arXiv:2603.09212v1 [eess.AS] 10 Mar 2026

口语情感的声学与语义建模

为获得
Doctor of Philosophy学位而
提交的论文，属于
Faculty of Engineering

BY

Soumya Dutta

印度科学理工学院电气工程系，班加罗尔 – 560012 （印度）

2026年3月

原创性声明

本人**Soumya Dutta**，学号为**04-03-00-10-12-20-2-19108**，特此声明，题为

口语情感的声学与语义建模

所呈现的论文是本人在**电气工程系**于**印度科学理工学院**2021-2026年期间完成的原创工作。

本人签字确认：

- 本人未对任何数据或结果进行篡改。
- 我未剽窃任何知识产权，并已清楚标明和注明他人的贡献。
- 我已明确致谢所有合作研究和讨论。
- 我已知悉，任何虚假声明都将导致严厉的纪律处分。
- 我已知悉，这项工作可能会接受学术不端行为筛查。

日期：　　学生签名

作为上述论文的导师，我证明据我所知上述陈述属实，并已尽到应有的审慎义务，以确保论文的原创性。

导师姓名：　　导师签名

© Soumya Dutta 2026年3月
版权所有

谨献给

我的父母和姐妹

感谢他们无条件的支持

我的妻子

感谢她的陪伴

致谢

若没有许多人的指导、支持与鼓励，我不可能完成这篇论文。

首先，我谨向导师 Sriram Ganapathy 博士致以最诚挚的谢意，感谢他让我有机会在声学模式学习与提取实验室（LEAP）工作。他给予我充分的自由，让我探索一个对我们双方而言都全新的研究方向，使我得以在学术和独立性方面不断成长。在过去五年中，曾有很长一段时间，尽管持续付出努力，却始终未能发表论文。我尤其感谢 Sriram 在那段时期给予我的耐心、鼓励与始终如一的支持。我将永远珍视我们之间的讨论——无论是技术性的还是非技术性的——它们深刻地塑造了我的思维与观点。

我还要感谢印度政府教育部（MoE）通过总理研究奖学金（PMRF）提供的资助。同样，我也非常感谢高通创新奖学金（QIF）对我研究工作的支持。此外，我还要感谢我在三星研究院和Adobe研究院实习期间的导师们。

我要感谢 LEAP 实验室的同事们——Samrat、Ameen、Prachi、Shreyas、Anurenjan 等人，感谢我们在实验室共度的时光。在关键时刻，那些简短的鼓励常常让我在这段旅程中始终保持动力。我也感谢我的合著者 Avni、Smruthi、Viveka、Varada，以及目前的合作者 Payal 和 Paras，感谢他们富有启发性的讨论和共同努力，正是这些促成了这项工作。

若没有在B餐厅结识的那群可爱的朋友——Debarpan、Aniket、Arnab、Sambit、Ujjwal Da和Akash，我在印度科学研究所的生活便不完整。除了Debarpan（他也是我的实验室同事）和Sambit，其他人都在2022/2023年离开了印度科学研究所，但我们在餐厅和宿舍里度过的漫长时光，以及那些对话——我常常兴致勃勃地占据着讨论的主导——仍是我在印度科学研究所最珍贵的回忆之一。也要衷心感谢我在来到印度科学研究所之前结识的朋友：Arup、Ayan、Suman、Kaustav D.和Kaustav C.。随着时间流逝，我们的联系

致谢

已大幅减少，但知道仍有一些关心我的人只需一个电话就能联系上，总是令人欣慰。

最后，我想感谢我的家人。过去五年来，父母和姐姐始终给予我支持与鼓励，对此我深怀感激。知道他们一直支持着我，或许已成了我习以为常的事，但我永远不会把这份支持视为理所当然。

谨以此献给我的妻子 Abhipsa——我们以朋友的身份进入印度科学研究所，并在这段旅程中成为人生伴侣。在过去五年里，我们见证了彼此频繁的低谷与偶尔的高峰，也希望我们携手变得更加坚强。期待在未来的岁月里，与她共同建立美好的伴侣关系。

摘要

情绪是基本的心理状态，为日常互动赋予人性，塑造信任、投入和社会联结。随着由大语言模型驱动的人工智能系统日益融入日常生活，如何使其准确识别、刻画并生成细粒度的人类情绪，仍是一项关键挑战。尽管情感表达本质上具有多模态特征，本文聚焦于通过口语传达的情绪。本文探讨如何对声学与语义信息进行联合建模，以促进语音情感理解与情感合成。

本论文的第一部分聚焦于通过预训练进行情感感知表征学习。现有针对语音和文本的自监督学习目标主要强调语言重建，往往对情感识别至关重要的情感线索表征不足。本论文提出了融合声学与语义监督的预训练策略，从而获得与口语情感特征更为契合的表征。此外，论文还提出了一种用于基于文本的情感识别的语音驱动有监督预训练流程，即使缺乏人工标注文本语料库，也能构建大规模的情感感知文本模型。

第二部分探讨了现实对话场景中的口语情感识别。对话情感识别（ERC）面临上下文依赖、说话人轮次和模态不平衡等挑战，因此需要在对话轮次间以系统化方式整合声学与语义线索。论文提出了基于跨模态注意力和专家混合融合的层次化建模框架，以捕捉对话结构和模态特定动态，并在多个基准数据集上取得稳健表现。

论文的最后一部分研究语音合成中的情绪风格迁移。本文提出了一种全新的情绪风格迁移定义，适用于无文本、非平行的语音到语音场景，其目标是在保留源说话人身份和语言内容的同时，将情感说话风格从参考话语迁移到源话语。通过

摘要

将情感迁移视为解耦表示学习问题，所提出的方法能够在现实条件下实现可控的情感变换。大量客观评估和主观评估表明，该方法能够有效实现情感迁移、说话人保持和内容不变性。论文进一步表明，情绪风格迁移可用作数据增强策略，以提升低资源环境下的情感识别性能。

总体而言，这篇论文强调了当前语音与语言技术（包括近期的大语言模型）在从口语中稳健地建模情感方面的关键局限性。通过采用声学—语义建模视角，所提出的方法有助于赋予未来的人工智能系统更加可靠的情感理解能力和表达能力。论文最后概述了将这项工作扩展至临床相关应用的方向，例如自动抑郁检测，凸显了口语情感感知建模更广泛的影响。

基于本论文的出版物

同行评审期刊论文

1. S. Dutta、S. Balaji、S. Ganapathy，《一种用于对话多模态情感识别的专家混合模型》，*Computer Speech & Language*（已接收发表）。

2. S. Dutta, S. Ganapathy，《利用内容和声学表征进行语音情感识别》，*IEEE Transactions on Audio, Speech and Language Processing*，第 33 卷，第 3678-3689 页，2025年。3. S. Dutta, A. Jain, S. Ganapathy，《无文本非平行语音到语音情感风格迁移》，*IEEE Transactions on Audio, Speech and Language Processing*（审稿中）。

同行评审会议论文

1. S. Dutta, S. Balaji, Varada R, V. Salinamakki, S. Ganapathy，《ABHINAYA——自然条件下语音情感识别挑战赛系统》，*Proc. Interspeech2025*年：第 4663-4667 页。2. S. Dutta, S. Ganapathy，《用于对话多模态情感识别的大语言模型监督预训练》，*Proc. ICASSP 2025*：第 1-5 页。3. S. Dutta, S. Ganapathy，《结合说话人解耦的零样本音频到音频情感迁移》，*Proc. ICASSP 2024*：第 10371-10375 页。4. S. Dutta, S. Ganapathy，《采用可学习前端和自注意力的多模态Transformer情感识别》，*Proc. ICASSP 2022*：第 6917-6921 页。

预印本

1. S. Dutta、S. Ganapathy，《HCAM——用于多模态情绪识别的分层交叉注意力模型》，*arXiv preprint arXiv:2304.06910 (2023).*

目录

致谢 ii摘要 iv基于本论文的出版物 vi目录 viii图表目录 xii表格目录 xvi

1 引言 **1**

1.1 情感理解与合成.11.2 口语中的声学与语义.31.3 问题陈述.41.3.1 口语中的情感感知预训练 . . . . . 41.3.2 多模态情感识别.51.3.3 语音到语音场景中的情感风格迁移 . . . . . 51.4 贡献概述 .61.5 本论文其余部分的路线图 .7

2 背景研究 **9**

2.1 情感建模的语音与文本特征.92.1.1 语音情感识别的特征.92.1.2 文本情绪识别的特征.102.2 情感感知预训练的相关研究.112.2.1 面向情感的语音模型预训练.11

2.2.2 面向情感的文本模型预训练.122.3 多模态情感识别的相关研究.122.3.1 对话情感识别.132.3.2 语音情感识别：多模态视角 132.4 富表现力语音合成的相关研究.142.4.1 文本转语音.152.4.2 语音转换.152.4.3 风格迁移.162.5 数据集.172.6 章节总结.20

3 情感——　口语感知预训练 2　3

3.1 语音情感识别预训练.243.1.1 所提方法.243.1.2 实验与结果.283.1.3 讨论.323.1.4 语音情感识别预训练总结 383.2 文本情绪识别预训练.393.2.1 所提方法.393.2.2 实验与结果.403.2.3 文本模型预训练总结.413.3 章节总结.42

4 多模态　　口语多模态情感识别4　3

4.1 对话情感识别（ERC）.444.1.1 问题描述.454.1.2 多模态 ERC 的层次化建模.464.1.3 ERC 的混合专家建模.524.1.4 ERC 建模总结.664.2 自然情境下的情绪识别.664.2.1 所提方法.674.2.2 自然条件下语音情感识别总结.754.3 章节总结.75

5 情感迁移　　语音到语音系统中的情感迁移 7 7

5.1 提出的方法.78

5.1.1 总体概览.785.1.2 内容编码器.795.1.3 说话人编码器.805.1.4 情感风格因素.805.1.5 语音重建.835.1.6 情感风格迁移.84

5.2 实验与结果 . 85

5.2.1 数据集与预训练 . 85

5.2.2 实现 86

5.2.3 评估设置 86

5.2.4 客观评估 指标 . 87

5.2.5 与基线方法比较 . 88

5.2.6 客观结果 89

5.2.7 主观结果 测试集 91

5.2.8 任意到任意情感风格迁移. 92

5.2.9 对加性噪声的鲁棒性 . 93

5.3 讨论 94

5.3.1 语音分词器和合成器的选择. 94

5.3.2 时长预测器的影响. 95

5.3.3 对抗损失的重要性. 95

5.3.4 在语音情感识别中的应用 . 96

5.4 本章小结.97

6 总结与未来拓展 99

6.1 贡献.996.2 局限性.1016.3 未来研究方向.1026.4 章节总结.104

A 架构参考组件 105

A.1 自监督语音模型.105A.1.1 PASE+ .105A.1.2 HuBERT.105

A.1.3 WavLM.105
A.1.4 data2vec.106A.2 SONAR.106A.3 Whisper.106A.4 SALMONN.107

参考文献 109

图表目录

1.1 多模态对话模型在口语情感理解方面存在局限性的示例。尽管六位人工标注者一致认为该话语表达了恐惧，一致率为 100%，Gemini 3（启用思考功能）却预测其情感为快乐，并给出了错误的解释。. 21.2 论文贡献概要.61.3 论文各章节路线图.7

3.1 所提出的 CARE 模型框图。模型的声学编码器以 PASE+ 特征为目标进行训练。蓝色块表示被冻结的组件或不含可学习参数的组件。对于语义编码器，Transformer 层被冻结，而卷积适配器则经过训练。由于声学编码器的输出维度为 768，因此添加了一个全连接层，以匹配 PASE+特征的 256 维度。推理时不会使用该全连接层以及语义编码器之后的平均池化模块。 . . 253.2 使用不同声学目标时 CARE 的性能。以 eGeMAPS 作为特征的模型采用与 PASE+ 基线类似的方式进行训练。所有数值均为 5 次随机初始化的平均值.373.3 使用不同语义目标时 CARE 的性能。所有数值均为 5 次随机初始化的平均值.373.4 WavLM 在 MSP-PODCAST 上持续预训练时的性能比较。图中还给出了 CARE 的声学编码器与通用编码器组合的性能，以供参考。此处，RAVDESS 指 RAVDESS-Song 数据集.383.5 推理模型规模（以百万为单位的参数量）与语音情感识别性能（8 个数据集上的加权 F1 分数平均值）的散点图。与现有方案相比，CARE 实现了更好的权衡。更多详情请参见表格 3.2和3.4及相关讨论.39

