也约在 500 epoch 后趋于稳定；基准模型 CNN 的准确率最低，模型趋于稳定也较晚，大约在 550 epoch。EMO-DB 数据集上 3 个模型所获得的准确率曲线如图 9 所示，观察该图可发现本文模型不仅准确率相对较高且收敛地也较早。

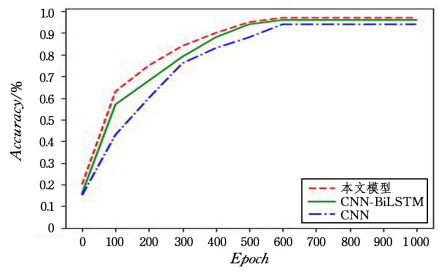


图 8 RAVDESS 数据集上的准确率曲线对比

Fig. 8 Accuracy curve comparison graph on RAVDESS dataset

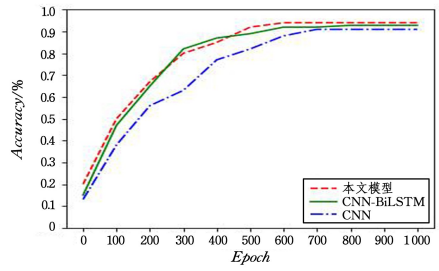


图 9　EMO-DB 数据集上的准确率曲线对比图

Fig. 9　Accuracy curve comparison graph on EMO-DB dataset

为了进一步分析本文模型在情感分类上的有效性,给出了在 RAVDESS 数据集上本文模型所得到的混淆矩阵,如图 10 所示。该图显示了真实情感和预测情感的混淆程度,并展示了每个类别中有多少情感被准确或错误地预测出来。实际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 10 可以观察到伤心类的准确率最高达到了 100%,生气类的准确率最低为 93%,生气类最易与惊喜类产生混淆,其次是平静类。

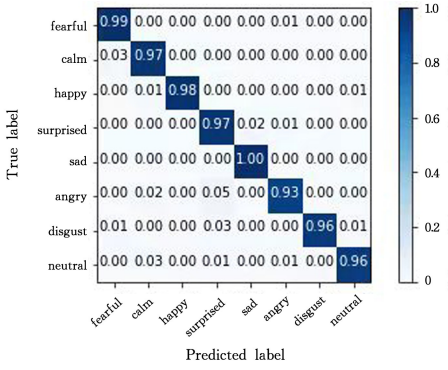


图 10　本文模型在 RAVDESS 数据集上的混淆矩阵

Fig. 10　Confusion matrix of the proposed model on RAVDESS dataset

在系统的测试过程中,给出了在 RAVDESS 数据集上 F1_score 准确度矩阵的分类报告,如表 2 所列。表中列出了该模型的强度,并使用百分比说明了总体预测准确度。由表 2 可以看出 F1_score 的平均值为 0.97,其中厌恶类与开心类的 F1_score 最高为 0.99,中立类最低为 0.94。

表 2　RAVDESS 数据集上的分类报告

Table 2　Classification report on RAVDESS dataset

	Precision	recall	F1-score
angry	0.97	0.99	0.98
calm	0.96	0.97	0.97
disgust	1.00	0.98	0.99
fearful	0.94	0.97	0.96
happy	0.98	1.00	0.99
neutral	0.95	0.93	0.94
sad	1.00	0.96	0.98
surprised	0.98	0.96	0.97
accuracy			0.97

在 EMO-DB 数据集上本文模型所得到的混淆矩阵如图 11 所示。由该图可以看出,中立类的识别率最高为 1,而无聊类最低,只有 0.90。还观察到无聊类易与伤心类产生混淆,高兴类的准确率也相对较低,其不仅易与伤心类产生混淆还易与厌恶类产生混淆。EMO-DB 数据集上的分类报告如表 3 所列。由该表可以看出本文模型在 EMO-DB 数据集上的平均 F1_score 为 0.96,与 RAVDESS 数据集上得到的分类报告相比 F1_score 较低,观察该表可知,生气类和高兴类的 F1_score 最高为 0.99,无聊类和中立类的 F1_score 最低为 0.92。

