邵清, 陈世杰. 一种用于语音情感识别的高弹性多模态架构[J]. 智能计算机与应用,2026,16(8):186-194. DOI:10.20169/j.issn.2095-2163.25022004

一种用于语音情感识别的高弹性多模态架构

邵 清, 陈世杰
(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: 由于语音情感识别本身具有高度抽象且主观性强的特征,在单模态层面极具挑战性,这使得多模态方法的探索日益增多。先前的多模态研究大多聚焦于提升识别准确性,由此引发了模型规模不受控制的增长,尤其是在深度学习兴起之后,这种对资源的需求限制了许多应用场景。本文从新的视角出发,致力于高效利用多模态信息,进而提出一种新颖的高弹性架构。该架构允许以极为简便的并行方式构建复杂的多模态模型。无论模态如何,其都能自由灵活地组合不同模型,并且可以纳入多任务模型。同时提供了一种基于概率分布阐释该架构有效性的方法,并明确了使其发挥作用的条件。此架构能够良好地利用多模态信息,在 IEMOCAP 数据集上取得了 83.84% 的先进准确率,相较于现有的多模态方法提升了 4.93%,相较于单模态方法提升了 7.22%。

关键词: 语音情感识别; 多模态; 多任务; 高弹性
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0186-09

A highly-elastic multimodal architecture for speech emotion recognition

SHAO Qing, CHEN Shijie
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China)

Abstract: Due to the highly abstract and subjective characteristics of speech emotion recognition itself, it is extremely challenging at the single-modal level, which has led to an increasing number of explorations of multi-modal methods. Most previous multi-modal studies have focused on improving recognition accuracy, which has triggered the uncontrolled growth of model scale. Especially after the rise of deep learning, this demand for resources has limited its deployment in many application scenarios. From a new perspective, this paper is committed to the efficient utilization of multi-modal information and then proposes a novel highly elastic architecture. This architecture allows for the construction of complex multi-modal models in an extremely simple parallel manner. Regardless of the input modality, it can freely and flexibly combine different models and can incorporate multi-task models. At the same time, a method for explaining the effectiveness of this architecture based on probability distributions is provided, and the conditions for making it work are clarified. This architecture can make good use of multi-modal information and has achieved an advanced accuracy rate of 83.84% on the IEMOCAP dataset, which is an improvement of 4.93% compared with existing multi-modal methods and an improvement of 7.22% compared with single-modal methods.

Key words: speech emotion recognition; multi-modal; multi-task; highly-elastic

0 引 言

在过去十年中,随着互联网以及信息与通信技术的发展,可交互设备的数量呈爆发式增长。作为人机交互的核心技术之一,语音情感识别受到了越来越多的关注[1],是一项长期存在的挑战,尽管深度学习极大地推动了语音情感识别任务的发展,但语音数据本身处理难度较高,且仍然面临着许多困难。语音数据包含大量高度抽象的信息,包括语义、说话者身份和情感等[2]多个维度。使用深度学习方法处理这些高层次信息会显著加剧过拟合问题。因此,与语音相关的任务通常从手动特征提取开始,

基金项目: 国家自然科学基金(61472256,61170277); 上海理工大学科技发展基金(16KJFZ035,2017KJFZ033); 上海理工大学沪江基金(A14006)。

作者简介: 邵 清(1970—), 女,博士,硕士生导师。主要研究方向:深度学习。
通信作者: 陈世杰(1999—), 男,硕士。主要研究方向:深度学习。Email:slngen@163.com。

收稿日期: 2025-02-20

例如梅尔频率倒谱系数(MFCC)[3-4]和滤波器组[5]，然后将这些特征输入端到端的深度学习模型中以完成下游任务[6]。

幸运的是，借助自然语言处理(NLP)和计算机视觉(CV)所开辟的道路，无监督端到端语音表征学习最近取得了里程碑式的进展[7]。无监督表征学习模型通常采用预训练-微调框架开展应用，在预训练阶段输入大量未标注的数据以学习表征模式，然后在微调阶段通过监督学习来执行下游任务[8]。从 Wav2vec 系列[9]开始，越来越多有效的语音表征学习方法被提出，例如 HuBERT[10] 和 WavLM[11]。在预训练过程中，这些模型使用 Libri-Light[12]、GigaSpeech[13]和 VoxPopuli[14]等语音数据集，这类数据集的总时长可达数万小时，已经超过了一个人一生所能听到的语音时长，并且数据集规模仍在不断扩大。

这些模型通常具有相似的结构，首先，使用堆叠的卷积神经网络(CNN)对语音数据进行降维并提取初步特征，然后使用一个上下文网络对这些特征进行转换，以获得最终有效的表征。然而，不同模型在预训练阶段的优化目标存在差异，这会导致学习到的表征模式各不相同。也为利用多模式表征提升系统性能提供了基础条件。

对于语音情感识别任务，在现有的研究工作中已经进行了相应的尝试，例如文献[15]融合了通过不同特征提取方法所获得的语音表征，即梅尔频率倒谱系数、频谱图和 Wav2vec 所提取的特征。尽管已经取得了一些改进，但仍有更大的潜力可挖掘，因为在文献[16]中只有一种基于预训练的特征提取方法，其余均为手动提取方法，导致表征模式较为单一；并且其特征融合方法是建立在深度网络之上的，意味着过拟合问题在很大程度上仍然会影响性能。其他类似的研究工作[17-18]也存在同样的问题。