3.6 使用自动语音识别系统生成预训练数据的转录文本，并由大语言模型（LLM）将其标注为正面、负面或中性情感。这些“银标”及其对应的文本转录构成了 RoBERTa-large 模型的监督训练数据集.403.7 RoBERTa-large 模型在不同数据集上的性能。采用不同的大语言模型，从语音转录文本中生成伪情绪标签。文中还报告了未经监督微调的预训练 RoBERTa 的性能.414.1 来自 ERC 数据集 MELD 的一段对话示例。此示例展示了多名说话者在对话中互动时情绪如何变化 . . . . . 454.2 所提出模型的框图。图中还标出了训练的三个阶段.464.3 所提出模型中使用的协同注意力网络。它由两个子块组成——交叉注意力块和自注意力块.484.4 (a) 单模态特征提取模块的训练；(b) MiSTER-E 的完整流程。训练流程的其余部分时，语音和文本嵌入模块均被冻结。针对两种模态分别训练两个上下文添加网络，并同时训练一个多模态网络。最后，训练一个混合专家门控网络，以预测每个话语的情绪类别。 . . . 524.5 (a) 上下文添加网络（用于语音模态）和 (b) MiSTER-E 中使用的多模态网络。这两个块的输入均源自单模态特征提取模块。 TIN 表示时间 Inception 网络， MHA 表示多头注意力.544.6 IEMOCAP 和 MELD 上的模型性能比较。(a) MiSTER-E 中不同混合专家门控策略的效果。(b) 单模态语音大语言模型和大语言模型在两个数据集上的性能.624.7 (a) 不同数据集上各专家的权重分布；(b) 各专家和 MiSTER-E 在两个数据集上的性能 . . . . . 634.8 (a) IEMOCAP 的混淆矩阵。(b) IEMOCAP 和 MELD 上按类别划分的 F1 分数，展示了不同情绪间的差异.644.9 MELD 中的一项案例研究，其中情绪在整个对话过程中不断变化。该对话包含 7 个话语，MiSTER-E 正确预测了其中 6 个.65

4.10 Abhinaya 语音情感识别系统各组件示意图。我们使用三种类型的模型：仅语音（S1、S2）、仅文本（T1、T2）和语音-文本（ST1）。只有 T1 用于零样本设置。模型使用的文本是由 Whisper 生成的自动语音识别转录文本.68 4.11 训练和验证数据中 8 种情感类别的分布 ..... 70 4.12 使用自动语音识别转录文本，在零样本设置下评估不同大语言模型 [141, 145, 8, 146]所得的验证宏F1分数（%）。所考察的 LLaMA模型为 Instruct版本。同时也展示了基线性能。..... 73 5.1 (a) S2S-ZEST 训练和 (b) 风格迁移范式概览。情感风格因素以不同颜色标示。风格迁移过程中，源语音标记经过内容编码器和说话人编码器，而时长预测器、基频重建模块和情感分类器模块则接收来自参考语音的输入.78 5.2 用于提取说话人嵌入的模型。GRL 表示梯度反转层。训练期间，蓝色块保持冻结。..... 80 5.3 音高轮廓重建模块——将说话人嵌入 (s) 与帧级情感嵌入 E 相加，构成键值序列；源语音词元嵌入 (C) 构成查询序列。交叉注意力模块的帧级输出经过使用一维卷积神经网络的位置前馈网络，以重建基频轮廓 (ˆf)。交叉注意力模块的架构也一并展开显示，以供参考。..... 81 5.4 分析阶段从语音中提取的不同因素。情感分类器采用说话人对抗损失进行训练。帧级嵌入 (E)、说话人嵌入 (s) 和语音标记 (t) 用于重建基频轮廓 (ˆf)。此外，话语级情感嵌入 (¯e) 与去重词元 $t' = \{t_1, ..., t_{T'}\}$ 一同用于预测每个词元的时长 (ˆd)。所有蓝色块在训练期间均保持冻结，而黄色块则参与训练。灰色块不含任何可学习参数。83

5.5 情感风格迁移——从参考语音中提取帧级 (E^{ref}) 和话语级 ($\bar{e}^{ref}$) 嵌入。使用源词元 t'、说话人向量 s 和情感嵌入 $\bar{e}^{ref}$ 进行时长预测。利用这些预测时长 d^{conv} 生成重复词元序列 t^{conv}。结合此词元序列、E^{ref} 和说话人嵌入 s，预测基频轮廓 $\hat{f}^{conv}$。最后，将词元序列、说话人和情感嵌入以及预测的基频轮廓输入 BigVGAN 模型，生成转换语音.84 5.6 不同测试设置下的主观评估。所用缩写：MOS—平均意见分，SMOS—相似度平均意见分。图中也显示了 95% 置信区间.90 5.7 不同测试设置下的主观评估。SMOS 表示相似度平均意见分。★ 表示基线与 S2S-ZEST 之间的分数差异具有统计显著性（$p < 0.05$） ..... 92 5.8 不同情感对之间的平均情感相似度。..... 93 5.9 使用不同语音分词器和合成器组合时，重新合成语音的词错误率.94 5.10 两位说话人 I 和 II 的说话人嵌入 t-SNE图，分别展示有无对抗损失的情况。图中颜色对应 ESD 中的 5 种情感类别.96

表格目录

2.1 本论文所用数据集及其在不同章节中作用的概述。 213.1 评估数据集概述。最后一列表示训练集/测试集中的说话人是否重叠.283.2 与其他下游数据集相关研究的比较。 # 在预训练中包含下游数据集的模型。结果中以 **粗体**和下划线分别标出最佳模型和次佳模型。所有数值均为基于 5 次随机初始化计算的加权 F1 分数（显示均值和标准差）。括号中还显示了非加权平均召回率。 . . . . . 303.3 MSP-IMPROV 数据集的结果。CCC 表示一致性相关系数，V、A、D 分别表示效价、唤醒度和支配度。 323.4 不同下游数据集上的基线结果，以加权 F1 分数表示。所有数值均为下游网络 5 次随机初始化的平均值。本表还展示了 CARE 各个组成部分的结果.33

3.5 修改语义编码器后在下游数据集上的结果（加权 F1 分数）.353.6 使用不同声学编码器初始化方式时，在下游数据集上的结果（加权 F1 分数）.364.1 不同数据集上的结果（加权 F1 分数）.504.2 不同训练范式下系统的加权 F1 分数 514.3 问题形式化与模型描述中使用的符号.53

表格目录

4.4 使用加权 F1 分数比较不同方法在 IEMOCAP、MELD 和 MOSI 数据集上的表现。所有报告的结果均为语音-文本系统的结果。 ##我们报告了训练 SMIN 时未使用外部情感数据的结果。 # 我们在自己的实验设置下运行了作者提供的公开实现。上标结果表示在适用情况下，3 次随机初始化的均值和标准差.604.5 对部分基线方法进行重新设计，比较其使用和不使用大语言模型特征时的结果。结果表明，即使对于 OPENSMILE 和 wav2vec 这类非大语言模型表征，该方案也适用.614.6 不同模型在平衡验证集和测试集数据上的性能（%）。未报告某些模型在测试数据上的性能，因为这些模型未提交至排行榜。 ∗ 表示由组织者在官方竞赛截止日期之后评估的测试分数.714.7 使用不同损失函数时模型的性能。所有结果均为平衡验证数据上的宏平均 F1 分数（%）。**加粗** 表示每个模型的最佳损失函数.73Abhinaya 及不同模型组件在平衡验证数据上的逐类别宏平均 F1 分数（%）。**加粗** 表示每个情绪类别的最佳模型.74使用 Abhinaya 中不同模型组合进行多数投票的结果（%）.755.1 情感风格迁移的各种评估设置.865.2 客观评估结果。此处，WER 表示词错误率，情感相似度表示根据 emotion2vec 嵌入计算的情感相似度，说话人相似度表示根据 Resemblyzer 嵌入计算的说话人相似度。词级皮尔逊相关系数指词语速率皮尔逊相关系数。不同测试设置见表 5.1。 ∗ 表示这些条目未**标为粗体**或加下划线，因为 StarGANv2-EST 不适用于 7 种测试设置中的 2 种。 885.3 中性到情感设置下对叠加 MUSAN 噪声的鲁棒性. . . . . 935.4 使用 S2S-ZEST 和不使用持续时间预测的 S2S-ZEST，在 ESD 数据集的四种测试设置下，WER、情感相似度及词级皮尔逊相关系数（PCC）.95

表格目录

5.5 使用 S2S-ZEST、未采用情感对抗损失的 S2S-ZEST（$\lambda_{adv}^{emo} = 0$）以及未采用说话人对抗损失的 S2S-ZEST（$\lambda_{adv}^{spk} = 0$）在 ESD 数据集的四种测试设置下的 WER、情感相似度和说话人相似度.965.6 不同训练条件下采用或不采用数据增强时，IEMOCAP 4 类和 6 类分类的加权 F1 分数。Orig. 表示使用全部训练样本；100 和 200 表示每个类别仅使用 100 或 200 个样本来训练模型。No Aug. 表示未使用增强数据，而使用合成样本的实验标记为 Aug.96

6.1 Google 自动语音识别系统在所使用的三个数据集上的词错误率（WER）（%）。所用缩写：快乐：*Hap.*，中性：*Neu.*，愤怒：*Ang.*，兴奋：*Exc.*，沮丧：*Fru.*，厌恶：*Dis.*，积极：*Pos.* 和消极：*Neg.* 对于 IEMOCAP，数字指合并后的训练集和验证集。 . . . . 1036.2 我们的系统在测试时使用数据集上的自动语音识别转录文本所得的加权 F1 分数 . 103

第1章

引言

情感在人类交流中发挥着核心作用，影响着信息在社会互动中的表达、感知和回应方式。人类可以通过多种方式表达情感，例如面部表情 [1]，语音 [2]，手势 [3]，生理信号 [4]，，或综合运用这些方式。在这些模态中，口语在交互式交流中传达情感方面尤为重要。随着对话式人工智能系统日益采用语音界面，近期研究 [5] 表明，与文本交流相比，用户在表达或查询情绪状态时明显更偏好语音交互。支持此类交互不仅需要具备从语音中理解情感的能力，还需要具备生成具有情感表现力的语音的能力。

1.1 情感理解与合成

尽管语音与语言技术取得了长足进展，但从口语中稳健地理解情感仍是一项艰巨任务。基于大语言模型（LLM）的对话代理在文本交互方面展现出强大的能力，并日益应用于情感支持和心理健康相关查询等对情感敏感的场景 [6, 7]。这些模型越来越多地以多模态方式进行训练，从而能够处理语音和文本输入 [8]。

然而，仅能获取语音输入并不能保证可靠的情感理解。图1.1 展示了一个典型的失败案例：尽管存在声学线索，大语言模型（LLM）仍将表达恐惧的口语话语错误地推断为表达快乐。值得注意的是，六位彼此独立的人工标注者均将同一话语标注为恐惧，一致率达 100%，这表明人类听者对其情感意图的判断毫无歧义——



图1.1：多模态对话模型在口语情感理解方面局限性的示例。尽管六位人工标注者一致认为该话语表达的是恐惧，一致率为 100%，Gemini 3（启用思考功能）却预测其表达的是快乐，并给出了错误的解释。

而模型却做出了错误预测，并给出了有缺陷的解释性推理。这凸显了当前系统的一项局限：情绪推断往往依赖浅层语义启发式或不完整的声学推理，而非对语音中与情绪相关的声学和语义因素进行有原则的建模。这一局限不仅影响情感理解，也会影响语音生成：要生成情感恰当的口语回应，就必须明确控制语音声学特征与其语言内容之间的相互作用。

这表明，语音情感的理解与合成都面临一个核心挑战：情感意图并非由语言内容或声学实现单独决定，而是源于二者的相互作用。应对这一挑战，需要原则性地理解声学与语义信息如何在语音中被共同编码和感知。

