面向精准育人的多模态学生言语情感识别方法研究

刘振中,张夷楠,王翠萍

(洛阳师范学院 心理健康教育与咨询中心,河南 洛阳 471934)

摘要:本文面向国家“精准育人”战略需求,针对课堂场景中学生面部表情变化细微,导致现有基于计算机视觉的情感识别方法面临识别困难的现实瓶颈,创新性地从学生言语维度出发,提出了一种融合多维度声学特征与文本特征的多模态学生言语情感识别方法,并构建了多模态感知神经网络.其中,在该网络的声学编码器中设计了注意力统计池化层,能够有效提取学生言语中的关键帧级别特征和序列级特征.实验结果表明,所提出的方法中的不同模块均能有效提升情感识别性能,且在识别效果上显著优于现有单模态方法.该方法有效融合了学生言语中的声学和文本的信息,突破了单模态感知的局限性,能够为“精准育人”实践提供可靠的技术支持,切实提高育人成效.

关键词:精准育人;学生言语情感;声学特征;多模态情感识别

中图分类号:O413 **文献标志码**:A **文章编号**:1000-2367(2026)04-0098-08

随着国家对人才培养要求的逐步提高,“精准育人”已成为新时代教育高质量发展的核心诉求[1].课堂教学本质上是一个充满情感互动的师生交流过程,在这一过程中,通过观察并分析学生的言语表达,教师能够即时了解学生的认知水平、情绪状态等高价值信息[2].其中,识别并分析学生课堂言语中隐含的情感状态能够为教师评估学生的学习投入度、课堂参与度提供直接依据[3],同时也是反映学生心理健康状态的关键指标[4].可以发现,精准识别并分析课堂中学生的言语情感,能够为“精准育人”提供高效的技术支持.

目前,在情感识别领域,由于个体的面部表情具有直观、外显等特点,现有研究倾向于通过面部表情来捕捉个体情感.如 Dukić等[5]通过 Inception-v3 和 ResNet34 构建了一种用于视频分析的实时情感预测方法,该方法用于分析视频中学生的情感状态.Cámbara 等[6]探索了一种基于面部识别的学生情绪识别方法,用于智能教室支持的学习场景.然而,课堂是具有一定环境约束的场景,学生会下意识地管理或隐藏自己的真实情感.因此,仅从学生的面部表情来识别其情感状态可能会导致识别无果.从言语的角度来看,学生在进行课堂发言时会无意识地在言语中携带自身情感状态,从学生言语的角度探索学生情感识别方法似乎更有利于快速了解学生的情感状态[7].同时,现有的基于深度学习的学生情感识别方法普遍仅基于文本数据进行建模以识别学生情感,忽略了学生言语中声学特征对言语情感的贡献[8].

基于此,本研究在观察了大量课堂视频的基础上,提出了一种多模态学生言语情感识别方法.此外,在声学特征中,并非所有语音帧在情感识别任务中都具有相同的贡献,一些关键帧能更好地反映学生的言语情感特征.基于这些发现,本文设计并提出了一个能够利用多种多维度声学特征和文本特征的 EmoBERT 神经网络,EmoBERT 由声学编码器、文本编码器和特征解码器共同组成.通过引入注意力池化层,声学编码器能够同时捕捉语句级特征和帧级特征.所提出的方法在公共数据集上进行了评估,实验结果在一定程度上证明了该方法在课堂场景中识别学生言语情感的有效性和优越性.

收稿日期:2025-10-30;**修回日期**:2025-12-16.

基金项目:国家社会科学基金教育学青年项目(CBA230291);河南省软科学研究计划项目(242400410444).

作者简介(通信作者):刘振中(1980—),男,河南开封人,洛阳师范学院副教授,研究方向为高等教育,E-mail:149148069@qq.com.

引用本文:刘振中,张夷楠,王翠萍.面向精准育人的多模态学生言语情感识别方法研究[J].河南师范大学学报(自然科学版),2026,54(4):98-105.(Liu Zhenzhong,Zhang Yinan,Wang Cuiping.Research on multimodal student speech emotion recognition methods for precise education[J].Journal of Henan Normal University(Natural Science Edition),2026,54(4):98-105.DOI:10.16366/j.cnki.1000-2367.2025.10.30.0002.)