此外，从不同表征模式的视角来看，不少方法其实可以实现统一，如多模态方法和多任务方法。多模态方法和多任务方法的通用框架如图 1 所示。多模态方法通常使用不同的特征提取器来处理不同模态的输入，然后将提取到的隐藏特征输入到一个特征融合网络中，以完成下游任务。由于输入的模态不同，每个特征提取器会学习到一种独特的表征模式，随后这些表征模式会被用于同一个下游任务。相比之下，多任务方法通常共享一个单一的特征提取器，期望共享的隐藏特征的表征模式能够同时适用于所有不同的下游任务。从这个角度来看，可以发现多模态方法和多任务方法都是在特定层融合具有不同表征模式的特征。本文提出的方法正是基于这一思路，设计了一个固定层来融合特征，以便让多模态和多任务框架在一个架构下实现统一。

多任务和多模态方法在语音情感识别任务中取得了巨大成功，两类方法都能很好地利用多种模态信息[19-20]。最主流的方法是同时利用语音和文本信息。例如，文献[21]在训练阶段同时执行语音情感识别任务和自动语音识别(ASR)任务，在推理阶段仅执行语音情感识别任务，从而提高了识别准确率。除了同时利用语音和文本信息之外，还有部分多任务方法会利用额外的抽象信息，比如说话者身份信息[22-23]。文献[24-25]通过多模态方法融合语音和文本特征，也取得了比单模态情况下更好的结果。然而，现有这些多模态方法都依托单个复杂模型完成特征利用，使得模型对计算资源的要求越来越高，比如需要更高的计算能力、更高效的内存等。这些缺点极大地限制了该类方法的实际应用，因此本文基于上述分析与提出的框架，提出了一种新颖的高弹性架构，该架构能够同时处理多模态与多任务方法，能更好地利用不同表征模式的特征，同时支持解耦部署以适配灵活多变的实际应用场景。本文的主要贡献如下：

(1) 在公开数据集上取得了当前最优的准确率结果；

(2) 提供了一个统一多模态与多任务方法的新视角，并通过实验验证了该视角的有效性；

(3) 提出了一种高弹性架构，并针对投票机制进行了定量的实验。

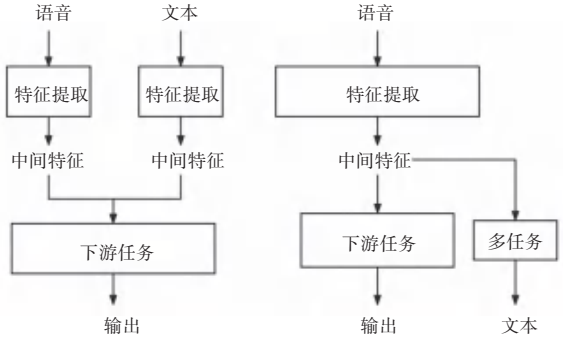


图 1　多模态架构(左)和多任务架构(右)

Fig. 1　Multimodal (left) and multi-task framework (right)

1　模型结构

在本节中，本文将介绍模型架构的详细信息。该架构由任意数量的处理线路(ProcessLine，PL)和一个投票机制(Vote Mechanism，VM)组成。PL 用

于以不同的表征模式方式处理不同模态或相同模态的输入,并生成一个名为 PL(PLout)输出的中间结果。用于融合不同表征模式特征的特定层,被称为投票机制,其通过整合来自各个 PL 的所有符合条件的 PLout 来获得最终结果。模型架构的整体概览如图 2 所示。

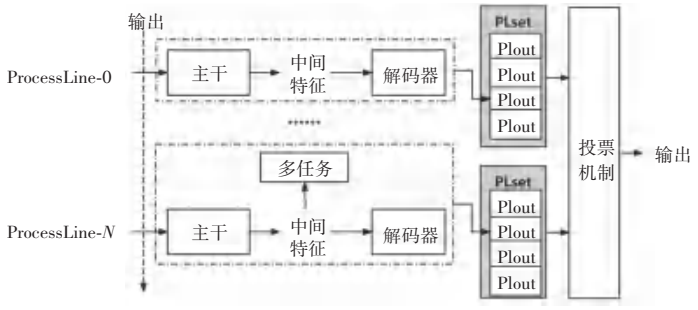


图 2 模型架构概览

Fig. 2 The overview of proposed architecture

1.1 处理线路(PL)

每个 PL 代表了一种用于语音情感识别任务的处理方法。PL 内部包含 3 个核心模块,即主干网络(Backbone)、解码器(Decoder)和 PL 集合(PLset)。主干网络是针对特定模态输入的特征提取器,能够生成具有独特表征模式的隐藏特征。隐藏特征被输入到下游组件中,若接入语音情感识别任务的解码器,则解码器会为情感标签给出分数以便进行进一步分析,或者输入到其他多任务模块中。当主干网络和解码器完成参数更新后,如在一个训练步骤中,本文会将更新后的模型整体打包成一个整体,称为 PLout,并存储在容量有限的 PLset 中。PL 集合中的那些 PLout 将有权通过投票机制对最终输出进行投票。PLset 被视为一个解耦层,当本文想要获得最新的输出时,投票机制不会对 PLset 以上的部分产生影响。

该解耦特性具备如下优势:

并行性:由于 PL 彼此完全独立,可以毫无限制地并行运行,例如在各种不同的设备中运行,运行速度可以不同,甚至可以在不同的时间运行。

高弹性:在选择 PL 方面没有任何限制,所有类型的有效方法都可以使用,并且本文可以根据具体情况来选择 PL 的组合,无论是需要更高的效率、更低的功耗,还是更高的准确率。具体来说,假设总共 N 个 PL,那么该架构可支持的 PL 组合形式总数为:

$$C_N^1 + C_N^2 + \cdots + C_N^N = 2^N - 1 \tag{1}$$

这对应了所有非空的 PL 组合形式,体现了架构的高度灵活性。

本文为 PLset 设计了两种维护策略:交互模式(Interflow mode)与同步模式(Synchro mode)。交互模式需要手动设置 PLset 的大小,即每个 PLset 中保留的 PLout 的最大数量。一旦 PL 更新且 PLset 已满时,新的 PLout 将其评估结果与 PLset 中现有的 PLout 进行比较,如果新的 PLout 比其中任何一个都好,那么较差的那个 PLout 将被淘汰。这种贪婪策略的设计是为了保持解耦特性。同步模式是交互模式的简化版本。鉴于交互模式需要保存 PLout,并且投票机制的计算消耗与 PLout 的数量成正比,本文将 PLset 的大小减小到 1,并用来自 PL 的任何新的 PLout 替换 PLset 中较旧的 PLout。由于随机性增加,同步模式下的结果在性能上比交互模式稍差,但仍然是可以接受的。具体细节请参考实验部分。

1.2 投票机制

每个 PLset 都可以推荐一个 PLout 用于投票。如果各个 PLout 处于同一个 PLset 中,那么这些 PLout 是平等的。因此,存在一定数量的投票组合。例如,总共有 N 个 PL,每个 PL 的 PLset 大小为 M,那么最多会产生 M^N 种投票组合。然而,一旦某个 PL 更新,PLset 最多只会更新一个 PLout,也就是说,大多数投票组合是重叠的。因此,在所有 PL 都更新一次的情况下,最多需要更新次数如下:

$$N \times (N - 1) \times M \tag{2}$$

表现最佳的 PLout 投票组合将被记录下来作为最终的投票委员会,用于产生最终的输出结果。

本文直接将所有 PLout 的分数相加作为投票结果,除了具有效率方面的优势外,这种设计还有更多优点:

1) 可扩展性

该架构能够轻松适应类似的任务。例如,如果应用场景只需要 3 种情感类别,如乐观、消极和中性,本文需要设计一个带有三分类分数的解码器。当另一个应用场景需要更详细的分析时,可以通过

扩展标签来重复使用，比如将“消极”细分为“悲伤”和“焦虑”。相反，在一个对类别要求较少的系统中，无需重新训练，只需合并一个 PL 中相似的标签即可。

2）缓解过拟合

过拟合是基于深度学习的方法面临的最大问题之一。具有超强学习能力的模型会损害其泛化能力，尤其是在完全端到端的模型中[29]。因此，通过手动规则适当限制学习能力将有助于缓解过拟合问题。

1.3　原理理解

在本节中，将讨论该架构的有效性，并提供一个基于数学分布的视角来理解这一架构。

假设现在有 PL_a 和 PL_b，两个 PL 的输出分布分别为 D_{PLa} 和 D_{PLb}。训练和评估数据集的标签分布为 D_{train} 和 D_{eval}。所以训练就意味着让 D_{PLa} 和 D_{PLb} 接近 D_{train}。

图 3 展示了最终稳定状态的两种可能示例，一种是更趋向于正向的情况，另一种是更趋向于负向的情况。每个区域都表示一种数学分布，并且重叠区域与投票机制中的理念相契合。有几个细分区域值得关注：

1）有效重叠区域

$$D_{PLa} \cap D_{PLb} \cap D_{eval} \tag{3}$$

这个区域能够增强架构的部分稳定性。

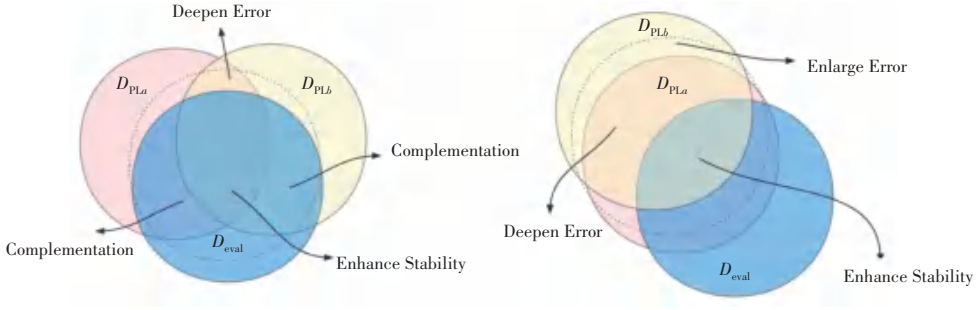


图 3　积极（左图）和消极（右图）情况的分布示例

Fig. 3　Distribution examples for positive (left) and negative (right) situation

2）互补区域

$$(D_{PLa} \cup D_{PLb} - D_{PLa} \cap D_{PLb}) \cap D_{eval} \tag{4}$$

该区域是提升架构性能的最重要条件。多个 PL 可以相互补充，从而增强正向影响区域相对于负向影响区域的优势，但这种提升并不稳定，因为该区域的概率密度（重叠区域）与其他负向影响区域相等，甚至更低。