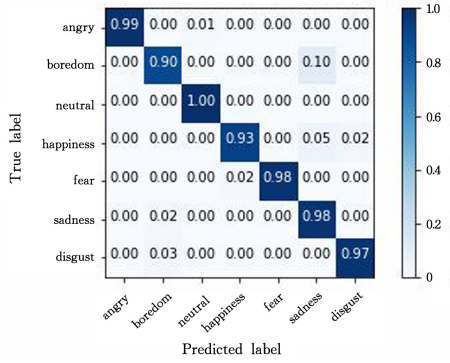


图 11　本文模型在 EMO-DB 数据集上的混淆矩阵

Fig. 11　Confusion matrix of the proposed model on EMO-DB dataset

表 3　EMO-DB 数据集上的分类报告

Table 3　Classification report on EMO-DB Dataset

	Precision	recall	F1-score
angry	1.00	0.99	0.99
boredom	0.96	0.90	0.92
disgust	0.97	1.00	0.98
fear	0.97	0.93	0.95
happiness	1.00	0.98	0.99
neutral	0.87	0.98	0.92
sadness	0.97	0.97	0.97
accuracy			0.96

经实验验证,本文模型在 RAVDESS 数据集与 EMO-DB 数据集上在不同的评价指标下都可以获得较高的准确率,其中在 RAVDESS 数据集上的准确率、F1_score 和混淆矩阵均高于在 EMO-DB 数据集上所获得的。

3.2.3　改进前后 MFCC 对比

为验证改进 MFCC 在融合多头注意力机制的并行模型上的实验效果,分别在不同数据集上做了原始 MFCC 与改进后 MFCC 准确率的对比,如图 12 所示。从图 12 可以看出经改进后的 MFCC 可使准确率有所提高,在 RAVDESS 数据集上提高了 0.81%,EMO-DB 数据集上提高了 1.87%。

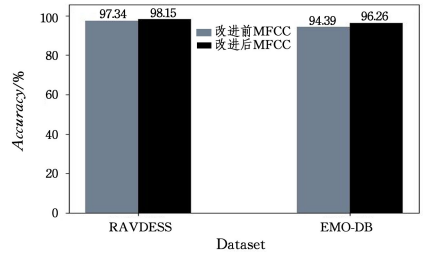


图 12　改进前后 MFCC 准确率对比

Fig. 12　MFCC accuracy comparison before and after improvement

RAVDESS 数据集上改进 MFCC 得到的混淆矩阵如图 13 所示。其中平静类、惊喜类、伤心类与生气类都达到了百分百的预测,无错误预测,厌恶类的准确率最低只有 0.95,其最易与惊喜类发生混淆,其次是害怕类和平静类。F1-score 分类报告如表 4 所列,从该表可以看出改进后的 MFCC 特征在 RAVDESS 数据集上整体都有着较高的准确率,其中平静类与开心类的 F1-score 可以达到 1,害怕、伤心以及惊喜类的 F1-score 较低,为 0.97。

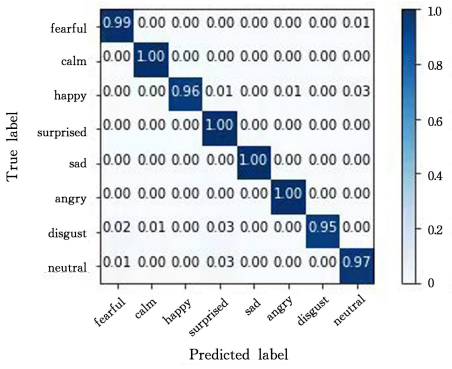


图 13 RAVDESS 数据集上改进后 MFCC 混淆矩阵

Fig. 13 Improved MFCC confusion matrix on RAVDESS dataset

表 4 改进 MFCC 在 RAVDESS 上的分类报告

Table 4 Improved MFCC classification report on RAVDESS

	Precision	recall	F1-score
angry	0.97	0.99	0.98
calm	0.99	1.00	1.00
disgust	1.00	0.96	0.98
fearful	0.94	1.00	0.97
happy	1.00	1.00	1.00
neutral	0.98	1.00	0.99
sad	1.00	0.95	0.97
surprised	0.97	0.97	0.97
accuracy			0.98

EMO-DB 数据集上改进 MFCC 得到的混淆矩阵与分类报告如图 14 和表 5 所示。本文方法在 EMO-DB 数据集上的效果差于在 RAVDESS 数据集上的效果。从图 14 的混淆矩阵可以看出厌恶类的预测最高可达到 1，而中立类的最低，只有 0.89，且其易与生气类发生混淆。从表 5 可知本文方法在 EMO-DB 数据集上的平均 F1-score 为 0.95。