1　数据预处理

首先，本研究从课堂视频中分离出音频轨道，并通过基于短时能量和过零率的语音活动检测方法（voice act detection，VAD）以音频中的静默片段为分割依据对音频进行分割[9]．在这一过程中，本研究将每个分割片段的最小和最大长度限制在 1 至 12 s 之间．语音片段可以表示为 $S=\{S_1,S_2,\cdots,S_n\}$．

其次，为了使音频特征更加接近人耳感知，本研究获取了学生言语的梅尔频谱图特征．首先将语音统一处理为 16 kHz．音频进行预加重以补偿声音传播过程中的能量损失，并将音频划分为 25 ms 的短帧，为语音信号添加 10 ms 位移的汉明窗．随后，通过傅里叶变换将音频从时域转换至频域，并通过一组梅尔尺度滤波器组对频谱进行处理，通过对数运算获得对数梅尔滤波器．最后，得到了维度为 50 的梅尔频谱图特征．然而，仅依赖梅尔频谱图难以完全反映学生言语的声学表现[10]．研究表明，基础的低级声学描述符能够更全面地反映话语的情感，使用来自不同音频域的低级声学描述作为多维度声学特征可以为学生言语情感识别提供更有效的信息[6]．因此，本研究加入了不同音频域的多维度声学特征，如时域、频域、能量域和感知域等（如表 1所示），从而为模型提供更充分的声学情感线索．

表 1　多维度声学特征描述

Tab. 1　Auxiliary acoustic feature description

音频域	多维度声学特征	特征描述
时域	起音时间	反映了一个音节从其起始点达到其能量峰值所需的时间、速率．
频域	频谱质心	反映了语音的明亮度．频谱质心越低，语音就越暗沉、低沉．
能量域	均方根能量	反映了在一定时间段内语音的平均能量．
感知域	尖锐度	描述声音的尖锐程度．

最后，本研究利用 Modelscope 提供的语音转录模型[11]，将学生言语的音频数据转录为文本，转录后的文字可被表示为 $T=\{S_1,S_2,\cdots,S_n\}$．由于用于分类任务的文本编码器需要一个[CLS]标记来表示输入的起始位置，发送给文本编码器的文本输入可被表示为 $T=\{S_{\mathrm{CLS}},S_2,\cdots,S_n\}$．

2　EmoBERT 模型

本研究所提出的 EmoBERT 的结构如图 1 所示．EmoBERT 包含了声学编码器、文本编码器以及一个特征解码器．首先，声学编码器和文本编码器接收本研究所设计的多维度声学特征和文本特征，并分别对声学特征和文本特征进行深度编码，将编码后的高维特征发送给解码器．然后，利用融合后的声学特征和文本特征以共同表征学生言语中的情感线索，解码器对融合后的特征进行解码．最后，由解码器输出学生的言语情感类别．

2.1　文本编码器

本研究选择 BERT[12] 作为文本编码器．BERT 是一个由多个 Transformer[13] 架构组成的大型预训练模型，并在海量的先验知识上进行了训练．BERT 具有出色的自然语言表示能力，能够为下游任务提供语言表示支持．所使用的 BERT 预训练模型版本为中文，包含 12 个 Transformer 块和 1.1 亿个参数．在文本编码器中，使用 BERT 来获取文本的高维特征，其维度为 768．

2.2　声学编码器

声学编码器被设计用于抽取学生言语的多尺度声学特征，其包含 1 个一维卷积块、注意力统计池化和全连接块．其中，一维卷积块包含两个一维卷积层，在一维卷积块中，为了使模型能够更细致地理解学生声学特征与情感之间的联系，本研究首先采用了不同尺度的卷积核来提取学生言语的帧级别特征；随后利用注意力统计池化层，从众多帧级别特征中筛选最有价值的帧级别特征；最后，利用全连接层对学生言语中的序列级特征进行建模．

具体而言，在声学编码器中，第 1 个一维卷积层接收人工设计的多维度声学特征 S_a，并捕获帧级别特征 x_a，其输入通道为 80，输出通道为 512，卷积核大小为 5．随后，注意力统计池化层接收由一维卷积块捕捉到的

帧级别特征 x_a,并计算加权平均值和加权标准差以选择 x_a 中与情感联系最为紧密的声学特征.在注意力统计池化层中,首先,将由一维卷积层抽取到的帧级别特征中的关键特征计算为标量 s_t.

$$s_t = v^t f(Wh_t + b) + k,$$

其中 $f(\cdot)$ 是一种非线性激活函数,此处本研究使用了 tanh 函数.$h_t(t=1,\cdots,T)$ 是第 t 个时间步长时最后一帧层的激活阈值.本研究使用 Softmax 函数对 s_t 进行归一化,从而得到注意力得分 α_t,并且 α_t 是池化层的权重,用于计算加权平均向量 $\boldsymbol{\mu}$,

$$\boldsymbol{\mu} = \sum_t^T \alpha_t h_t.$$

然后,通过计算加权标准差 σ,使注意力统计池化层更关注有价值的语音帧,其计算过程为:

$$\sigma = \sqrt{\sum_t^T \alpha_t h_t \odot h_t - \boldsymbol{\mu} \odot \boldsymbol{\mu}},$$