对于具有相同模态输入的 PL 而言，这个区域会非常有限，但当输入为语音和文本等不同模态时，其能发挥更大的作用。

3）错误重叠区域

$$D_{PLa} \cap D_{PLb} - D_{eval} \tag{5}$$

该区域会加深架构的预测误差，其与“有效重叠区域”结构相似，均为不同 PL 之间的重叠部分，区别在于二者是否与 D_{smal} 重叠。上述特征使得两类区域存在一种零和博弈关系。在很大程度上，这个区域的负面影响程度取决于每个 PL 的分类性能，这也是本文设计贪婪策略来维护 PLset 的原因之一。

4）扩大错误区域

$$D_{PLa} \cup D_{PLb} - D_{PLa} \cap D_{PLb} - D_{eval} \tag{6}$$

该区域会扩大架构的错误部分。例如，单个 PL 可能会给出一个置信度较低的正确结果（即输出分数较低），而其他 PL 给出的错误结果只需要稍高一点的置信度就可以掩盖这个正确结果。

显然，可以总结出影响该架构性能的 4 个因素：

（1）训练数据集与评估数据集之间的相似度；

（2）每个独立 PL 的性能表现；

（3）PL 的数量规模；

（4）PL 输出分布的差异性。

从概率视角分析，这些因素越有利，最终获得的结果就越好。值得注意的是，在准确率性能方面存在一个理论理想上界 UB：

$$(D_{PLa} \cup D_{PLb}) \cap D_{eval} \tag{7}$$

然而，受限于实际数据分布的复杂性，现有方法无法直接求解 UB 的精确值，但可以通过间接的方式得到一个近似值。假设总共有 N 个 PL，第 n 个处理线路（PL）的正确预测集合记为 $C^{(n)}$，那么 UB 可以近似地表示为下式：

$$UB \approx | C^{(1)} \cup C^{(2)} \cup \cdots C^{(n)} | \tag{8}$$

换言之，在掌握所有独立 PL 的分类性能后，就可以计算出整个架构的近似上限 UB。此外，PL 表征

模式的相似性可通过交并比(IoU)进行量化评估:

$$\mathrm{IoU}=\frac{|C^{(1)}\cap C^{(2)}\cap\cdots C^{(n)}|}{|C^{(1)}\cup C^{(2)}\cup\cdots C^{(n)}|}\approx\frac{|C^{(1)}\cap C^{(2)}\cap\cdots C^{(n)}|}{UB} \tag{9}$$

IoU 的值越低,代表各个 PL 之间的相似性就越小。

2 配置 PL

本节描述了 PL 的具体配置细节。对于每种模态输入,本文选择了两个主干模型。对于语音输入,选择了 Wav2vec 2.0 和 WavLM;对于文本输入,则选择了 BERT[26] 和 XLNet[27]。所有这些主干模型都是预训练模型,以便高效提取具有差异且富含语义信息的表征。由于主干模型具备足够的学习能力,因此在主干模型之后的解码器采用了一个简单的线性分类层。本文还参考了文献[21]实现了一种多任务方法,其主干模型为 Wav2vec 2.0。因此,总共有 5 个 PL,分别简称为 S1、S2、T1、T2 和 M0。

2.1 S1 (Wav2vec 2.0)

Wav2vec 2.0 作为一种自监督的语音表征提取器取得了重大进展,能够从原始语音数据中提取特征。其包含 3 个模块:首先由堆叠的卷积神经网络(CNN)根据固定大小的感受野从原始语音中提取初步的隐藏表征;然后有一个基于 Transformer 的上下文网络,将生成的高级特征作为最终的语音表征结果。为实现自监督预训练,该模型额外引入了另一个量化模块,用于将来自上下文网络的上下文表征进行离散化处理。模型基于对比学习任务开展预训练,目标是识别被掩码的真实量化上下文表征。在预训练之后,可在上下文表征之后连接解码器执行下游任务,即进行微调。

2.2 S2 (WavLM)

Wav2vec 系列模型在语音识别领域取得了巨大成功,然而对于其他语音处理任务的探索却相对有限。WavLM 是首个针对全栈语音处理任务的预训练模型,其提出了掩码语音去噪方法,以提高模型的鲁棒性。该方法通过手动模拟嘈杂或重叠的语音作为输入,并对掩码区域的原始语音进行伪标签预测。这种方法使得 WavLM 在更多非语音识别(ASR)任务上取得了更好的结果。

WavLM 的整体结构与 Wav2vec 2.0 非常相似,本文中二者细节上的差异无需重点关注,只需确保二者具有不同的表征模式,即可验证所提出架构的有效性。

2.3 T1 (BERT)

BERT 是自然语言处理(NLP)领域中最经典的预训练模型之一。该模型在推动预训练和微调方法的发展方面发挥了重要作用。事实上,Wav2vec 系列、WavLM 等语音自监督预训练模型,在很大程度上都以 BERT 为参考。因此,二者存在许多相似之处,如 Transformer 架构、掩码方法以及模型规模等。

BERT 可以输出与一个词元(token)相对应的上下文表征,其中词元代表一个单词或具有某种特殊含义的标识(如用于表征整个句子的特殊词元)。下游任务可基于这些上下文表征来实现,如利用整个句子的表征进行情感识别。

2.4 T2 (XLNet)

BERT 是一个自编码器模型。其预训练过程是从掩码输入中重建原始数据,也被称为掩码语言建模(MLM)。然而,这一范式隐藏了掩码词元和未掩码词元相互独立的假设,使得该模型存在预训练与微调之间的差异问题。