1.2 口语中的声学与语义 e

贯穿本论文，我们始终通过这一声学—语义视角来看待语音中的情感建模，并探究这两类信息来源如何被有效学习、结合与解耦，以服务于情感理解与合成。

虽然声学和语义线索共同促进情感理解，但如何学习能够跨数据集泛化的表征，仍是语音情感识别的一项核心挑战。wav2vec [9], HuBERT [10], 和 WavLM [11]等自监督语音模型已展现出在广泛语音任务上的强大泛化能力。然而，它们的预训练目标主要针对语言重建进行优化，却忽略了对情感表达至关重要的细粒度声学变化。这一差距促使人们探索情感感知预训练策略，以联合建模与情感线索相一致的声学与语义因素。

除了话语级情感识别，许多现实应用还需要理解情绪如何随着对话进程而变化。这种设置通常称为对话情感识别（ERC） [12], ，涉及对多个话语和说话人之间上下文依赖关系的建模。尽管结合声学与语义信息已被证明对多模态情感识别 [13, 14, 15] 和抑郁症检测 [16, 17], 等任务有效，但由于对话场景具有很强的上下文依赖性，准确建模情绪仍颇具挑战。此外，现有的大多数对话情感识别数据集都是在受控或表演条件下收集的，因此难以反映自然语音中的多样性。

除了情感理解之外，要生成情感恰当的语音，还需要控制声学与语义因素在输出信号中的实现方式。这促使我们研究解耦表示学习 [18], ，其目标是分离语言内容、说话人身份和情感说话风格等组成部分。尽管这种因子分解已应用于说话人识别 [19] 和自动语音识别 [20], 等任务，但其最主要的应用领域是语音转换（VC），其目标是在保留语言内容的同时改变说话人身份 [21]。基于这一思路，可以设想一种可控语音合成的新范式：仅将情感说话风格从参考信号迁移到目标话语，同时保持说话人身份和语言内容不变。这项任务称为情绪风格迁移 [22], ，需要以合理的方式将语音分解为声学和语义组件。

在这篇论文中，我们以这些进展为基础，探讨推动口语情感建模发展的三个相互关联的方向。首先，我们研究情感感知的表征学习技术，这类技术通过联合利用声学与语义信息，捕捉语音中与情感相关的变化。

其次，我们探讨如何有效利用语音和文本表征进行口语情感识别，尤其关注对话和自然场景。第三，我们探索声学与语义因素的解耦，以实现语音中可控的情感变换，并以情感风格迁移任务为例。这些方向共同构成了一个统一的框架，用于理解和合成口语中编码的情感信息。

1.3 问题陈述

Th本文考虑的不同问题陈述将在下文详细介绍 l.

1.3.1 语音语言中的情感感知预训练

大多数针对语音和文本的预训练算法都会对原始信号或文本的部分内容进行掩码，并训练模型预测被掩码的部分。这种范式因掩码语言建模（MLM）而广为人知 [23], ，推动自监督表征学习取得了显著进展。然而，现有预训练目标主要针对语义重建进行优化，这使模型偏重与内容相关的信息，而忽视了富有表现力的声学变化。由于语音中的情感意图通常通过韵律、能量和时间动态来传达，这限制了此类表征在情感识别方面的效用。本文探讨两个与设计更契合情感识别的预训练目标相关的问题：

- 语义和声学线索都对情感理解至关重要；然而，当前的自监督学习框架鲜少对这些因素进行联合建模。尽管近期有些方法探索了面向语音的情感感知预训练 [24, 25], ，但仍忽视了语义与声学的互补作用。本文研究能够同时利用这两类因素的预训练策略，并提出一种统一方案，用于语音中的情感感知自监督学习。

- 近期研究表明，可以通过监督或蒸馏，将特定任务知识从大型预训练模型迁移到较小的模型或特定任务模型中 [26]。我们提出一种互补的问题设置：首先将口语话语转录为文本，并使用大语言模型（LLM）推断情感标签并为文本添加标注。随后，利用生成的带情感标签的文本语料库预训练语义表征，以用于情感识别；即使在

无法获取人工标注的情感数据时，也能进行情绪感知文本预训练。

1.3.2 多模态情感识别

所研究的问题 属于口语情感识别的范畴 :

•对话情感识别（ERC）是一个序列到序列分类问题，其中，由两个或更多说话人的多个话语组成的对话被映射为情感标签序列。该任务要求对长期上下文依赖以及说话人之间的交互进行建模。现有的大多数方法采用端到端架构处理 ERC，同时执行上下文建模和多模态融合。在本论文中，我们通过层次化建模或模块化网络架构，采用将上下文建模与多模态融合明确分离的框架来研究 ERC。

• 传统的语音情感识别（SER）系统主要依赖声学特征，例如梅尔频谱图，或使用 OpenSMILE 等工具包提取的手工描述符 [27]。尽管一些先前的研究纳入了语言信息，但这些方法通常是在受控或模拟的设置下进行评估，并且能够获得真实转录文本。在本论文中，我们关注自然条件下的 SER，并探索使用 ASR生成转写，将直接从语音中提取的声学线索与语义线索结合起来。

1.3.3 语音到语音中的情绪风格迁移设置

随着语音合成在个人助理 [28]、娱乐 [29], 和机器人 [30], 等应用中的日益普及，对可控且富表现力语音生成的需求显著增加。因此，近期文本转语音（TTS）研究主要聚焦于对音高、能量、说话风格和情感等富表现力属性进行建模 [31]。虽然这些方法能够对合成语音进行细粒度控制，但通常依赖文本输入。

相比之下，本论文探索了一种不同的范式：情感风格迁移 （EST），应用于语音到语音场景。给定源语音信号和传达目标情感的参考语音信号，目标是将参考语音的情感说话风格迁移到源语音，同时保留源语音的说话人身份和语言内容。我们在无文本 且非平行 设置中研究这一问题，其中转录文本不可用，且不存在同一说话人以不同情绪说出相同内容的配对录音。这一设置带来了独特的挑战，并需要有效解耦情感、语义和说话人相关因素。

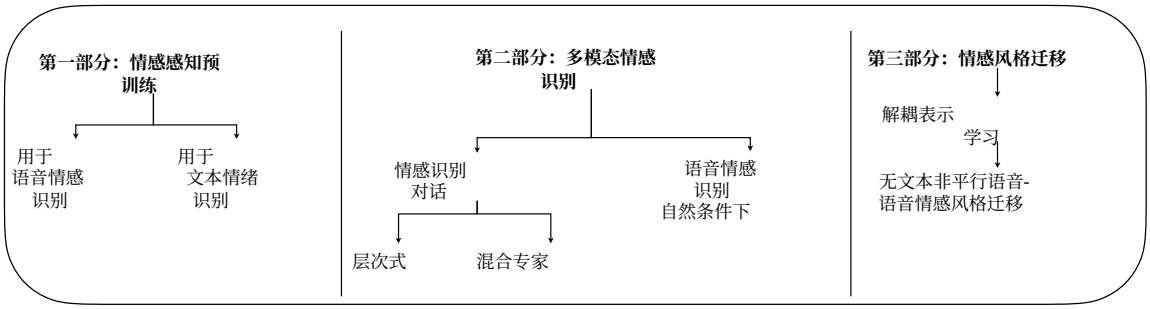


图 1.2：论文贡献总结

1.4 贡献概述

本论文沿着三个广泛且相互关联的研究方向做出贡献：（i）语音和文本模型的情感感知预训练，（ii）口语情感识别，以及（iii）语音到语音场景中的情感风格迁移。这些贡献的概览如图1.2。

- **情感感知预训练：**我们研究了如何设计更契合口语情感理解的预训练目标。我们不再局限于传统的掩码预测目标，而是探究能够显式考虑语音中声学与语义因素的预训练策略。结果表明，我们提出的方法在八个语音情感识别基准数据集上具有更好的泛化能力，同时保持了较高的计算效率。此外，我们提出了一种框架，利用自动语音转写和大语言模型生成情感标注文本数据，使得即使只有口语，也能对基于文本的模型进行情感感知预训练。这为人工标注文本资源有限的场景提供了一种实用的情感感知文本表示学习策略。

- **多模态情感识别**我们提出了层次化建模方法，以应对对话情感识别（ERC）的复杂性。我们研究了模块化融合策略，包括混合专家架构，以有效整合不同对话轮次中的声学和语义线索。我们提出的架构在广受欢迎的对话情感识别数据集上取得了新的最先进成果。此外，我们还研究了现实且自然的条件下的语音情感识别（SER），重点关注对数据不平衡的鲁棒性。作为自然条件下的语音情感识别的一部分

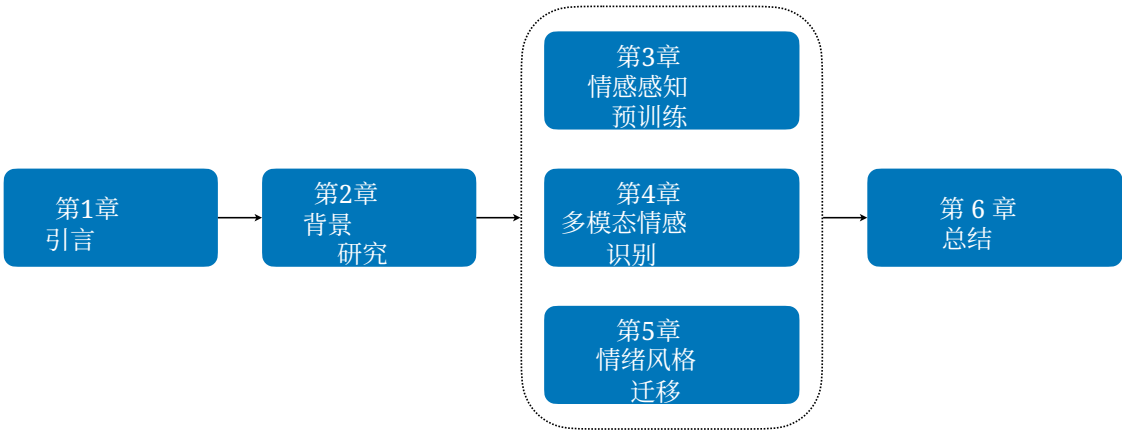


图 1.3：论文各章节的路线图

条件挑战在*Interspeech 2025*中，我们探索了语音大语言模型（SLLMs）、大语言模型（LLMs）以及 WavLM 等传统自监督语音模型在情绪分类中的应用。我们探索了不同的损失函数，以缓解类别不平衡问题。我们在挑战赛中的表现证明了所提出的建模策略在现实场景中的鲁棒性。

- **语音中的情绪风格迁移：**我们将情绪风格迁移构建为一个无文本、非平行语音到语音场景下的解耦表征学习问题。据我们所知，本研究是首批无需相同内容的平行情感录音即可实现零样本情绪风格迁移的工作之一。所提出的框架能够在保留说话人身份和语言内容的同时实现情感风格迁移，并泛化到未见说话人和未见情感组合。

1.5 论文其余部分的路线图

本论文各章节的组织结构如图所示：1.3。论文其余部分的结构如下。

第2 章介绍所使用的数据集以及应用于所研究不同问题的一些基线系统，为实验研究提供背景。

第 3 章介绍了一种情感感知预训练框架，该框架结合了声学和语义监督来学习语音表示。实验表明，所提出的预训练模型在多个语音情感识别基准上显著优于规模大得多的模型。本章还提出了一种用于基于文本的情感识别的基于大语言模型的监督预训练策略。

第 4 章聚焦于在对话和自然条件下进行口语情感识别的问题。针对对话情感识别（ERC），本文探讨了一种层次化建模范式。此外，本文提出了一种模块化专家混合框架，将源自语音大语言模型和文本大语言模型的声学与语义表征融合，用于 ERC。随后，我们讨论了为 *Interspeech 2025* 自然条件下语音情感识别挑战赛提出的系统。

第5章探讨仅语音、非平行设置下的情绪风格迁移问题。本章首先研究语音中语义、说话人和情感因素的解耦，随后提出一种情绪风格迁移框架。大量客观评估和主观评估表明，所提方法行之有效。此外，该方法还被用作数据增强策略，以提升情感识别性能。