其中⊙表示哈达玛积.注意力统计池化操作的最终输出是由 $\boldsymbol{\mu}$ 和 σ 两部分拼接而成的,这两部分均具有 512 个维度.因此最终输出是 $\boldsymbol{\mu}$ 和 σ 的总和,即 1 024.最后,通过全连接块对学生言语的序列级特征进行建模.全连接块包含两个全连接层,用于从注意力统计池化层的输出中提取序列级声学特征.第一个全连接层用于将特征维度降低至 512.随后,第 2 个全连接层接收输出,并保持特征在原始的 512 维度上不变.声学编码器的最终输出可以表示为 X_a.

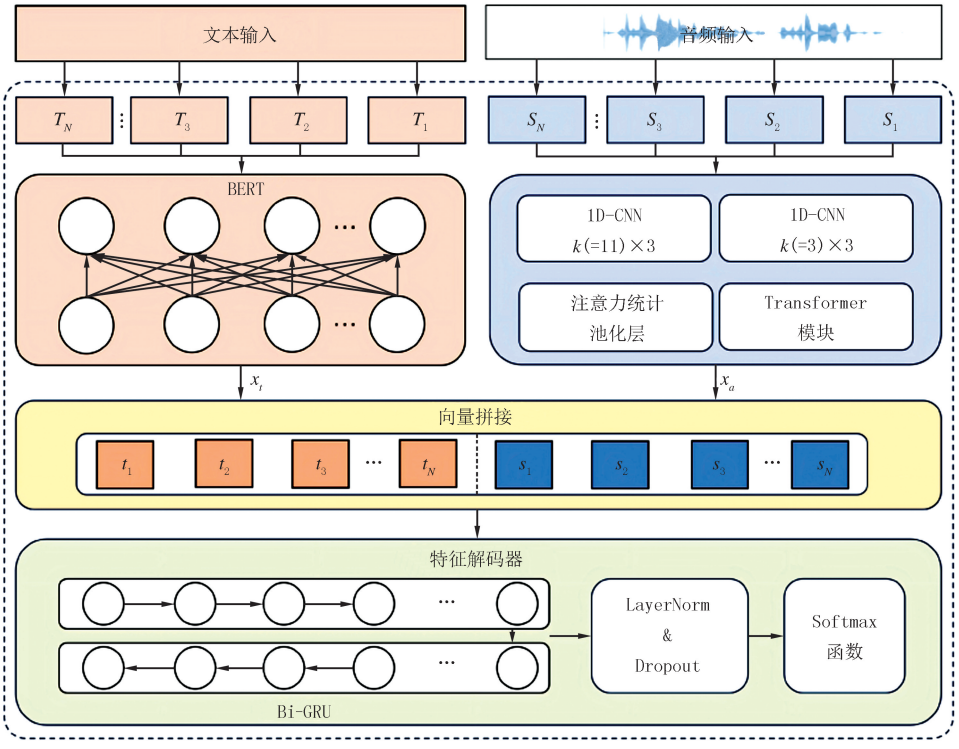


图1 EmoBERT结构图

Fig.1 The structure of EmoBERT

2.3 特征解码器

在本研究所设计的 EmoBERT 中,特征解码器被部署用于融合并解码来自声学编码器和文本编码器的输出.由于特征解码器接收的高维特征是由声学编码器和文本编码器中的注意力机制筛选得到的,因此在解码器中融合高维文本特征和高维声学特征时,本研究仅采用了简单的向量拼接操作,融合后的特征表示为 H_t,这一过程可被简单表示为:

$$H_t = C_e(X_t, X_a),$$

其中,H_t 的维度是 X_t 和 X_a 维度之和,即 1 280.在特征解码器中,将 H_t 视为一个序列. 为了有效建模声学

特征和文本特征的序列依赖关系，本研究采用了门控循环单元(gated recurrent unit，GRU)作为序列特征提取器.GRU 对序列中的时间依赖性具有极强的建模能力[14].因此，两层双向 GRU 被堆叠，并用以捕捉学生言语中的序列依赖性.在 GRU 中，隐藏层的维度为 256.同时，为了避免过拟合，本研究在隐藏层之间设置了 0.3 的 dropout. $c_i^s \in \mathbf{R}^{d_u}$ 表示每个融合特征，由 GRU 表征的输出 $H_i^s \in \mathbf{R}^{2d_u}$ 可以表示为：

$$H_i^s, h_i^s = \overleftrightarrow{\mathrm{GRU}}(c_i, h_{i-1}^s),$$

其中，$h_i^s \in \mathbf{R}^{d_u}$ 表示 GRU 的第 i 个隐藏层的状态. 最终，利用 Softmax 函数输出学生言语情感的类别.