XLNet 采用自回归方法来解决上述缺陷。自回归意味着只对单向上下文进行编码,在预训练阶段其不会对输入进行掩码处理,因此不存在预训练和微调之间的差异问题,但与此同时,对于深度双向上下文的处理效果不够理想。为此,XLNet 还提出了一种排列语言建模方法来改善自回归的这一问题,这种操作使得每个位置的上下文都可以由来自左侧和右侧的词元组成,即每个位置都能利用到所有位置的上下文信息。

2.5 M0 (多任务)

在获取隐藏特征之后,该特征被用于两项任务,其一是语音情感识别(SER),其二是采用联结主义时间分类[28](CTC)方法的自动语音识别[29](ASR)任务。该方法不仅能很好地利用多模态特征,还提升了模型的能力,也就是说,模型能够同时执行自动语音识别任务。

2.6 训练和推理

具体来说,给定语音输入为 X_S,文本输入为 X_T,并且情感标签记录为 Y。每个 PL 的输出为

$$\begin{cases}\mathrm{S1}(X_s)\\ \mathrm{S2}(X_s)\\ \mathrm{T1}(X_T)\\ \mathrm{T2}(X_T)\\ \mathrm{M0}(X_s)\end{cases} \tag{10}$$

使用交叉熵损失 $\mathcal{L}_{CE}$ 和 CTC 损失 $\mathcal{L}$ 分别作为语音情感识别和自动语音任务的优化目标，每个 PL 的语音情感识别任务的优化目标为：

$$\begin{cases} L_{\mathrm{S1}} = \mathcal{L}_{CE}(\mathrm{S1}(X_S), Y) \\ L_{\mathrm{S2}} = \mathcal{L}_{CE}(\mathrm{S2}(X_S), Y) \\ L_{\mathrm{T1}} = \mathcal{L}_{CE}(\mathrm{T1}(X_T), Y) \\ L_{\mathrm{T2}} = \mathcal{L}_{CE}(\mathrm{T2}(X_T), Y) \\ L_{\mathrm{M0_SER}} = \mathcal{L}_{CE}(\mathrm{M0}(X_S), Y) \end{cases} \tag{11}$$

对于 M0 而言，还存在另一个自动语音识别（ASR）任务的优化目标，将其记为：

$$L_{\mathrm{M0_ASR}} = \mathcal{L}_{\mathrm{CTC}}(\mathrm{M0}(X_s), X_T) \tag{12}$$

因此，M0 的训练目标为：

$$L_{\mathrm{M0}} = L_{\mathrm{M0_ASR}} + \alpha L_{\mathrm{M0_SER}} \tag{13}$$

其中，α 是一个超参数，用于调整多任务的权重。整个架构的优化目标可以计算为：

$$L = L_{\mathrm{S1}} + L_{\mathrm{S2}} + L_{\mathrm{T1}} + L_{\mathrm{T2}} + L_{\mathrm{M0}} \tag{14}$$

所有这些优化目标彼此相互独立，并且其中的每一个都代表一个完整的 PL。本文可以根据实际需求和应用场景对这些目标进行组合或拆分，适配不同设备或不同时间段内的部署要求。

在推理阶段，投票机制会作用于所有可用的处理 PL 集合（PLset）。该过程可以分为两个步骤。第一步是从每个 PL 独立地获取结果，此时，这些结果可以作为初步结果，如在要求低延迟的在线场景中；第二步是将所有分数相加，以此作为最终结果。

3　实验

3.1　数据集和评估标准

为将结果与大多数方法进行比较，本文选择了 IEMOCAP[30] 数据集。IEMOCAP 面向多模态、多说话人的情感分析任务构建，其中包括语音、文本转录、视频以及面部信息的动作捕捉数据。总共有 10 名演员，两人一组，在特定的情感场景下表演了 5 场戏。参照文献[18,21-22,31-34]，本文选取了 5 531 条带有“快乐”、“愤怒”、“中性”、“悲伤”和“兴奋”标签的话语，并将“兴奋”归并到“快乐”类别中。因此，最终有 4 个类别，数据集样本分布如图 4 所示。



图 4　IEMOCAP 数据集的数据分布

Fig. 4　The data distribution of IEMOCAP dataset

实验采用 10 折交叉验证的方式。对于每一折，单独留出一位说话人的数据作为测试集，使用其余九位说话人的数据进行训练，这样训练集和测试集之间就不会有说话人重叠的情况，并且第 k 折的测试集具有特定的规模。这种划分方法增加了稳健性评估的权重，也就是说，是在与说话人无关的情况下进行评估。

实验采用加权准确率（WA）作为核心评价指标，计算公式为：

$$\mathrm{ACC} = \frac{1}{10}\sum_{k=1}^{10} \mathrm{ACC}^{(k)} \tag{15}$$

其中 $\mathrm{ACC}^{(k)}$ 是第 k 折的准确率：

$$\mathrm{ACC}^{(k)} = \frac{|C^{(k)}|}{N_{\mathrm{eval}}^{(k)}} \tag{16}$$

本文设计了另外 3 个指标来衡量该架构的有效性，其中上一节提到的交并比（IoU）表示各 PL 之间的差异程度，而上限 UB 则表示该架构理论上的最高准确率。为更直观地说明，本文采用百分比的形式来表示上限，即：

$$\mathrm{UBP} = \frac{1}{10}\sum_{k=1}^{10} \frac{\mathrm{UB}^{(k)}}{N_{\mathrm{eval}}^{(k)}} \tag{17}$$