最后，第6 章总结了本论文的主要贡献，并讨论了其局限性。本章最后概述了未来可能的研究方向，包括将情感识别和预训练方法拓展至临床相关任务，例如自动抑郁检测。

第2章

背景研究

本章介绍与本论文实验研究相关的背景知识和前期工作。我们首先回顾情感识别中常用的语音和文本表征，涵盖从手工描述符到任务驱动和自监督特征提取器的多种方法。随后，我们综述语音和文本模型的情感感知预训练相关工作，并着重指出其在捕捉情感线索方面的局限性。接下来，我们回顾对话情感识别（ERC）的方法，尤其关注整合声学与语义信息的方法。现有研究主要聚焦于受控或表演式数据集，而我们也讨论了代表性工作如何借助多模态线索推进情感识别。在情感合成方面，我们首先回顾通过文本转语音（TTS）和语音转换（VC）实现表现性语音生成的相关研究，这些研究为本论文探讨的问题奠定了重要基础。随后，我们重点介绍情感风格迁移（EST）的前期工作。该任务旨在保留语言内容和说话人身份的同时迁移情感表现力。最后，我们介绍本论文不同实验设置中使用的数据集。

2.1 情感建模的语音和文本特征

这里讨论用于情感识别的语音和文本表征。

2.1.1 语音情感识别特征

多年来，人们探索了多种用于语音情感识别的特征，从手工描述符到大规模预训练模型。这些特征大致可分为以下几类：

- **手工制作的 声学特征**：众所周知，语音信号的各种声学特征 n

与情感表达相关 [2]。基于这一观察，时域特征（如语速）、频域特征（如基频（F_0））以及振幅相关特征（如强度）已被广泛用于语音情感识别（SER） [32, 33]。随着时间的推移，学界对若干特征集进行了标准化，其中最著名的是通过Interspeech副语言挑战赛实现的标准化 [34, 35]。然而，这些特征集通常维度较高，因此推动了更紧凑的表征形式的发展，例如eGeMAPS特征集，该特征集由88个经过精心选择的参数组成 [36]。这些手工设计的特征集可通过openSMILE工具包获取 [27] ，并已广泛用于各种设置下的语音情感识别 [37, 38, 39, 40]。

- **可学习的特征提取器**：由于不同的语音任务通常需要特定于任务的表征，已有多项研究探索直接从原始语音波形中学习特征提取器。早期工作以可学习组件取代固定的梅尔滤波器组表征 [41]。随后，多项研究进一步提出直接从原始音频中学习卷积滤波器 [42, 43, 44, 45]。研究表明，这类可学习表征在包括语音情感识别在内的多种语音处理任务中，表现优于人工设计的特征。

- **自监督模型**：尽管人工设计和可学习特征提升了受控条件下的性能，但在多样化的真实场景中，它们往往不足以满足语音情感识别的需求。因此，人们开始通过自监督学习开发通用语音表示。具有代表性的方法包括对比预测编码（CPC） [46], PASE [47, 48], wav2vec [9, 49], HuBERT [10], 和 WavLM [11]。这些模型能够学习丰富的表征，捕捉语音的声学和语言属性。此类模型的预训练版本和微调变体持续显著提升了语音情感识别的性能，并已广泛应用于近期研究中 [50, 51, 52, 53]。

2.1.2 文本情感识别的特征

与语音情感识别（SER）类似，从文本中进行情感识别的任务也受到了研究界的广泛关注。用于此任务的特征大致可分为以下几类：

- **基于词典的特征**：英语领域已开发出大量情感词典，包括 WordNet [54], WordNet-Affect [55], 和 SenticNet [56], 等。

这些资源将词语或概念与预定义的情绪类别关联起来。基于这些关联，通常使用词频-逆文档频率（TF-IDF）向量来表示文本，并将其用于情绪分类。此前的多项研究表明，基于词典的方法取得了令人满意的效果 [57, 58,59]。

- **自监督模型**：Mikolov 等人提出了最早的分布式文本表征学习方法之一，即 word2vec 模型。与 GloVe 等类似的嵌入方法一样，这些模型从大型文本语料库中学习固定维度的词表示，并已广泛用于文本情绪识别。虽然这些嵌入优于基于词典的方法，但它们存在一个重要局限：每个词都由一个静态向量表示，且不受其上下文用法的影响。

为了解决这一局限，研究者通过 ELMo、GPT 和 BERT 等模型引入了上下文化词表示，这些模型均基于 Transformer 架构。在大型文本语料库上训练后，这些模型能够提供依赖上下文的表征，并在包括文本情感识别在内的广泛下游任务中表现出良好的泛化能力。多项研究探讨了如何利用这些模型进行特征提取和微调，以提升性能。关于基于 Transformer 的文本情绪识别方法的全面讨论，请参阅近期的综述研究。

2.2 情感感知预训练的相关工作

本节回顾了以提升情感理解为目标，对语音和文本模型进行预训练的相关工作。与通用的自监督学习范式不同，这些方法通过监督、蒸馏或任务特定损失函数，引入与情感相关的归纳偏置。

2.2.1 情感语音模型预训练

- **Chen 等人** [24] 提出了一种情绪感知语音表征学习框架，基于从预训练 WavLM-large [11]模型进行知识蒸馏。蒸馏损失仅应用于音高和能量较高或较低的帧，假设这些区域包含更多情绪内容信息。这种有针对性的蒸馏策略得到了一种紧凑的学生模型，在三个语音情感识别基准上均取得了出色的表现。

- **Ma 等人** [25] 同样采用了知识蒸馏范式，并使用 data2vec [72]作为教师模型。他们的方法同时包含话语级损失和帧级损失：前者对齐池化后的学生表征与教师表征，后者则在帧的掩码子集上最小化学生表征与教师表征之间的均方误差（MSE）。最终模型在多个语音情感识别基准上展现出竞争力，包括多语言环境下的基准测试。

尽管这些方法效果显著，但它们主要依赖从大型教师模型中隐式迁移情感线索，而未明确建模声学与语义因素在情感理解中的互补作用。

2.2.2 情感文本模型预训练

文本模型通常采用掩码语言建模（MLM）目标进行预训练 [23], ，模型需据此预测句子中被掩码的词语。虽然此类目标能够生成通用文本表征，但鲜有研究探索旨在明确提升文本模型情感理解能力的预训练策略。本小节回顾了这一方向的既有研究。

- **Tian 等人** [73] 提出了一种情感感知预训练框架，该框架会对一小组已知具有情感倾向的种子词赋予更高的掩码概率。除掩码语言建模损失外，预训练期间还引入了极性预测和方面-情感对预测等辅助目标。尽管该研究主要关注情感分析，而非细粒度情感识别，但结果表明，所提出的策略能够提升 RoBERTa 等预训练模型的性能 [74]。

- **Sosea 等人** [75] 提出了一种类似的方法，在预训练期间以更高的概率掩码情感和情绪词典中的词语。研究表明，这种称为情绪掩码语言建模（eMLM）的策略能够提升下游情绪分类任务的表现。

与这些由词典驱动的掩码策略不同，本文探讨利用大语言模型从口语中获取监督信号进行情绪感知文本预训练，从而能够在经过整理的文本语料库之外实现可扩展的情绪标注。

2.3 多模态情感识别的相关工作

现在我们来讨论 与口语多模态情绪识别相关的既有研究 .

2.3.1 对话中的情感识别

在本小节中，我们回顾对话情感识别（ERC）方面的既有研究。该研究通过纳入对话上下文和说话人交互，将话语级情感识别加以扩展。

- **Poria 等人** [13] 提出了 ERC 问题最早期的解决方案之一，采用双向长短期记忆网络 [76]框架进行上下文建模和多模态融合。尽管单模态特征提取器并未与上下文模型联合训练，但该研究表明，将上下文建模与多模态融合解耦，可能比训练单一统一网络更有效。

- **Yun 等人** [77] 提出了一种知识蒸馏框架，用于将非文本模态与文本表征对齐。所提方法利用语音和视觉模态，并明确纳入说话人信息，以应对对话数据中说话人数量和交互模式的变化。

- **Li 等人** [78] 通过联合预测情感类别和情绪变化，对连续话语间的情绪转变动态进行建模。明确对对话轮次中的情绪演变进行建模后，在多个对话情绪识别基准数据集上均取得了更好的表现。

尽管这些研究显著推动了 ERC 的上下文建模和多模态建模，但现有方法大多在端到端架构中，将表征学习、上下文建模和多模态融合紧密耦合。因此，声学与语义线索对情感识别各自的贡献往往难以解释或控制。相比之下，本论文从模块化层次视角研究 ERC， [79], 明确将话语级表征学习与对话上下文建模和多模态融合分离开来。通过这种方式，我们能够更系统地探究声学与语义信息如何促进对话中的情感理解，以及如何通过混合专家模型等灵活的融合机制有效结合这些线索。

2.3.2 语音情感识别：多模态视角

语音情感识别已经使用广泛的声学特征和建模技术进行了深入研究。早期研究主要关注单独的声学线索，而近期进展则凸显了纳入从口语中提取的语义信息的重要性

。在本小节中，我们回顾以语音情感识别为主题的既有研究，重点关注整合声学和语义线索的方法，因为这一视角与本论文所 प्रस्तुत的贡献最为相关。大多数此类方法依赖于人工转录文本，而语音情感识别数据集通常会提供这些文本。在缺少此类转录文本的场景下，可以使用自动语音识别（ASR）系统将语音信号转换为文本，从而实现这些多模态技术的应用。

- **Lee 等人** [80] 纳入了语义信息，具体是通过引入情感显著性这一针对语料库内词语的概念。单个词语与情感类别之间的互信息被用于量化词语与特定情绪的相关性。与早期工作类似，该任务仅限于将情绪分类为正面和负面类别。

- **Jin 等人** [81] 通过引入特定情感的词语共现统计，扩展了词袋表示，从而得到一种情感向量表征。这些语义特征与声学信息相结合后，在语音情感识别基准上取得了更好的性能。

- **Sun 等人** [82] 提出了一种近期的基于 Transformer 的多模态框架，分别使用 BERT [23] 和 wav2vec 2.0 [49], 处理文本与语音。为应对各模态之间信息内容的差异，作者引入了一种解耦表征学习策略，以分离共享特征与模态特定特征。所提出的方法在基准数据集上展现出更优的语音情感识别性能。

尽管上述方法展示了整合声学与语义信息的优势，但它们主要是在受控或表演语音数据集上进行评估的，在这些数据集中，录音条件、说话风格和情感表达都经过精心安排。因此，在自然条件下（其特点包括自发语音、背景噪声、说话人多样性以及情感分布不均衡）的表现和鲁棒性仍基本未得到探索。相比之下，本论文明确聚焦于自然场景下的语音情感识别，并探讨当这些假设不再成立时，如何有效利用声学与语义线索。

2.4 富表现力语音合成的先前研究

近年来，合成语音中的情感与表现力控制受到了广泛关注，这一趋势由对话代理、娱乐等应用推动，

以及人机交互。本节回顾了富表现力语音合成的相关研究，重点关注文本转语音、语音转换和情感风格迁移。

2.4.1 文本转语音

在富表现力文本到语音（TTS）领域，已有大量研究，其目标是在合成语音时控制说话人身份、韵律和情感表现力等属性。本小节回顾了从参考语音信号中提取说话风格的代表性研究。

- **Min 等人** [83] 提出了早期的基于参考的富表现力文本到语音框架，该框架从参考语音信号中推断说话风格。作者引入了风格自适应层归一化（SALN），其中，文本嵌入由从参考语音中提取的学习到的风格向量所预测的增益和偏置参数进行调制。

- **Huang 等人** [84] 指出，使用单个全局向量来表征富表现力风格过于受限。为此，他们提出了一种多层次风格建模方法，融合了音素级、帧级和话语级风格表示。研究表明，该系统能够有效地泛化到未见过的参考说话人。

- **Li 等人** [85] 采用潜在扩散模型 [86] 从参考语音中捕捉富表现力的风格信息。学习到的潜在风格表示实现了高质量的富表现力语音合成，在感知评估中取得了可与人类语音媲美的表现。