3　结果与讨论分析

3.1　数据集与客观评价指标

本研究采用了交互式情感动作捕捉(interactive emotional motion capture，IEMOCAP)[15]和多模态情感线数据集(multimodal emotion-line dataset，MELD)[16]作为实验数据集，IEMOCAP 和 MELD 均为多模态言语情感识别任务中较为常用的数据集，能够有效评估本研究所提出方法的性能.IEMOCAP 数据集包含有脚本化和即兴的表达情感的讲话内容，共分为 5 个部分.IEMOCAP 数据集包含 3 种数据形式：视频、音频和文本，总共涵盖 10 种情感，其中，常被广泛用于实验的 4 种情感类型分别为愤怒、快乐、中性以及悲伤，共计 5 584 条数据.MELD 数据集共计 11 089 段脚本化对话，具有 3 种形式：视频、音频和文本，包含愤怒、厌恶、快乐、沮丧、中性、惊讶、害怕等 7 种情感.

为了评估本研究所提出的方法在实际课堂场景中的效果，本研究采集了 235 个在智能教室中录制的高质量课堂视频，并对视频数据进行了处理，以构建多模态学生情感数据集(multimodal teacher emotion dataset，MTED).MTED 中的数据涵盖了生物、信息技术和语文 3 个学科，每个视频中至少包含了两位学生的发言.两名教育技术专业的研究生对数据集进行了手动标注，其 Kappa 系数 0.77，表明标注结果的一致性程度较高.筛选并处理后的 MTED 包含了 6 个类别的学生言语情感，分别为：兴奋、快乐、中性、困惑、沮丧和惊讶，共计 6 314 条数据.

未加权准确率(unweighted accuracy)和加权平均 F1 分数(weighted average F1，WA-F1)被广泛用于评估模型在多模态情感识别任务中的效果.因此，为了评估 EmoBERT 在 IEMOCAP、MELD 和 MTED 数据集上的效果，本研究采用了先前研究中使用的评估指标.针对 IEMOCAP 中的 4 类言语情感，采用 UA_4 以及 WA-F1 作为评价指标；针对 MELD 中的 7 类言语情感，采用 UA_7 以及 WA-F1 作为评价指标；针对自建数据集 MTED 的 6 类言语情感，采用 UA_6 以及 WA-F1 作为评价指标，UA_N 的计算方法如下所示：

$$UA_N = \frac{\sum_{i=1}^{N} t_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} t_{ij}},$$

其中，N 表示类别数量，t_{ij} 表示被标注为类别 i，但预测为类别 j 的数量. 需要注意的是，本研究在公开数据集 IEMOCAP 和 MELD 中所采用的 BERT 版本为英文版的 $\mathrm{BERT}_{\mathrm{base}}$，在自建数据集 MTED 中所采用的 BERT 版本为中文版的 $\mathrm{BERT}_{\mathrm{base}}$，二者结构一致，参数量相同，且隐藏层大小均为 768.

3.2　实验结果

本研究在 IMEOCAP 和 MELD 公开数据集上对所提出方法的效果进行了评估.同时，本研究选取了 IEMOCAP 和 MELD 上的最新方法作为基准模型，并与本研究的方法进行了比较.由表 2 中的实验结果可以看出，本研究提出的 EmoBERT 模型超越了当前较为先进的方法，并在 IEMOCAP 数据集上取得了最佳性能，其 UA_4 和 WA-F1 分别达到了 78.6%和 77.8%.由表 3 中的实验结果可以看出，所提出的 EmoBERT 在 MELD 测试中的表现优于多种现有较为先进的方法，其 UA_7 和 WA-F1 分别为 66.2% 和 64.7%.

3.3　消融实验及结果

为了探究共同组成 EmoBERT 模型的不同模块在学生言语情感识别任务中的贡献程度，本研究在 MTED 数据集中设计并实施了消融实验，实验结果如表 4 所示.

表 2 模型在 IEMOCAP 中的实验结果

Tab. 2 Experimental results of the model in IEMOCAP %

方法	模态	UA_4	WA-F1
DBT[17]	文本+音频	74.0	—
CHFusion[18]	文本+音频+视频	76.5	76.8
Makiuchi[19]	文本+音频	73.2	—
SAWC[20]	文本+音频	76.8	76.9
EmoBERT	文本+音频	78.6	78.8

表 3 模型在 MELD 中的实验结果

Tab. 3 Experimental results of the model in MELD %

方法	模态	UA_7	WA-F1
bc-LSTM[21]	文本+音频	57.5	56.4
DialogueGCN[22]	文本	59.5	58.1
DialogueCRN[23]	文本	60.7	58.4
MM-DFN[24]	文本+音频	62.5	59.5
EmoBERT	文本+音频	66.2	64.7