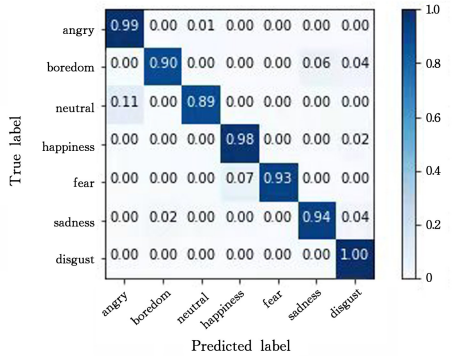


图 14 EMO-DB 数据集上改进后 MFCC 混淆矩阵

Fig. 14 Improved MFCC confusion matrix on EMO-DB dataset

表 5 改进 MFCC 在 EMO-DB 上的分类报告

Table 5 Improved MFCC classification report on EMO-DB

	Precision	recall	F1-score
angry	0.96	0.99	0.97
boredom	0.98	0.90	0.93
disgust	0.96	0.89	0.93
fear	0.93	0.98	0.95
happiness	1.00	0.93	0.96
neutral	0.94	0.94	0.94
sadness	0.88	1.00	0.94
accuracy			0.95

经实验可以得出改进后的 MFCC 特征相对于原始 MFCC 在本文模型上的准确率有所提升，在 RAVDESS 数据集上提高了 0.81%，在 EMO-DB 数据集上提高 1.87%。

3.3 与现有方法比较

文献[19]通过添加高斯噪声及改变速度等方法来增加样本量实现数据扩充，并搭建了 CGRU 模型，在 RAVDESS 数据集和 EMO-DB 数据集上分别得到了 89.1%和 88.7%的准确率。文献[20]利用添加噪声并改变音调进行数据增强后搭建 CNN 模型，在 RAVDESS 数据集上得到了 98%的准确率。文献[21]在 RAVDESS 数据集和 EMO-DB 数据集上采用数据增强和特征融合在 CNN 模型上得到了 90.6%和 93.3%的准确率。文献[22]构建了两个卷积神经网络和长短期记忆(CNN-LSTM) 网络(一个 1D CNN LSTM 网络和一个 2D CNN LSTM 网络)分别从语音和 log-mel 频谱图中学习局部和全局情感相关特征，在 EMO-DB 数据集上获得了 95.89%的准确率。文献[23]使用梅尔谱图和梅尔频率倒谱系数作为音频描述方法，并提出了全卷积神经网络架构作为分类器，在 RAVDESS 数据集上达到了 75.28%的平均准确率，在 EMODB 数据集上有 92.71%的平均准确率。

将本文方法与不同基线最先进的语音情感识别方法进行了比较，以验证所提方法的有效性和稳健性。在 RAVDESS 与 EMO-DB 数据集上对所提模型进行对比分析，结果如表 6 所列。由表 6 可以看出本文方法无论是在 RAVDESS 数据集还是 EMO-DB 数据集上都可以得到较高的准确率。

表 6 本文方法与其他基线方法比较分析

Table 6 Comparative analysis of the proposed method and other baseline methods

(单位：%)

模型	RAVDESS	EMO-DB
文献[19]	89.10	88.70
文献[20]	98.00	—
文献[21]	90.60	93.30
文献[22]	—	95.89
文献[23]	72.28	92.71
本文方法	98.15	96.26

结束语 本文以语音情感识别为背景进行研究，为发挥网络的性能、避免过拟合，利用高斯白噪声与波形位移对数据集进行扩增。采用滑动平均滤波器滤除浊音信号的基音频率，并提取 MFCC 的一阶差分及二阶差分，将其与静态 MFCC 融合以得到更全面的语音信号情感特征。为解决当前模型识别率较低的问题，将改进后的 MFCC 送入融合多头注意力机制的并行模型，提取语音中的空间情感特征以及时序情感特征，并设计了一种联合损失函数保证了不同类别之间的特征可分离又使得类内的特征更加紧凑。在未来的研究中可以考虑将多种特征融合以便更充分地表示语音情感信息。

参 考 文 献

[1] DANNUO J, XIN H, JINGHAN X, et al. Design of Intelligent Vehicle Multimedia Human-Computer Interaction System[C]// IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2019

[2] ZHOU Y, SUN Y, ZHANG J, et al. Speech emotion recognition using both spectral and prosodic features[C]// International Conference on Information Engineering and Computer Science. IEEE, 2009: 1-4.