2.4.2 语音转换

表现力文本转语音系统依赖文本转录进行合成，而越来越多的研究则聚焦于在语音到语音设置下直接修改语音信号。语音转换（VC）的任务是将源语音信号中的说话人身份转换为参考说话人的身份，同时保留语言内容，且无需文本监督。这一问题的扩展形式称为富表现力语音转换（EVC），还旨在迁移参考语音的情感说话风格。本小节回顾了该领域具有代表性的研究。

- **Rizos 等人** [87] 提出了最早采用 StarGAN [88]框架进行说话人和情感联合转换的方法之一。该模型能够同时转换语音中的说话人身份和情感风格。作者进一步展示了

转换后样本作为语音情感识别数据增强策略的实用价值。

• **Maimon 等人** [89] 采用预训练的自监督语音模型，将语音分解为内容和风格表征。他们使用可学习的嵌入来捕捉说话人特定风格信息，并将其用于情感语音转换。该研究还指出，语速是实现有效情感风格迁移（VC）的关键因素。

• **Yao 等人** [90] 采用了基于参考语音的风格编码器，与 Min 等人提出的方法类似， [83]，用于在语音转换框架中建模说话风格。他们使用风格自适应层归一化（SALN）对转换过程进行条件化处理，与先前方法相比，提升了自然度和风格相似度。

2.4.3 风格迁移

情感风格迁移（EST）是指在保留源语音的语言内容和说话人身份的同时，将参考语音信号中的情感说话风格迁移到源语音信号中的任务。与富表现力语音转换（EVC）不同，EST 要求将情感风格与说话人身份明确解耦，因此问题更具挑战性。尽管已有少数研究在平行设置下探讨了这一任务[1]，本文提出了首批针对非平行纯语音设置下情感风格迁移问题的解决方案之一 [22]。本小节将回顾后续探索相关问题表述的研究。

• **Zhang 等人** [91]提出了 VEVO，这是一种统一框架，利用参考语音修改内容表征，同时根据参考说话人特征对声学表征进行条件化。由此得到的模型支持多种富表现力语音任务，包括情感风格迁移。尽管 VEVO 使用带转录文本的语音进行训练，但在推理时不需要文本输入。

• **Yao 等人** [92] 利用 NaturalSpeech 3 [93] 中的风格编码器从参考语音中提取风格表征。说话人嵌入由同一说话人的多段话语生成，并与风格表征解耦，从而能够在不同说话人之间迁移情感说话风格。然而，所提出的方法

[1]当训练期间可以获得来自同一说话人、内容相同但情感表达不同的语音时。

并未在训练期间源说话人或语言内容未见的条件下进行评估。

- **Yoon 等人** [94]提出了一个框架，能够对转换语音中的内容、音高、能量、节奏和说话人身份进行细粒度控制。情感风格迁移是该模型的一个自然应用，该模型采用韵律提取器，将情感特征从参考迁移到源语音。尽管引入了内容分类器以增强解耦，评估仍局限于训练期间可观察到源语音内容的场景，这一局限亦见于 [92]。

综合来看，关于表现力文本转语音、语音转换和情感风格迁移的既有工作凸显了可控语音合成方面的稳步进展。然而，现有方法大多要么依赖文本监督，要么假定存在平行情感数据，要么将情感风格与说话人身份或内容纠缠在一起。相比之下，本论文聚焦于仅语音、非平行设置下的情感风格迁移，在该设置中既没有对齐的情感话语，也没有可用的文本转写，并且必须对情感、内容和说话人因素进行显式解耦。

2.5 数据集

本论文使用的数据集涵盖多项任务，包括情感感知预训练、情感识别和情感风格迁移。下文将介绍每个数据集及其在论文中的主要作用，并在适用时注明其其他用途。不同数据集的摘要也列于表 2.1中。

- **MSP-PODCAST** [95] 是一个大规模情感语音语料库，由约 409 小时的播客录音组成。数据被标注为 10 类情感——愤怒、快乐、悲伤、中性、厌恶、恐惧、轻蔑、惊讶、其他和无——此外还标注了连续的效价、唤醒度和支配度分数。得益于其规模、自然录音条件和丰富的情感标注，该数据集既用于情感感知预训练（第3.1.2.1、3.2.2.1节），也用于评估自然情境下的语音情感识别（第 4.2.1.5节）。

- **IEMOCAP** [96] 是一个多模态数据集，包含 151 段录制的对话，分为五个场次，每个场次均包含一名男性说话人与一名女性说话人之间的双人互动。录音被切分为话语，共有 10,039 条话语，由人工评估者标注为 10 种情感类别之一：愤怒、快乐、悲伤、中性、沮丧、兴奋、恐惧、惊讶、厌恶，以及

“其他”。按照标准评估流程，我们考虑两种分类设置。第一种采用四分类设置，包含愤怒、快乐、悲伤和中性类别，其中快乐和兴奋类别合并，共计 5,531 条话语。第二种设置采用六分类任务，与先前研究相同 [97], ，使用前六种情感类别，共计 7,433 条话语。在两种设置中，Session 5 均留作测试集。在本文中，IEMOCAP 被用作基准数据集，用于评估情感感知预训练表示（第3.1.2.2、 3.2.2.2节）以及开展对话情感识别（ERC）实验（第 4.1.2.4、 4.1.3.3节）。

• **MELD** [98] 是一个多方对话情感识别数据集，源自电视剧 老友记。该数据集包含 1,433 段对话，共计 13,708 条话语。每条话语均被标注为七种情感类别之一：愤怒、喜悦、悲伤、恐惧、厌恶、惊讶和中性。该数据集划分为 1,153 段对话（11,098 条话语）用于训练和验证，以及 280 段对话（2,610 条话语）用于测试。在本论文中，MELD 被用作基准数据集，用于评估情感感知预训练表征（第 3.1.2.2、 3.2.2.2 节），并用于对话情感识别（ERC）实验；该数据集以更复杂的多方互动场景补充了 IEMOCAP 提供的双人对话场景（第 4.1.2.4、 4.1.3.3 节）。

• **CMU-MOSI** [99] 是一个情感分析数据集，由 93 段口语独白组成，并划分为 2，199 条话语，每条话语均标注了一个连续情感分数，其取值范围为 $[-3, 3]$。按照先前研究的做法，该任务被定义为一个二分类问题：情感分数处于范围 $[-3, 0)$ 内的话语标记为负面，而处于范围 $[0, 3]$ 内的话语则标记为正面。我们采用与 Lian等人 相同的数据集划分方案 [97], ，使用 49 段独白进行训练、13 段进行验证，其余 31 段用于测试，最终得到 1，188 条训练语句、325 条验证语句和 686 条测试语句。在本论文中，CMU-MOSI 用于评估情感感知预训练表示的泛化能力（第3.1.2.2、3.2.2.2节），以及独白场景下的多模态情感识别模型，并与 IEMOCAP 和 MELD 等对话基准测试形成互补（第4.1.2.4、4.1.3.3节）。

• **DAIC-WOZ** [100]是一个用于自动抑郁检测的基准数据集，包含参与者与访谈者之间进行的 189 次临床访谈。其中，107 次访谈被指定为训练集，35 次被指定为开发集。该数据集存在显著的类别不平衡，训练集中仅有 30 次访谈被标注为“抑郁”。Fol-

lowingprior work [17], ，从五个划分中的每次访谈随机抽取话语，并在开发集上报告性能。在本论文中，DAIC-WOZ 用于评估情感感知预训练语音表征在临床相关下游任务中的有效性，并考察其在标准情感识别基准测试之外的泛化能力（第 3.1.2.2节）。

• **RAVDESS-Song** [101] 是一个情感歌唱数据集，包含由 23 位不同歌手演唱的 1，012 段录音。每段录音都标注了六种情感类别之一：中性、平静、快乐、悲伤、愤怒和恐惧。我们通过创建 5 个不同的数据划分进行说话人无关评估。每个划分中，16 位歌手的录音用于训练，3 位歌手的录音用于验证，其余 4 位歌手的录音则留作评估。在本文中，RAVDESS-Song 用于评估情感感知预训练语音表征在歌声场景下的泛化能力，为口语研究提供了一个互补的情感评估场景（第 3.1.2.2节）。

• **CaFE** [102]是一个加拿大法语情感语音数据集，包含由 12 位说话人说出的 936 段话语。每段话语都标注了七种情感类别之一：中性、愤怒、厌恶、悲伤、惊讶、恐惧和快乐。与 RAVDESS-Song 类似，我们采用 5 个不同的数据划分进行说话人无关评估。每个划分中，8 位说话人的话语用于训练，其余 4 位说话人则平均分配到验证集和测试集。不同划分中的说话人分配采用随机方式。在本文中，CaFE 用于评估情感感知预训练语音表征的跨语言泛化能力，为英语情感识别数据集提供了一个非英语基准（第3.1.2.2节）。

• **Emo-DB** [103] 是一个德语情感语音数据集，包含 10 位说话人说出的 535 条话语。每条话语都标注了七种情感类别之一：中性、愤怒、厌恶、悲伤、无聊、恐惧和喜悦。与 CaFE 和 RAVDESS- Song 类似，我们通过 5 种不同的数据划分进行说话人无关评估。对于每种划分，使用 6 位说话人的话语进行训练，其余 4 位说话人则平均分配到验证集和测试集。不同划分中的说话人分配均为随机。在本文中，Emo-DB 用于评估情感感知预训练语音表征的跨语言泛化能力，以补充英语和法语情感识别基准测试（见 3.1.2.2）。

• **MSP-IMPROV** [104] 是一个视听情感数据集，包含 8,438 条话语

由 12名演员演绎，每条话语均标注了连续的效价、唤醒度和支配度（VAD）分数，取值范围为 $[1, 5]$。为确保评估不受说话人影响，我们采用留出两名说话人的协议，将数据集划分为 12 折。每一折中，使用 10 名说话人的话语进行训练，其余两名说话人的话语用于验证和测试集。最终性能取所有折的平均值。在本论文中，使用 MSP-IMPROV 评估情感感知预训练语音表征在连续情感预测任务中的有效性，具体包括效价、唤醒度和支配度估计（第 3.1.2.6节）。

- **ESD** [105] 是一个情感语音数据集，包含由 10 位英语母语说话人和 10 位汉语母语说话人说出的语句，涵盖五种情感类别：中性、快乐、愤怒、悲伤和惊讶。在本论文中，我们仅使用该数据集的英语子集，并遵循预定义的训练-验证-测试划分，每位说话人每种情感选取 300 条话语用于训练。因此，训练语句总计 15，000 条，另有 2，500 条未见话语（每位说话人每种情感 50 条）用于验证。本论文使用 ESD 训练情感风格迁移系统；但训练过程中并未利用该数据集的平行特性（即相同内容以不同情感说出）（第 5.2.1 节）。平行语句仅用于评估，以便在保持说话人身份和语言内容不变的情况下，对情感迁移质量进行受控评估（第 5.2.3 节）。

- **CREMA-D** [106] 是一个表演式情感语音数据集，包含演员说出 12 个固定句子的 7，442 段视听录音，涵盖六种情感类别：愤怒、悲伤、快乐、恐惧、厌恶和中性。本论文仅使用 CREMA-D 评估情感风格迁移系统（第 5.2.3 节）。

- **TIMIT** [107] 是一个音素平衡语音语料库，包含约 5 小时的语音，来自 630 位说话人，每位说话人都会朗读固定的 10 个句子。该数据集最初是为自动语音识别而设计的，不包含富表现力或情感语音。在本论文中，TIMIT 被用作评估情感风格迁移系统的中性、非表现性语音来源（第 5.2.3节）。