表 4 EmoBERT 在 MTED 中的消融实验结果

Tab. 4 The ablation experiment results of EmoBERT in MTED %

步骤	模型	UA_6	WA-F1	步骤	模型	UA_6	WA-F1
(a)	基线模型(完整的 EmoBERT)	82.1	81.7	(d)	替换特征解码器	79.1	78.5
(b)	去除多维度声学特征	81.5	81.3	(e)	去除声学特征	72.5	71.2
(c)	替换注意力统计池化层	80.3	80.2	(f)	去除文本特征	60.7	59.3

由表 4 中的数据可以看出,基准模型的准确度为 82.1%,WA-F1 为 81.7%,这一结果充分证明了所提出的方法在模拟课堂场景中识别学生言语情感的有效性.将消融实验的结果与基线模型进行对比可以发现,去除多维度声学特征之后,所提出方法的 UA_6 和 WA-F1 分别下降了 0.6%和 0.4%,该结果说明了本研究加入的多维度声学特征在识别学生言语情感中发挥了一定的辅助作用,同时也证明了构建更加细腻的数据处理方法能够增强模型对学生言语情感的感知能力;将注意力统计池化层替换为最大池化层之后,UA_6 和 WA-F1 分别下降了 1.2%和 1.1%,该结果证明了注意力统计池化层能够使模型更加关注声学特征中最有价值的音频帧,并在一定程度上减少了特征冗余;将特征解码器中的双向 GRU 结构替换为多层感知机之后,模型的 UA_6 和 WA-F1 分别下降了 1.2%和 1.7%;去除声学特征输入导致模型的 UA_6 和 WA-F1 分别下降了 6.6%和 7.3%,该实验模拟了缺乏音频模态的情况,说明声学特征对于言语情感表征的效果具有较大影响.去除文本特征输入对学生言语情感识别的影响最大,直接导致模型的 UA_6 和 WA-F1 分别下降了 18.4%和 11.9%,这在一定程度上模拟了缺乏文本模态的情况,并表明仅使用声学模态来识别学生言语情感时,识别的准确率会大大降低.

本研究通过混淆矩阵(如附录图 S1 所示)反映了去除 EmoBERT 的不同模块后,EmoBERT 在学生言语情感任务中的识别效果.

本研究实施的消融实验充分解释了所提出的方法中不同模块对学生言语情感识别的贡献.比较图 S1 中的(a)和(b),可以看出多维度声学特征在提高学生言语情感识别性能方面发挥了一定作用.其中,受多维度声学特征影响最大的类别是“沮丧”,其识别准确度提高了 6.4%.这在一定程度上反映了学生更倾向于通过更加细腻的声学变化来凸显“沮丧”情感.

通过比较图 S1 中的(b)和(c),可以发现本研究所设计的注意力池化层对识别不同学生言语情感有很大帮助,而“困惑”类别是注意力池化最大的影响因素.“困惑”这一类别受到注意力机制的影响最为显著,模型注意力的汇聚使该类别的识别准确度提高了 3.3%.这一结果在一定程度上说明了本研究所提出的注意力统计池化层对于捕捉学生言语的帧级别特征发挥了重要作用.

对比图 S1 中的(c)和(d),可以发现将解码器中的 GRU 结构替换为全连接层会对所有类别的识别效果产生影响,其中“惊讶”和“兴奋”的识别准确度分别降低了 3.5%和 3.8%.这一结果表明,无论是在声学特征还是在文本特征上,学生表达的言语情感均具有一定的序列关系.同时,这一结果也说明了所采用的 GRU 结构能够有效整合融合后的声学特征和文本特征,并捕捉其内在的序列依赖关系.

通过对比图 S1 中的(d)和(e)可以发现,去除声学特征输入之后,“沮丧”情感的识别准确度受到的影响最大,其识别准确度相较于基准模型降低了 16.8%,其次是“困惑”和“兴奋”,识别准确度分别下降了 13.1%和 10.9%.该结果在一定程度上说明了声学特征在学生表达“沮丧”“困惑”“兴奋”3 类情感时所发挥的重要

作用.

通过对比图 S1 中的(d)和(f)，去除文本特征输入之后，模型对 6 类学生言语情感的感知能力显著下降，尤其是“困惑”情感，下降幅度为 23.1%.其次，“快乐”和“中性”两类情感也受到了较大的影响，下降幅度分别为 21.7%和 20.1%.该结果证明了文本模态中的句法结构、词汇以及语境等信息在学生言语情感识别中发挥了关键作用.