此外，本文使用交叉验证下不同折之间准确率结果的标准差来表示该架构的稳定性，其计算方式为：

$$\Delta = \sqrt{\frac{\sum_{k=1}^{10} (\mathrm{Acc} - \mathrm{Acc}^{(k)})^2}{10}} \tag{18}$$

3.2　参数设置

超参数设置参考表 1。学习率模仿 Transformer 架构：

$$\mathrm{lr} = k \times d^{0.5} \times \min(\mathrm{epoch}^{-0.5}, \mathrm{warm}^{-1.5}) \tag{19}$$

表 1　参数设置

Table 1　Parameter settings

PL	k	d	warm	BS	AS	PLS
S1	2e-6	512	10	1	8	4
S2	2e-6	512	10	1	8	4
T1	1e-6	512	10	32	1	4
T2	3e-6	512	10	32	1	4
M0	1e-5	512	10	1	8	4

注：BS：BatchSize；AS：Accumulation Steps；PLS：PLSet Size

M0 中多任务的权重设置为 0.01。选择 AdamW 作为优化器。每个 PL 都训练 100 个迭代，并且为了便于相互比较，主干模型都采用类似的基础模型，这些模型的参数规模与其他模型相近，并且隐藏特征的维度均为 768 维。

其中,UBP 与 IoU 用于评价的 PL 协同投票的特性,因此单 PL 基线条件下该两项指标不参与计算。

3.3 实验结果

本文方法与现有文献中的最佳结果对比见表 2。在交互模式中,当所有 PL 同时生效时,达到了最佳的当前最优结果 83.84%,这比最佳基线结果高出 4.93%,与最佳的单模态情况相比提高了 7.22%。一些文献也报告了未加权准确率(UA)指标,不过,本文选择了其中较好的结果来进行比较。

表 2 不同方法对比结果

Table 2 Comparison results of different methods

方法	ACC/%
文献[18]	67.40
文献[32]	71.15
文献[31]	72.60
文献[9]	72.73
文献[21]	78.15
文献[26]	79.58
文献[22]	82.68
本文	83.84

交互模式和同步模式的结果分别列于表 3 和表 4 中。本文给出了所有 PL 组合情况的结果,即总共有 31 种情况。可以明显地注意到,随着 PL 数量的增加,ACC 和 UBP 都会持续提升,IoU 会持续下降,图 5 绘制了相同 PL 数量的 ACC,IoU 和 UBP 取平均值的结果,还可以直观看出,随着 PL 数量的增加,ACC 和 UBP 的差值也越来越大,即 ACC 会更快收敛,这符合本文之前的理论分析。

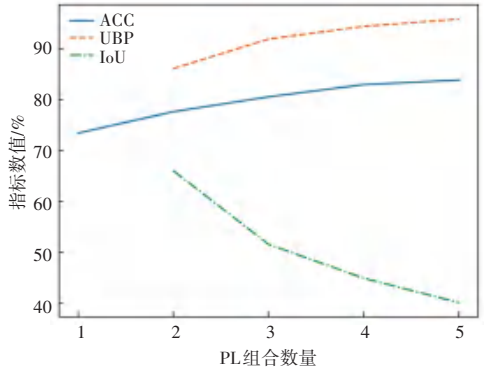


图 5 不同 PL 数量对应的 ACC、IoU 和 UBP

Fig. 5 ACC, IoU and UBP corresponding to different numbers of PL

除了准确率的提升之外,还发现当 PL 的数量增加时:1) 架构的稳定性有所提高,也就是说,标准差 Δ 会减小;2) PL 之间的差异程度增大,即 IoU 会降低;3)整个架构性能的上限会显著提高。

表 3 交互模式实验结果

Table 3 Results of interaction mode

PLs	ACC	Δ	UBP	IoU
Baseline				
S1	76.62%	3.29		
S2	73.59%	3.25		
T1	70.02%	2.21		
T2	68.0%	3.58		
M0	78.91%	2.35		
2-cross				
S1S2	78.55%	2.24	84.63%	73.24%
S1T1	78.96%	3.35	88.17%	64.01%
S1T2	76.51%	2.07	87.55%	60.51%
S1M0	80.73%	2.46	87.45%	73.49%
S2T1	77.99%	2.89	88.68%	59.41%
S2T2	76.04%	2.01	87.92%	56.22%
S2M0	80.36%	2.6	87.57%	70.15%
T1T2	70.3%	2.63	78.64%	72.18%
T1M0	79.86%	3.32	87.39%	67.89%
T2M0	77.01%	2.6	87.05%	63.34%
3-cross				
S1S2T1	81.15%	2.53	92.47%	51.8%
S1S2T2	81.2%	2.44	92.17%	48.72%
S1S2M0	81.19%	2.72	91.08%	61.21%
S1T1T2	77.38%	2.16	91.03%	51.53%
S1T1M0	82.45%	2.39	92.51%	55.19%
S1T2M0	82.23%	2.15	92.2%	52.33%
S2T1T2	77.49%	1.94	91.39%	48.05%
S2T1M0	82.5%	2.51	93.02%	52.25%
S2T2M0	82.42%	2.29	92.81%	48.47%
T1T2M0	77.51%	2.24	90.24%	55.38%
4-cross				
S1S2T1T2	82.91%	2.33	94.19%	42.37%
S1S2T1M0	82.65%	2.37	94.72%	47.64%
S1S2T2M0	83.28%	2.2	94.53%	44.33%
S1T1T2M0	83.28%	2.28	93.96%	46.97%
S2T1T2M0	82.39%	2.27	94.51%	43.34%
5-cross				
—	83.84%	2.13	95.81%	40.13