2.6 章节总结

本章全面介绍了与本论文所探讨的三个核心研究方向相关的既有工作背景：情感感知表征学习、多模态

表 2.1： Sum本论文所用数据集及其在不同章节中的作用概述。

数据集	Task	# 样本数	章节
MSP-PODCAST [95]	情感感知语音预训练 自然语音情感识别	~409 小时	3.1.2.1, 3.2.2.1 4.2.1.5
IEMOCAP [96]	预训练表示评估 对话情感识别（ERC）	10,039 条话语	3.1.2.2, 3.2.2.2 4.1.2.4, 4.1.3.3
MELD [98]	预训练表示评估 对话情感识别（ERC）	13,708 条话语	3.1.2.2, 3.2.2.2 4.1.2.4, 4.1.3.3
CMU-MOSI [99]	预训练表示评估 对话情感识别（ERC）	2,199 条语音	3.1.2.2, 3.2.2.2 4.1.2.4, 4.1.3.3
DAIC-WOZ [100]	抑郁症检测（表示评估）	189 次访谈	3.1.2.2
RAVDESS-Song [101]	预训练语音表征评估	1,012 段录音	3.1.2.2
CaFE [102]	预训练语音表征评估	936 条话语。	3.1.2.2
Emo-DB [103]	预训练语音表征评估	535 条语音	3.1.2.2
MSP-IMPROV [104]	效价–唤醒度–支配度预测	8,438 条语音	3.1.2.6
ESD [105]	情感风格迁移（训练；非平行用途） 情感风格迁移的评估	15,000 条语音	5.2.1 5.2.3
CREMA-D [106]	情感风格迁移评估	7,442 段录音	5.2.3
TIMIT [107]	情感风格迁移的评估	5 小时	5.2.3

口语情感识别，以及通过情感风格迁移实现的富表现力语音合成。

我们首先回顾了情感识别中语音与文本特征的发展历程，涵盖手工设计的声学描述符和基于词典的文本特征，以及大规模自监督模型和基于 Transformer 的模型。尽管这些表征推动了性能稳步提升，但它们在很大程度上依赖通用的预训练目标，而这些目标并未与口语中的情感线索明确对齐。

接下来，我们梳理了语音和文本模型中情感感知预训练的现有方法。以往工作主要通过知识蒸馏或修改掩码策略来引入情感归纳偏置。然而，这些方法往往孤立地关注声学或语义方面，因此需要采用统一的预训练策略，对这两个因素进行联合建模，以实现情感理解。

随后，我们回顾了多模态情感识别领域的研究，重点关注对话情感识别（ERC）。现有方法表明，上下文建模和多模态融合十分重要，但通常难以平衡模态主导性、可扩展性和鲁棒性。此外，大多数多模态情感识别基准都基于表演性或受控语音，这限制了其在真实、自然场景中的适用性

。

最后，我们讨论了富表现力语音合成方面的已有工作，涵盖文本转语音、语音转换和情感风格迁移。尽管基于参考的表现力合成已取得显著进展，但将情感风格与说话人身份及语言内容进行显式解耦仍然具有挑战——尤其是在非平行、仅语音的设定中。

综合来看，这一背景凸显了若干尚未解决的挑战：（i）缺乏能够联合利用声学与语义线索的情感感知预训练目标；（ii）需要能够处理对话结构与自然语音的稳健多模态情感识别框架；（iii）在没有平行监督的情况下实现语音中可控情感迁移的困难。

这些研究空白直接推动了本文后续章节所提出的方法和贡献。

第3章

面向口语的情感感知预训练

自监督预训练已成为现代语音与语言处理的基石，使模型能够从大量无标注数据中学习通用表征。在语音领域，wav2vec2.0 [49], HuBERT [10], 和 WavLM [11]在众多下游任务中展现出卓越的迁移能力，包括自动语音识别、说话人识别和语音情感识别。尽管取得了这些进展，预训练所采用的目标主要用于捕捉语言结构或短期声学规律，而非语音的情感特征。

然而，口语中的情感感知依赖于声学实现与语义语境之间微妙的相互作用。音高、能量和语速等韵律变化与语言内容相互作用，传达情感意图；这些线索在不同说话人、录音条件和数据集中往往呈现出不同的形式。因此，采用针对语义重建优化的目标进行预训练的模型，可能无法充分编码与情感相关的变化，导致其迁移到情感识别任务时泛化能力有限。

本章探讨旨在弥合这一差距的口语情感感知预训练策略。我们研究如何在预训练过程中联合利用声学与语义因素，以生成更适用于下游情感识别任务的表征。具体而言，我们探索侧重于语音信号中与情感相关区域的预训练目标，并分析这些目标对多个语音情感识别基准测试中泛化能力的影响。

除了语音表示之外，本章还探讨从口语中提取的文本模型的情感感知预训练

。我们不再仅仅依赖包含隐含情绪内容的大型文本语料库，而是采用一种语音驱动监督范式：将口语话语转录成文本，并使用大语言模型推断情绪标签后为其添加标注。随后，我们利用这些带情绪标签的文本语料库对基于文本的表示进行预训练，从而即使在带标签的文本数据稀缺的情况下，也能实现情绪感知语义建模。

3.1 语音情感识别的预训练

我们介绍一种用于语音情感识别（SER）的自监督模型，名为**C**内容与**A**声学**R**表征**E**情感（CARE） [108]。据我们所知，我们的方法是首个预训练自监督模型、并将语音的语义和声学组件整合起来的尝试。CARE采用双重编码框架处理语音信号：语义编码器将语音表示与句子级转录文本对齐，非语义编码器则将语音表示与PASE的低级声学特征对齐+ 模型 [48]。随后，将两个编码器的输出合并，并训练一个轻量级分类头来执行情感识别。

3.1.1 所提方法

3.1.1.1 背景

RoBERTa：Devlin 等人提出了一种文本表示模型，这是该领域的一项重要贡献。 [23] Liu 等人 [74] 在更大规模的文本数据语料库上训练了这一架构，且未采用下一句预测任务。这个预训练模型称为稳健优化的 BERT 方法（RoBERTa），研究表明，它在多项下游任务中的表现优于 BERT。

3.1.1.2 CARE 模型

我们提出一种双重编码方案（语义编码器和声学编码器），通过适合各自目标的不同监督信号来处理语音信号。各编码器所采用的监督方式详述如下：

语义监督：我们不假设预训练数据中存在真实文本转录。在这种情况下，预训练的自动语音识别（ASR）系统（Whisper-large-v3 [109]）可作为生成这些转录文本的替代方案。通常，与中性语音数据集相比，自动语音识别系统在情感语音上的词错误率（WER）更高 [110]。另一方面，播客录音提供了足够长的上下文和丰富多样的内容，适合用于语义

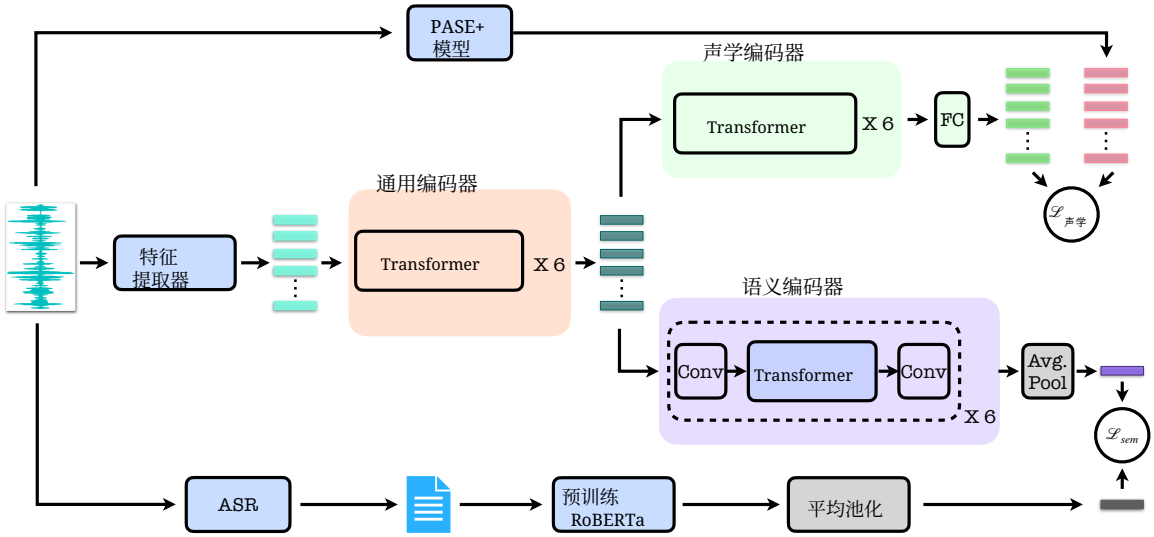


图 3.1：所提出的 CARE 模型框图。该模型的声学编码器以 PASE+ 特征作为目标进行训练。蓝色块表示冻结的组件或不含可学习参数的组件。对于语义编码器，Transformer 层被冻结，而卷积适配器则参与训练。由于声学编码器输出的维度为 768，因此附加了一个全连接层，以匹配维度为 256 的 PASE+ 特征。推理时不使用该全连接层以及语义编码器后的平均池化模块。

编码器。具体而言，我们观察到词错误率为 12。53%，对于语音情感识别任务而言，这或许是合理的。由于语义编码器的目的是使语音信号与其内容对齐，从而促进情感识别，因此可以采用自动语音识别式对齐损失。然而，正如 Fan 等人所确立的那样，句子级文本表示更适用于情感识别任务。 [111]因此，我们使用预训练 RoBERTa 模型从转录文本中提取上下文词级嵌入 [74] ，并对这些嵌入进行均值池化，以获得一个表示整段转录文本的特征向量。这些话语级嵌入作为我们 CARE 模型中语义编码器的监督信号，或称“教师”。我们将这些话语级嵌入记为 $\boldsymbol{y}_{text}$。

声学监督： 在先前的研究中，均值池化后的表征已被证明能够编码说话人身份、口音和语言等特征 [112]。然而，我们推测，语音情感通常包含在音高、节奏及其调制等细粒度声学属性中 [33]。因此，我们为声学编码器选择了帧级目标。

针对帧级声学目标的直接方法是对语音信号的部分进行掩码并加以重建。然而，先前的研究表明，随机掩码在情感识别上不如选择性掩码高能量或高基频区域有效，正如 Chen 等人 [24] 所证明的那样。基于这些观察，我们选择预测 PASE+

特征，这些特征涵盖滤波器组能量、音高以及其他对捕捉情感至关重要的低层描述符。具体而言，我们使用 256 维的帧级 PASE+ 特征作为 CARE 模型中声学编码器的目标。这些特征以因子 2 进行下采样，从而以 50 Hz 的频率生成目标描述符。我们将 PASE+ 模型的声学目标记为 $\boldsymbol{y}_{pase}$。

模型架构：语音信号首先经过一系列卷积层处理，这些卷积层旨在每20毫秒生成帧表示。随后是六层Transformer层的堆叠，构成通用编码器，在所提出的模型中同时服务于声学编码器与语义编码器两条通路。

语义编码器旨在将语音表示与其对应的生成的转录文本对齐。该编码器由六层Transformer层组成，并使用预训练文本表示模型的权重进行初始化。由于这些层是用文本数据训练的，语义编码器中的Transformer层无法泛化到语音表示。为解决这一问题，我们提出一种新颖的适配策略，即引入两个一维卷积块——分别置于每个Transformer 层之前和之后。

第一个块将来自通用编码器的语音表示调整为与基于文本的模型所期望的内部表示对齐。第二个块在Transformer处理后对这些表征进行细化。此外，遵循在文本模型中处理语音的既有做法 [113, 114, 115], 在由语义编码器中的Transformer层处理之前，语音序列的时间分辨率会被降低。具体而言，每个Transformer 层之前的卷积块将序列长度下采样为原来的三分之一，而其之后的块则以相同倍数进行上采样。每个卷积块由单个卷积层组成，卷积核大小为5，输入通道和输出通道均设为768，以匹配预训练Transformer层的维度。尽管先前工作已探索过使用卷积层对语音自监督学习模型进行适配 [116, 117], 但使用卷积适配器将预训练文本模型适配于语音任务，在本工作中尚属首次探索。重要的是，Transformer层本身在训练过程中不会被更新。最后，对语义编码器的输出表示进行平均池化，以生成话语级表示。

声学编码器同样由六个 Transformer 层组成，随后使用全连接层将其输出映射到 256 维，以匹配 PASE+ 特征目标。图 3.1 展示了 CARE 模型的框图。