通过对比结果可以发现，本研究在 EmoBERT 中设计的不同模块均对提高学生言语情感识别的准确度具有不同程度的正向贡献；相较于声学特征，文本特征对学生言语情感识别的贡献更大，说明文本特征蕴含了更多的情感线索.

3.4　讨论

3.4.1　声学变化是学生言语情感表达的重要组成部分

通过消融实验的结果可以看出，多维度声学特征是对声学特征更加细致的人工描述，能够更加细腻地反映学生表达言语情感时语调、声量等声学特征上的变化.当多维度声学特征被作为模型的输入时，能够有效提升模型对情感的感知能力.此外，当模型不接收声学特征输入时，受到影响最大的学生言语情感类别为“沮丧”“兴奋”和“困惑”，这也在一定程度上证明了学生在课堂中表达这 3 类情感时，往往会倾注更多的语调、音量的变化来凸显情感.

3.4.2　文本蕴含更多学生言语情感线索

通过消融实验的结果可以发现，在识别学生言语情感时移除文本输入，模型性能出现显著下降.这一结果表明，文本信息在识别学生言语情感过程中起到了不可替代的作用.相较于声学模态，文本可能承载了更明确的情感语义线索.具体而言，文本中蕴含的句法结构、情感极性词语以及语境等特征直接反映了学生的主观感受、意图与评价，这些特征均为情感判断提供了准确的判别依据.当失去文本输入时，受到影响最大的情感类别为“快乐”“中性”以及“困惑”，这在一定程度上证明了这 3 类情感的表达与句法结构、情感极性词语以及语境信息均有密切的联系.

3.4.3　多模态融合能够有效提升学生言语情感识别效果

通过实验结果可以发现，无论是公开数据集还是自建数据集，多模态融合方法的效果均优于单模态的方法，这一结果证明了多模态融合方法在学生言语情感识别中的先进性.同时，消融实验的结果表明，在利用多模态融合方法进行学生言语情感识别时，文本模态和音频模态均应被视为关键模态，其在情感识别任务中的贡献理应受到更多关注.在未来的研究中，进一步探讨文本模态与声学模态之间的深层次融合机制，似乎是提升学生言语情感识别效果的最有力的途径.

4　总结

针对真实课堂场景中的学生言语情感识别，得益于真实课堂场景中出现的丰富数据形式，本研究面向“精准育人”需求，提出了一种通过多模态融合以增强模型的情感感知能力的方法.在该方法中，本文设计并提出了多模态融合的 EmoBERT 神经网络，该网络不仅能够有效融合学生言语中的声学特征和文本特征，同时可以利用注意力统计池化机制选择最有价值的帧级别特征.在公开数据集上的实验结果证明了本研究所提出的方法优于现有方法，具有一定的先进性.在自建数据集 MTED 中的消融实验结果表明，本研究提出的方法能够更有效地识别课堂场景中的学生言语情感，能够为“精准育人”提供技术支持.同时，消融实验的结果证明了 EmoBERT 中的不同模块对学生言语情感识别任务具有不同程度的正向贡献.

然而，本研究仍存在一些不足之处：虽然本研究提出了一种不依赖计算机视觉的学生言语情感识别方法，但这并不意味着该方法具有显著的优势.在课堂教学过程中，学生言语的隐含情感状态与言语行为之间也存在着紧密联系.例如，当学生在理解新概念遇到障碍时，其困惑情绪会促使他们在课堂互动中提出疑问.这类提问行为通常伴随着特定的语言特征，如使用“我不太明白”“能否再解释一下”等试探性表达.因此，在未来的研究中，将学生的言语行为视为识别学生言语情感的重要线索之一，并探索行为对情绪识别的影响具有一定的研究价值.同时，探究如何有效融合学生的言语信息和视觉信息，并进一步完善多模态情感识别方

法,为识别学生情感提供更充分的数据支持.

附录见电子版(DOI:10.16366/j.cnki.1000-2367.2025.10.30.0002).

参 考 文 献

[1] 彭婧,陈强.从“选育”到“孕育”:中美基础学科拔尖创新人才培养模式比较与启示[J].创新科技,2025,25(11):11-22.
Peng J,Chen Q.From "selection and cultivation" to "nurturing":a compara-tive study of training models for innovative talents in fundamental disciplines between China and the United States[J].Innovation Science and Technology,2025,25(11):11-22.

[2] 方海光,洪心,孔新梅,等.基于课堂交互行为数据的教学教研融合研究:以网络画板和 iFIAS 的数学教学教研应用为例[J].中国电化教育,2022(5):115-121.
Fang H G,Hong X,Kong X M,et al.Research on teaching and research integration based on classroom interactive behavior data:a case:application of netpad and iFIAS in mathematics[J].China Educational Technology,2022(5):115-121.