注:Δ 这实验结果标准差;空格代表该指标在此条件下不适用,UBP、IoU 仅针对两个及以上 PL 的组合进行计算。

从多模态的角度来看,具有相同模态输入的 PL

的 IoU 更高，这会导致 UBP 与 AC 相对偏低。例如，S1 和 S2 的 IoU 高达 73.24%，而 S2 和 T2C 的 IoU 仅为 56.22%，这一实验结果与理论分析也是一致的。

表 4　同步模式实验结果

Table 4　Result of synchro mode

PLs	ACC	Δ
	Baseline	
S1	76.62%	3.29
S2	73.59%	3.25
T1	70.02%	2.21
T2	68.00%	3.58
M0	78.91%	2.35
	2-cross	
S1S2	77.31%	3.25
S1T1	78.45%	3.44
S1T2	74.11%	2.47
S1M0	79.78%	2.59
S2T1	77.06%	2.78
S2T2	74.26%	2.70
S2M0	79.37%	2.72
T1T2	69.44%	2.92
T1M0	79.10%	3.34
T2M0	74.28%	3.30
	3-cross	
S1S2T1	79.94%	2.75
S1S2T2	79.69%	2.55
S1S2M0	79.70%	3.09
S1T1T2	76.06%	2.00
S1T1M0	81.47%	2.89
S1T2M0	81.00%	2.21
S2T1T2	76.15%	1.78
S2T1M0	81.25%	2.61
S2T2M0	81.05%	2.19
T1T2M0	75.97%	2.28
	4-cross	
S1S2T1T2	80.79%	2.26
S1S2T1M0	81.33%	2.65
S1S2T2M0	81.59%	2.60
S1T1T2M0	80.94%	2.49
S2T1T2M0	80.91%	2.14
	5-cross	
	81.98%	2.51

至于多任务的情况，尽管在多任务方法中利用了多模态信息，使其比单模态的 PL 获得了更好的准确率，但与不同单模态 PL 的组合相比，它的 IoU 仍然较低。例如，S2 和 T2 的 IoU(56.22%)远高于与 M0 组合时的 IoU(48.47%)。

针对稳定性指标 Δ 展开分析，在同步模式下，PLouts 的组合是随机的，导致其稳定性和准确率比交互模式下更差。并且由于这种随机性，本文没有给出同步模式下的 UBP 和 IoU 指标。

4　结束语

该架构不仅适用于语音情感识别任务，理论上也可推广至各类分类任务。后续研究将尝试把该架构应用于图像分割等更通用的任务场景。

就准确率而言，尽管所提出的架构已经取得了优异的成果，但理论上限分析表明仍有很大的提升空间。后续可以从以下 3 个方面进一步开展工作：

(1)多样化解码器

当前 PL 中仅采用线性层网络作为解码器，未能充分发挥该架构的潜力。可以探索更多样化的结构，如多头机制、注意力机制等。

(2)智能投票机制

现有投票机制规则过于简单，对所有 PLset 赋予相同权重。后续可以在投票机制中设计一种更智能的权重分配机制，以便更好地利用不同 PL 的优势[35]。

(3)更强大的主干模型

目前选择的主干模型是基础规模的，根据文献[20,36]的结果表明，增大模型规模，如选择大规模模型，将进一步提升任务准确率。

参考文献

[1] 杨卓，麦尔渊，李惠聪，等. 增强现实上下文感知最新研究进展综述[J]. 计算机工程与应用，2024，60(14)：14-25.

[2] 罗莹莹. 多模态情感识别综述[J]. 计算机应用文摘，2024，40(19)：194-197.

[3] 张晓莉. 改进 MFCC 特征和 MLA 模型的语音情感识别[J]. 福建电脑，2024，40(1)：52-56.

[4] 李磊，朱永同，杨琦，等. 基于多任务学习与注意力机制的多层次音频特征情感识别研究[J]. 智能计算机与应用，2024，14(1)：85-94.

[5] 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)：22245-22269. DOI：10.1109/ACCESS.2021.3068045.

[6] Kaur S，Kulkarni N. Recent trends and challenges in human computer interaction using automatic emotion recognition：A review[J]. International Journal of Biometrics，2024，16(1)：16-43.