损失：一种语义 语义损失 L_{sem} 和帧级声学损失 L_{acoust}分别用于训练 g

语义编码器和声学编码器。将语义监督记为 $\boldsymbol{y}_{text}$，将语义编码器的输出记为 $\hat{\boldsymbol{y}}_{sem}$，语义损失为均方误差（MSE）损失：

$$L_{sem.} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{y}_{text}^{i} - \hat{\boldsymbol{y}}_{sem}^{i}||_2^2 \tag{3.1}$$

其中 N 表示批量大小。

对于帧级损失，令 $\hat{\boldsymbol{y}}_{acoust} \in \mathbb{R}^{N\times T\times D}$ 表示声学编码器的输出，其中 N、T 和 D 分别表示批量大小、每个话语的帧数以及表征的维度。损失定义如下：

$$L_{acoust.} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} ||\boldsymbol{y}_{pase}^{ij} - \hat{\boldsymbol{y}}_{acoust}^{ij}||_2^2 \tag{3.2}$$

其中$\boldsymbol{y}_{pase}$ 表示声学目标。预训练期间的总损失为

$$L_{tot.} = L_{sem.} + \lambda L_{acoust.} \tag{3.3}$$

其中 λ 根据验证性能确定。

推理： 为了评估模型在各种下游任务中的表现，我们采用 SUPERB 基准提出的范式 [118]。声学编码器中每个 Transformer 层的输出与语义编码器中该 Transformer 层之后卷积块的输出进行拼接。随后，将这些输出与通用编码器和卷积特征提取器的逐层输出结合起来。此过程共得到 13 个层表征：一个来自卷积特征提取器，六个来自通用编码器，另外六个来自拼接后的语义编码器和声学编码器。然后，将这些层表征的凸组合输入分类头。需要注意的是，在推理过程中，不使用声学编码器中的全连接层和语义编码器中的平均池化块。

在此设置中，下游任务唯一的可学习参数是凸组合的权重和轻量级分类头的权重。

表 3.1：评估数据集概述。最后一列表示训练/测试数据是否包含相同的说话人。

数据集	# 话语数	# 训练集	# 验证集	# 测试集	# 类别	评估	说话人独立性
IEMOCAP-4 [96]	5531	3860	430	1241	4	预定义	✓
IEMOCAP-6 [96]	7433	5230	580	1623	6	预定义	✓
MELD [98]	13706	9988	1108	2610	7	预定义	✗
CMU-MOSI [99]	2199	1188	325	686	2	预定义	✓
DAIC-WOZ [100]	10197	6003	2097	2097	2	预定义	✓
RAVDESS-Song [101]	1012	–	–	–	6	5 折交叉验证	✓
CaFE [102]	936	–	–	–	7	5 折交叉验证	✓
EmoDB [103]	535	–	–	–	7	5 折交叉验证	✓

3.1.2 实验与结果

3.1.2.1 预训练

MSP-PODCAST 语料库 [95] 用于预训练任务。共使用了 149，307 个样本，总计 230 小时的情感语音数据。其中，随机选取 80% 的数据作为训练集，其余 20% 作为验证集。使用 Whisper-large-v3模型生成转录文本（观察到的词错误率为 12。53%），并使用预训练 RoBERTa 模型对转录文本进行编码。通用编码器使用 WavLM-base模型的前 6 层初始化，声学编码器则使用同一模型的后 6 层初始化。卷积特征提取器也使用 WavLM-base模型初始化。语义编码器的 6 个 Transformer 层使用预训练 RoBERTa 基础模型最后 6 层的权重初始化，而卷积适配器则采用随机初始化。

3.1.2.2 下游任务

表格中总结了用于评估的不同数据集。3.1有关不同数据集的详细说明，请参阅第 2.5节。

3.1.2.3 损失与评估指标

交叉熵损失用于训练下游模型权重（凸组合权重和轻量级分类头参数）。

在测试中，我们使用加权 F1 分数作为评估指标，因为许多数据集存在类别不平衡（表 3.1）。将类别 c （含 N_c 个样本）的 F1 分数记为 $F1_c$，

加权 F1 分数为

$$WF1 = \frac{1}{\sum_{c=1}^{C} N_c} \sum_{c=1}^{C} N_c \times F1_c \tag{3.4}$$

我们还报告所有情况下的非加权平均召回率（UAR），即各类别召回率分数的平均值。

3.1.2.4 实现细节

预训练： 在预训练期间，MSP-PODCAST 中的所有语音话语都会填充或随机裁剪至 5 秒。模型的学习率设为$1e$-5，批量大小为 128，并使用 AdamW [119]作为优化器。模型总共训练 200,000 步，并根据验证集上的表现选择最佳模型参数，用于评估下游数据集。我们尝试了不同的 λ（公式 3.3）取值，以平衡预训练期间的两项损失。设置为 $\lambda = 0.1$会导致性能下降，而将其增大至 $\lambda = 10$ 则不会比$\lambda = 1$带来显著提升。因此，在后续所有实验中，我们固定 $\lambda = 1$ 。

微调与评估： 对于下游任务训练，语音信号会被裁剪至最长 30 秒，或填充至最短 1 秒。对于抑郁检测数据集 DAIC-WOZ，每个语音片段的时长为 10 秒 [17]。

通用、语义和声学编码器中每一层的输出维度均为$T \times 768$，其中 T 表示以 50 Hz 采样的语音信号中的帧数。对于 CARE 模型，由于拼接了语义编码器和声学编码器各 6 层的输出，组合后的输出维度为 $6 \times T \times 1536$。为与此维度保持一致，卷积特征提取器的输出以及通用编码器 6 层的输出均进行复制，从而得到维度为 $7 \times T \times 1536$的特征。这 13 层的表征通过凸组合方式进行组合，并使用可学习权重生成维度为$T \times 1536$的特征。随后，沿时间维度对特征进行均值池化，为每个音频文件生成一个 1536 维向量。该向量输入由两层前馈神经网络构成的分类头，该网络采用 ReLU 激活 [120]。与 SUPERB 框架一致，仅在各下游数据集上训练用于层表征凸组合的权重，以及两层全连接分类头中的权重 [118]。

我们使用的批量大小为32，学习率为 $1e$-4，并将模型训练50个训练轮次。两层分类头的隐藏维度设置为256。AdamW 优化器

表 3.2：与其他下游数据集相关研究的比较。# 预训练中包含下游数据集的模型。结果以**粗体**和下划线标示，分别表示最佳和次佳模型。所有数值均为基于 5 次随机初始化计算的加权 F1 分数（显示均值和标准差）。括号中还显示了非加权平均召回率。

数据集	WavLM [11] 参数:94M	HuBERT [10] 参数:94M	data2vec [72] 参数:94M	emotion2vec [25] 参数:94M	SONAR [121] 参数:600M	SALMONN [115] 参数:70亿参数	SALMONN [115] 参数:13B	CARE 参数:160M
IEMOCAP-4	65.9$^{\pm 0.5}$(67.2)	65.0$^{\pm 0.2}$(68。0)	62.7$^{\pm 0.7}$(64.0)	67.5$^{\pm 0.6\#}$(69。0)	59.4$^{\pm 0.4}$(61。0)	**75.8**$^{\pm 0.6\#}$(76.9)	72.9$^{\pm 2.3\#}$(74.9)	69.4$^{\pm 0.5}$(70.1)
IEMOCAP-6	51.7$^{\pm 0.5}$(48.4)	50.7$^{\pm 0.9}$(46.5)	46.0$^{\pm 0.4}$(42.3)	54.1$^{\pm 0.6\#}$(51.9)	43.5$^{\pm 0.2}$(41。0)	**59.3**$^{\pm 1.4\#}$(55.7)	58.1$^{\pm 1.6\#}$(55.2)	55.0$^{\pm 0.4}$(52.1)
MELD	45.6$^{\pm 0.4}$(24.3)	45.3$^{\pm 0.6}$(24.0)	41.9$^{\pm 0.5}$(23。1)	47.6$^{\pm 0.3\#}$(27.4)	43.2$^{\pm 0.2}$(23。3)	**53.3**$^{\pm 0.7}$(33.4)	52.6$^{\pm 0.4}$(32.8)	48.1$^{\pm 0.8}$(28.8)
CMU-MOSI	64.1$^{\pm 0.8}$(64。2)	62.5$^{\pm 0.6}$(62。5)	59.7$^{\pm 0.4}$(58。9)	66.5$^{\pm 0.6}$(65。9)	74.6$^{\pm 0.3}$(73。9)	**78.0**$^{\pm 0.7}$(77.0)	72.8$^{\pm 1.0}$(72.0)	66.7$^{\pm 1.0}$(66.2)
DAIC-WOZ	63.2$^{\pm 1.5}$(61.5)	65.9$^{\pm 2.0}$(61.9)	67.8$^{\pm 1.4}$(65。7)	61.6$^{\pm 0.7}$(61.0)	64.3$^{\pm 0.4}$(63.7)	62.6$^{\pm 3.4}$(60.4)	64.7$^{\pm 3.0}$(61。1)	**68.5**$^{\pm 2.1}$(67。1)
RAVDESS	50.5$^{\pm 3.6}$(49。1)	53.5$^{\pm 1.1}$(55.7)	38.5$^{\pm 5.2}$(40.8)	48.5$^{\pm 1.0}$(51。0)	11.8$^{\pm 2.0}$(10。8)	50.2$^{\pm 1.3}$(54。2)	51.9$^{\pm 3.6}$(53.4)	**60.1**$^{\pm 1.6}$(62.0)
CaFE	66.6$^{\pm 2.6}$(69。0)	66.5$^{\pm 4.5}$(69.1)	48.8$^{\pm 4.3}$(51。0)	59.3$^{\pm 3.8}$(62.6)	5.7$^{\pm 1.4}$(7.1)	59.9$^{\pm 2.0}$(62。8)	69.8$^{\pm 3.3}$(71。4)	**77.0**$^{\pm 1.5}$(78。1)
EmoDB	66.5$^{\pm 4.8}$(68.5)	66.9$^{\pm 3.9}$(68.2)	48.9$^{\pm 3.1}$(49.9)	64.4$^{\pm 2.6}$(66.7)	10.2$^{\pm 2.0}$(12。6)	82.8$^{\pm 2.9}$(85.3)	82.2$^{\pm 4.0}$(84。1)	**83.4**$^{\pm 2.0}$(83。9)
Avg.	59。3(56。5)	59。5(57。0)	51.8(49.5)	58。7(56。9)	39。1(36。7)	65。2(63。2)	65.6(63.1)	**66.0**(63.5)

在此同样使用。所有模型（包括CARE与基线系统）均使用相同的分类后端。因此，该设计可对不同表征进行公平比较[1]。

3.1.2.5 CARE 的性能

使用所提出的 CARE 模型的表征，在 8 个下游数据集上的结果如表 3.2 所示。这些基线模型根据推理时使用的参数数量分为两组：基础模型（参数规模 < 200M）和大型模型（> 500 M），其中也包括基于大语言模型的模型。针对每个类别得出以下观察结果：

基础模型： 我们将 HuBERT [10], WavLM [11], data2vec [72] 和 emotion2vec [25]表征作为此类别的基线模型。在这些基线系统中，emotion2vec 还在 IEMOCAP 和 MELD 数据集上进行了预训练，这在一定程度上解释了它在这些数据集的下游任务中取得更好结果的原因。虽然 CARE 在 CMU-MOSI 上的表现与 emotion2vec 相近，但在其他数据集上，其表现优于所有基础模型。在平均意义上，所提出的 CARE 相较于表现最佳的基线模型（HuBERT）实现了 15.6% 的相对提升。

大型模型： SONAR [121]被选为该类别的语音编码器。对于六个英语数据集，使用预训练英语语音编码器[2]，而法语和德语语音编码器分别用于CaFE和EmoDB数据集。类似

[1]代码可在此处获取：https://github.com/iiscleap/CARE。

[2]https://dl.fbaipublicfiles.com/SONAR/spenc.eng.pt

于CARE后端，SONAR编码器的层表征被线性组合，并在下游任务上训练分类头。尽管SONAR的参数规模几乎是CARE的四倍，我们提出的模型在除CMU-MOSI之外的所有数据集上均优于SONAR。