[3] 李协吉,孙洪涛,高佩琪.教师—学生积极情感表达一致性对中学生学习投入的影响:基于多项式回归与响应面分析[J].教师教育研究,2024,36(4):72-80.
Li X J,Sun H T,Gao P Q.The influence of teacher-students' Positive emotional expression consistency on middle school Students' Learning engagement:based on polynomial regression and response surface analysis[J].Teacher Education Research,2024,36(4):72-80.

[4] 张屹,郝琪,陈蓓蕾,等.智慧教室环境下大学生课堂学习投入度及影响因素研究:以“教育技术学研究方法课”为例[J].中国电化教育,2019(1):106-115.
Zhang Y,Hao Q,Chen B L,et al.Research on college students'classroom engagement and its influencing factors in smart classroom environment:using educational technology research method course as an example[J].China Educational Technology,2019(1):106-115.

[5] Dukić D,Sovic Krzic A.Real-time facial expression recognition using deep learning with application in the active classroom environment[J].Electronics,2022,11(8):1240.

[6] Cámbara G,Luque J,Farrús M.Convolutional speech recognition with pitch and voice quality features[EB/OL].[2025-09-12].https://arxiv.org/abs/2009.01309.

[7] 黄庭培,郑秋梅,李世宝.教师和学生的课堂行为互动及优化策略[J].教育理论与实践,2016,36(32):54-56.
Huang T P,Zheng Q M,Li S B.Interaction between teachers and students'classroom behavior and optimization strategies[J].Theory and Practice of Education,2016,36(32):54-56.

[8] 杨伊,陈昌来,陈兴冶.基于多模态语料库的教师评价类话语特征研究[J].教师教育研究,2024,36(5):22-29.
Yang Y,Chen C L,Chen X Y.A study on the characteristics of teacher evaluation discourse based on multimodal corpus[J].Teacher Education Research,2024,36(5):22-29.

[9] 郭子漾,李国勇.基于双门限的语音端点检测算法改进[J].计算机应用,2025,45(S1):101-105.
Guo Z Y,Li G Y.Improvement of speech endpoint detection algorithm based on dual-threshold[J].Journal of Computer Applications,2025,45(S1):101-105.

[10] Wang C Y,Ren Y,Zhang N,et al.Speech emotion recognition based on multi-feature and multi-lingual fusion[J].Multimedia Tools and Applications,2022,81(4):4897-4907.

[11] Gao Z F,Zhang S L,McLoughlin I,et al.Paraformer:fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition[EB/OL].[2025-09-12].https://arxiv.org/abs/2206.08317.

[12] Devlin J,Chang M W,Lee K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,Volume 1(long and Short Papers).Minneapolis:[s.n.],2019:4171-4186.

[13] Vaswani A,Shazeer N,Parmar N,et al.Attention is all you need[EB/OL].[2025-09-13].https://proceedings.neurips.cc/paper_files/paper/2017/hash3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[14] 孙志,王冠.自监督对比学习的 CNN-GRU 语音情感识别算法[J].西安电子科技大学学报(自然科学版),2024,51(6):182-193.
Sun Z,Wang G.CNN-GRU speech emotion recognition algorithm for self-supervised comparative learning[J].Journal of Xidian University(Natural Science),2024,51(6):182-193.

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

[16] Poria S,Hazarika D,Majumder N,et al.MELD:a multimodal multi-party dataset for emotion recognition in conversations[EB/OL].[2025-09-13].https://arxiv.org/abs/1810.02508.

[17] Yi Y F,Tian Y,He C,et al.DBT:multimodal emotion recognition based on dual-branch transformer[J].The Journal of Supercomputing,2023,79(8):8611-8633.

[18] Majumder N，Hazarika D，Gelbukh A，et al.Multimodal sentiment analysis using hierarchical fusion with context modeling[J].Knowledge-Based Systems，2018，161：124-133.

[19] Makiuchi M R，Uto K，Shinoda K.Multimodal emotion recognition with high-level speech and text features[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU).Cartagena：IEEE，2021：350-357.

[20] Santoso J，Yamada T，Ishizuka K，et al.Speech emotion recognition based on self-attention weight correction for acoustic and text features[J].IEEE Access，2022，10：115732-115743.

[21] Poria S，Cambria E，Hazarika D，et al.Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th Annual Meeting of the Association forComputational Linguistics(volume 1：Long Papers).[S.l.]：ACL，2017：873-883.

[22] Ghosal D，Majumder N，Poria S，et al.DialogueGCN：a graph convolutional neural network for emotion recognition in conversation[EB/OL].[2025-09-16].https：//arxiv.org/abs/1908.11540.