[7] Rathi T，Tripathy M. Analyzing the influence of different speech data corpora and speech features on speech emotion recognition：A

review[J]. Speech Communication, 2024, 162:103102. DOI:10.1016/ j. specom. 2024. 103102.
[8] 陶建华, 陈俊杰, 李永伟. 语音情感识别综述[J]. 信号处理, 2023, 39(4):571-587.
[9] Nasersharif B, Namvarpour M. Exploring the potential of Wav2vec 2. 0 for speech emotion recognition using classifier combination and attention-based feature fusion[J]. The Journal of Supercomputing, 2024. DOI:10. 1007/s11227-024-06158-x.
[10] Chakhtouna A, Sekkate S, Adib A. Unveiling embedded features in Wav2vec2 and HuBERT models for speech emotion recognition [J]. Procedia Computer Science, 2024, 232:2560-2569. DOI:10. 1016/j. procs. 2024. 02. 074.
[11] ChenS, Wang C, Chen Y, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518. DOI:10. 1109/JSTSP. 2022. 3188113.
[12] Kahn J, Zheng W. Libri-Light: A Benchmark for ASR with limited or no supervision [C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 7698-7693. DOI: 10. 1109/ICASSP40776. 2020. 9052942.
[13] Chen G, Chai S, Wang G, et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio [C]//Proceedings of Interspeech. 2021: 3585 - 3589. DOI: 10. 21437/Interspeech. 2021-1965.
[14] Wang C, Morgane R, Lee A, et al. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation [C]//Proceeding of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021:993-1003. DOI:10. 18653/v1/2021. acl-long. 80.
[15] Wang Y, Lu C, Lian H, et al. Speech swin-transformer: exploring a hierarchical transformer with shifted windows for speech emotion recognition [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2024: 766 - 770. DOI: 10. 1109/ICASSP48485. 2024. 10447726.
[16] ZouH, Si Y, Chen C, et al. Speech emotion recognition with co-attention based multi-level acoustic information[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2022: 6907 - 6911. DOI:10. 1109/ICASSP43922. 2022. 9747095.
[17] Xu M, Zhang F, Khan S U. Improve accuracy of speech emotion recognition with attention head fusion[C]//Proceedings of 2020 10th Annual Computing and Communication Workshop and Conference. Piscataway, NJ: IEEE, 2020: 1-6. DOI: 10. 1109/CCWC47524. 2020. 9031193.
[18] LiR, Wu Z, Jia J, et al. Dilated residual network with multi-head self-attention for speech emotion recognition[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019: 6690 - 6694. DOI:10. 1109/ICASSP. 2019. 8683177.
[19] 陶建华, 范存航, 连政, 等. 多模态情感识别与理解发展现状及趋势[J]. 中国图象图形学报, 2024, 29(6):1607-1627.
[20] George S M, Ilyas P M. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise [J]. Neurocomputing, 2024, 568:127123. DOI:10. 1016/j. neucom. 2023. 127123.
[21] Cai X, Yuan J, Zheng R, et al. Speech emotion recognition with multi-task learning [C]//Proceedings of the Interspeech 2021. 2021: 4508-4512. DOI:10. 21437/Interspeech. 2021-2048.
[22] Lian Z, Tao J, Liu B, et al. Context-dependent domain adversarial neural network for multimodal emotion recognition [C]//Proceedings of the Interspeech 2020. 2020: 394-398. DOI: 10. 21437/Interspeech. 2020-2165.
[23] Li H, Tu M, Huang J, et al. Speaker-invariant affective representation learning via adversarial training[EB/OL]. (2019-11-04)[2026-07-03]. DOI:10. 48550/arXiv. 1911. 01533.
[24] Poria S, Cambria E, Hazarika D, et al. Multi-level multiple attentions for contextual multimodal sentiment analysis [C]// Proceedings of 2017 IEEE International Conference on Data Mining. Piscataway, NJ: IEEE, 2017:1027-1032.
[25] Padi S, Sadjadi S O, Manocha D, et al. Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models[EB/OL]. (2022-02-18)[2026-07-03]. https://arxiv. org/abs/2202. 08974.
[26] Makhmudov F, Kultimuratov A, Cho Y I. Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures[J]. Applied Sciences-Basel, 2024, 14(10): 17. DOI:10. 3390/app14104199.
[27] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized autoregressive pretraining for language understanding [C]// Proceedings of Advances in Neural Information Processing SystemsNeurIPS, 2019:5753-5764.
[28] Yao Z, Kang W, Yang X, et al. CR-CTC: Consistency regularization on CTC for improved speech recognition[EB/OL]. (2024-10-13)[2026-07-03]. https://arxiv. org/abs/2410. 05101.
[29] Prabhavalkar R, Hori T, Sainath T, et al. End-to-end speech recognition: A survey[J]. IEEE/ACM Transactions On Audio, Speech, and Language Processing, 2024:32:352-351.
[30] 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.
[31] Lu Z, Cao L, Zhang Y, et al. Speech sentiment analysis via pre-trained features from end-to-end ASR models[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 7379 - 7383. DOI:10. 1109/40776. 2020. 9052937.
[32] Li P, Song Y, Mcloughlin I, et al. An attention pooling based representation learning method for speech emotion recognition [C]//Proceedings of Interspeech. 2018: 1242. DOI: 10. 21437/Interspeech. 2018-1242.
[33] Wu X, Liu S, Cao Y, et al. Speech emotion recognition using capsule networks [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019:6695-6699.
[34] Wang Y, Boumadane A, Heba A. A fine-tuned Wav2vec 2. 0/ HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding[EB/OL]. (2022-10-03)[2026-07-03]. DOI:10. 48550/arXiv. 2111. 02735.
[35] Zhao F, Zhang C, Geng B. Deep multimodal data fusion[J]. ACM Computing Surveys, 2024, 56(9):36. DOI:10. 1145/3649447.
[36] Matasyoh N M, Zeineldin R A, Mathis-UllrichF. Optimising speech recognition using LLMs: An application in the surgical domain[J]. Current Directions in Biomedical Engineering, 2024, 10(1):45-48. DOI:10. 1515/cdbme-2024-0112.