[3] LIU Z T,XU J P,WU M,et al. Overview of speech emotion feature extraction and dimensionality reduction methods [J]. Journal of Computer Science,2018,41(12):2833-2851.

[4] SHIMIZU T,ONAGA H. Study on acoustic improvements by sound-absorbing panels and acoustical quality assessment of teleconference systems[J]. Applied Acoustics,2018,139(1):101-112.

[5] VERVERIDIS D,KOTROPOULOS C,PITAS I. Automatic emotional speech classification[C]// IEEE International Conference on Acoustics,Speech,and Signal Processing. IEEE,2004:1-593.

[6] SCHULLER B,RIGOLL G,LANG M. Hidden Markov model-based speech emotion recognition[C]// IEEE International Conference on Acoustics,Speech,and Signal Processing(ICASSP'03). IEEE,2003.

[7] LIU Y R,ZHANG X Y,CHEN G J,et al. VMD improves the emotional speech feature extraction of GFCC [J]. Journal of Xi' an University of Electronic Science and Technology,2019,46(5):24-30.

[8] 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.

[9] LEE J,TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition [J]. Interspeech,2015,5(1):10-13.

[10] VERKHOLYAK O V,KAYA H,KARPOV A A. Modeling Short-Term and Long-Term Dependencies of the Speech Signal for Paralinguistic Emotion Classification[J]. SPIIRAS Proceedings,2019,18(1):30-56.

[11] YU H,JI Y,LI Q. Student sentiment classification model based on GRU neural network and TF-IDF algorithm[J]. Journal of Intelligent and Fuzzy Systems,2021,40(2):2301-2311.

[12] LI D D,SUN L Y ,XU X L,et al. BLSTM and CNN Stacking Architecture for Speech Emotion Recognition[J]. Neural Processing Letters,2021,53(6):1-19.

[13] CHEN Q P,HUANG G M. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition[J]. Engineering Applications of Artificial Intelligence,2021,102:104277.

[14] CHEN W L,SUN X. Speech emotion recognition based on MFCCG-PCA [J]. Journal of Peking University(Natural Science Edition),2015,51(2):269-274.

[15] LU W,DAI B J,LI H,et al. The influence of pitch frequency information in MFCC on speaker recognition system performance [J]. Journal of China University of Science and Technology,2009,39(8):859-863,884.

[16] DONG Y F,SU H,LIU B,et al. Model level fusion dimension emotion recognition method based on multi-headed attention mechanism[J]. Journal of Signal Processing,2021,37(5):885-892.

[17] LIVINGSTONE S R,RUSSO F A,JOSEPH N. The ryerson audio-visual da-tabase of emotional speech and song(RAVDESS):a dynamic,multi-modal set of facial and vocal expressions in North American English[J]. PlosOne,2018,13(5):e0196391.

[18] CIRAKMAN O,GUNSEL B. Online speaker emotion tracking with a dynamic state transition model[C]// 23rd International Conference on Pattern Recognition(ICPR). IEEE,2016:307-312.

[19] ZHENG Y,CHEN J N,WU F,et al. Research and Implementation of Speech Emotion Recognition Based on CGRU Model[J]. Journal of Northeastern University(Natural Science Edition),2020,41(12):1680-1685.

[20] PURI T,SONIM,DHIMAN G,et al. Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network. [J]. Journal of healthcare engineering,2022,2022:8472947.

[21] JAHANGIR R,TEH Y W,MUJTABA G,et al. Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion[J]. Machine Vision and Applications,2022,33(3):1-16.

[22] ZHAO J F,MAO X,CHEN L J. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical Signal Processing and Control,2019,47:312-323.

[23] GARCÍA-ORDÁS M T,ALAIZ-MORETÓN H,BENÍTEZ-ANDRADES J A,et al. Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network[J]. Biomedical Signal Processing and Control,2021,69:102946.

CUI Lin, born in 1984, Ph.D, associate professor. Her main research interests include arrayecture signal processing and speech signal processing.

CUI Chenlu, born in 1997, master. Her main research interest is speech emotion recognition.