[23] Hu D，Wei L W，Huai X Y.DialogueCRN：contextual reasoning networks for emotion recognition in conversations[EB/OL].[2025-09-13].https：//arxiv.org/abs/2106.01978.

[24] Hu D，Hou X L，Wei L W，et al.MM-DFN：multimodal dynamic fusion network for emotion recognition in conversations[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP).Singapore：IEEE，2022：7037-7041.

Research on multimodal student speech emotion recognition methods for precise education

Liu Zhenzhong，Zhang Yinan，Wang Cuiping

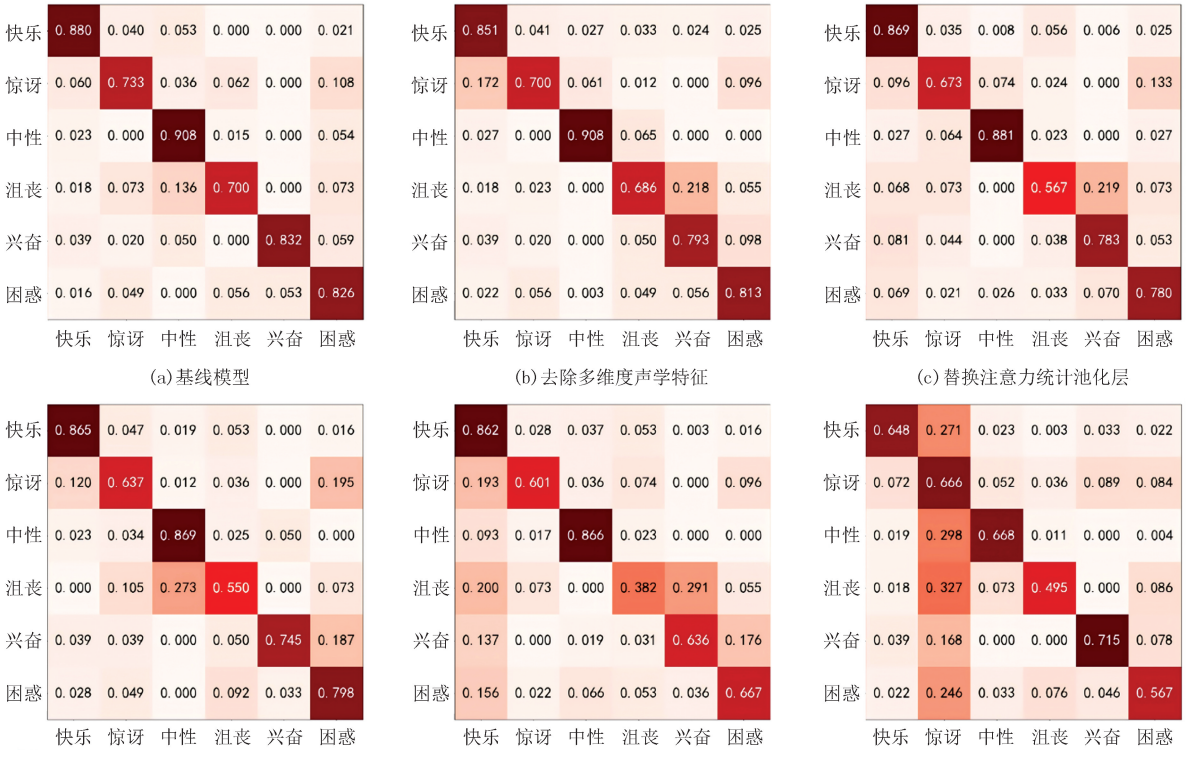
(Center for Mental Health Education and Counseling，Luoyang Normal University，Luoyang 471934，China)

Abstract：This study addressed the national strategic requirement of precision education by focusing on the subtle changes in students' facial expressions in classroom，which pose significant challenges to existing computer vision-based emotion recognition methods. This study innovatively approach the issue from the dimension of student speech，proposing a multimodal student speech emotion recognition method that integrates multidimensional acoustic and textual features，and constructing a multimodal perception neural network. Specifically，an attention statistical pooling layer is designed in the acoustic encoder of this network，effectively extracting key frame-level and sequence-level features from student speech. Experimental results demonstrate that the different modules within the proposed method significantly enhance emotion recognition performance，outperforming existing unimodal methods in recognition effectiveness. This method effectively integrates acoustic and textual information from student speech，overcoming the limitations of unimodal perception，and provides reliable technical support for the practice of precision education，thereby substantially improving educational outcomes.

Keywords：precise education；students' speech emotion；acoustic features；multimodal emotion recognition

[责任编校　陈留院　杨浦]

附录



图S1　EmoBERT在消融实验中的混淆矩阵

Fig.S1　The confusion matrix of EmoBERT in the ablation experiment