基于大语言模型的模型：SALMONN 的两个版本 [115]（70亿参数和 13B）[1] 被作为基于大语言模型的模型示例。这类模型通常应用于零样本设置；然而，由于不同数据集中的情感类别各不相同，其零样本性能并不稳定。例如，SALMONN-13B 模型在 IEMOCAP-4 数据集（该模型在此数据集上训练）上取得了 68。75% 的加权 F1 分数，但在 MELD 上仅为 24。06%。因此，为了公平比较，基于大语言模型的评估也采用与 CARE 和其他基线模型相同的框架。通过凸组合聚合所有层的内部表示（SALMONN 13B 有 41 层，SALMONN 7B 有 33 层），并针对每个下游数据集训练分类头（与 CARE 类似）。与 emotion2vec 类似，SALMONN 的预训练数据包含 IEMOCAP，因此其在 IEMOCAP-4 和 IEMOCAP-6 上的表现优于 CARE。较大的模型规模和丰富的预训练数据使 SALMONN 在 MELD 和 CMU-MOSI 数据集上的表现分别比 CARE 高出 10% 和 34%（相对提升）。然而，在其余四项任务中，CARE 均胜过 SALMONN 模型，在 RAVDESS-Song 和 CaFE 数据集上分别取得了 17% 和 24% 的相对提升。值得注意的是，尽管 SALMONN 的预训练使用了音乐数据集，CARE 仍成为 RAVDESS-Song 数据集上的最佳模型。

要点： 1）总体而言，CARE 在八个数据集上的平均表现最佳，甚至超过了参数量近乎其 80 倍的 SALMONN 13B 模型。尽管基于大语言模型的模型在域内情感识别数据集上表现出优势，但它们在域外任务上的表现会下降，这表明其在多种任务和多语种情感语音中的泛化能力有限。2）CARE 优于 WavLM、HuBERT 和 data2vec 等语音自监督学习模型，这一结果并不意外，因为这些模型是在非情感数据上训练的（相关实验见第 3.1.3.7 节）。3）值得注意的是，尽管 CARE 仅使用英语语音进行训练，但它在 CaFE 和 EmoDB 数据集上的表现优于多语种 SONAR 模型。这表明我们的预训练技术在语音情感识别的域外任务中具有泛化能力。

3.1.2.6 情感属性预测

情感识别可以表述为一个回归问题，即预测特定话语的效价、唤醒度和支配度 [122]。我们使用 MSP-IMPROV [104] 来

[1] https://huggingface.co/tsinghua-ee/SALMONN

表 3.3：MSP-IMPROV 数据集的结果。CCC 表示一致性相关系数，V、A、D 分别表示效价、唤醒度和支配度。

方法	CCC-V	CCC-A	CCC-D
WavLM-base [11]	0.51	0.64	0.47
emotion2vec [25]	0.5	0.61	0.49
SALMONN-7B [115]	0.53	**0.67**	<u>0.52</u>
SALMONN-13B [115]	<u>0.56</u>	0.65	0.51
CARE	**0.57**	<u>0.66</u>	**0.53**

实现这一目的。关于该数据集的详细信息，请参阅第2.5 节。

我们使用一致性相关系数（CCC）作为指标。将真实值的均值和方差分别记为 μ_g、 σ_g^2，将预测分数的均值和方差分别记为 μ_p、 σ_p^2，则一致性相关系数定义为

$$CCC = \frac{2\rho\sigma_g\sigma_p}{\sigma_p^2 + \sigma_g^2 + (\mu_g - \mu_p)^2} \tag{3.5}$$

在式3.5中， ρ 是真实值与预测分数之间的皮尔逊相关系数。在训练下游模型时，CARE及其他模型的表征采用与分类数据集类似的方式进行聚合。随后连接一个两层回归头，其隐藏维度为256，输出维度为3（三个属性各占1维）。目标是提高效价、唤醒度和支配度每个维度上真实值与预测值之间的一致性相关系数。该数据集以及其他基线模型的结果如表3.3所示。我们注意到，对于该任务，CARE嵌入在效价和支配度属性上取得了最佳结果，而在唤醒度方面，SALMONN-7B模型的表现略优。

3.1.3 讨论

3.1.3.1 与基线系统的比较

考虑了四种基线系统（表3.4）：-

PASE+： 对于每个下游数据集，提取 PASE+ 特征，并训练一个分类网络，以预测每个话语的情绪类别，与 CARE 的做法类似。推理时使用的参数总数为 8M。

Whisper： 对于每个下游数据集，Whisper-large-v3模型的 33 个编码器层所产生的表征 [109] 通过可学习权重进行线性组合。随后，在这些表征之上训练一个两层分类头，用于情感识别任务。

表3.4：不同下游数据集上的基线结果（以加权 F1 分数衡量）。所有数值均为下游网络经过 5 次随机初始化后的平均值。本表还展示了 CARE 各组成部分的结果。

数据集	PASE+ [48] **参数**:8M	Whisper [109] **参数**:800M	Whisper [109]+ RoBERTa [74] **参数**:1.6B	教师-融合 **参数**:1.6B	语义 + 通用编码器 **参数**:110M	声学 + 通用编码器 **参数**:94M	CARE **参数**:160M
IEMOCAP-4 [96]	56.68	56.40	61.97	69.49	66.44	65.91	**69.39**
IEMOCAP-6 [96]	41.38	40.62	49.28	56.61	53.05	52.09	**55.02**
MELD [98]	35.86	40.11	**49.29**	49.72	47.37	46.98	48.05
CMU-MOSI [99]	50.69	55.60	**75.14**	74.12	64.23	64.17	66.74
DAIC-WOZ [100]	66.84	62.08	64.04	67.63	66.32	66.89	**68.49**
RAVDESS-Song [101]	46.05	34.20	9.58	48.48	55.23	56.17	**60.11**
CaFE [102]	52.86	19.22	13.59	53.42	69.23	71.62	**76.98**
EmoDB [103]	62.59	24.77	14.98	66.75	75.42	78.63	**83.41**
Avg.	51.62	41.63	42.23	60.78	62.16	62.81	**66.02**

推理期间使用的参数总数为 800M。

Whisper+RoBERTa： 转录文本由 Whisper-large-v3模型生成，随后由预训练 RoBERTa模型处理。RoBERTa 的内部表示通过可学习权重进行线性组合，随后训练一个两层分类头。推理期间共使用 1.6B 个参数。

教师融合： PASE+ 和 Whisper+RoBERTa 表征进行拼接，并针对每个下游数据集训练一个两层分类头。此基线在推理时也共有 1.6B 个参数。

要点：1）在 8 个数据集上的平均表现方面，CARE 比声学监督信号高出 29。76%（相对提升）。这一改进可归因于 CARE 的参数规模大于 PASE+ 模型。2）CARE 的表现优于 Whisper 和 Whisper+RoBERTa 系统，相对提升分别为 41。79% 和 41。18%。这表明，尽管基于 Whisper 的基线系统规模大得多，CARE 对声学与语义信息的结合仍能有效实现情感识别。3）在 MELD 和 CMU-MOSI 上，CARE 的表现不如 Whisper+RoBERTa 基线。在这些数据集上，众所周知，基于文本的模型显著优于仅使用语音的系统 [123, 97]。在 Whisper+RoBERTa 设置中，RoBERTa 模型是在由 Whisper-large-v3（规模为 1。6B 的模型）生成的转录文本上进行微调的。相比之下，CARE 是一个更小的模型（160M），且在推理时不直接使用自动语音识别转录文本。为进一步说明模型规模方面的公平性，我们将用于自动语音识别的 Whisper-large-v3 替换为 Whisper-base，再进行 RoBERTa 建模。此时，MELD 上的表现从 49。29% 降至 46。02%，CMU-MOSI 上则从 75。14% 降至 71。91%。这凸显了准确转录文本的重要性

在文本信息情感含量丰富的设置中，准确的转录文本和强大的模型容量至关重要。4）尽管教师融合基线在若干涉及英语语音的数据集上表现出色，但 CARE 的平均表现仍比该基线高出 5。24 个百分点。这也说明了 CARE 使用知识蒸馏进行预训练的原因：仅使用 10% 的参数，便超越了融合基线。

3.1.3.2 两个编码器的重要性

表 3.4 展示了在下游数据集上仅使用声学编码器或语义编码器之一，并结合通用编码器时 CARE 的性能。为了评估语义编码器与通用编码器的组合，我们使用卷积特征提取器、通用编码器和语义编码器输出的 768 维表征，不包括声学编码器的输出。同样，在评估声学-通用编码器组合时，不纳入语义编码器的表征。值得注意的是，虽然 CARE 的参数量（160 M）高于 WavLM 或 emotion2vec 等模型，但这两种组合在推理时的参数量相近。语义-通用编码器组合在推理时的参数量为 110M，而在每个下游数据集上进行评估时，声学-通用编码器的参数总量为 94M。

要点： 1）在 8 个数据集上的平均表现中，声学编码器与通用编码器表征的组合比表现最佳的自监督学习模型（HuBERT）高出 8.08%（相对提升）（见表 3.2）。鉴于参数量相近，这一表现说明我们的预训练方法具有优势。2）在三个域外数据集上，声学-通用编码器组合优于语义-通用编码器组合。3）在所有数据集上，CARE 中两个编码器的组合表现最佳，这表明尽管两个编码器各自的性能相当，但它们能够捕捉语音信号的不同特征。

3.1.3.3 语义编码器的修改

为了评估我们为语义编码器所做设计选择的适用性，我们对架构进行了三项修改：

1. **CARE-No init.:** 移除卷积适配器后，语义编码器中的 Transformer 层采用随机初始化（而非使用预训练 RoBERTa 权重）。
2. **CARE-Trans.:** 移除卷积适配器，同时更新 RoBERTa Transformer 层。

表 3.5：修改语义编码器后在下游数据集上的结果（加权 F1 分数）。

数据集	方法	WF1
IEMOCAP（四分类）	CARE-No init.	65.71
	CARE-Trans.	65.89
	CARE-FT	68.16
	CARE	**69.39**
IEMOCAP(六分类)	CARE-No init.	50.96
	CARE-Trans.	51.19
	CARE-FT	52.37
	CARE	**55.02**
MELD	CARE-No init.	45.16
	CARE-Trans.	46.57
	CARE-FT	47.93
	CARE	**48.05**
CMU-MOSI	CARE-No init.	62.16
	CARE-Trans.	65.97
	CARE-FT	64.27
	CARE	**66.74**
DAIC-WOZ	CARE-No init.	64.57
	CARE-Trans.	65.93
	CARE-FT	67.43
	CARE	**68.49**

3. **CARE-FT：** 保留卷积适配器，同时更新语义编码器中的所有参数（卷积适配器和 Transformer 权重）。

这些修改的结果见表3.5。

要点： 1）使用 RoBERTa 初始化 Transformer 权重对 CARE 的性能至关重要。随机初始化语义编码器会导致其在全部五个数据集上的性能下降，这表明随机初始化的语义编码器难以回归到语义监督信号。2）移除卷积适配器（在 CARE-Trans 中）会对性能产生负面影响，凸显了我们采用基于卷积的适配技术来对齐语音表示与 RoBERTa 的 Transformer 层的必要性。3）更新语义编码器中的 Transformer 层会降低性能。由于 RoBERTa 是在文本上预训练的，使用语音数据进行微调会削弱其效果。

3.1.3.4 声学编码器和通用编码器的初始化

如第 3.1.1 节所述，CARE 的通用编码器和声学编码器均使用 WavLM-base 模型的权重进行初始化。我们展示了将该初始化方式改为 i) 随机初始化、ii) HuBERT-base [10] 或 iii) data2vec-base [72] 时方法的结果（表 3.6）。**主要结论：** 1) 使用 data2vec 初始化会降低模型性能，这可能是因为 data2vec 的基线性能低于 HuBERT 和 WavLM（见表

表 3.6：声学编码器采用不同初始化时，在下游数据集上的结果（加权 F1 分数）。

数据集	随机初始化	HuBERT 初始化	data2vec 初始化	WavLM 初始化
IEMOCAP-4	66.72	67.65	66.76	**69.39**
IEMOCAP-6	51.47	51.40	52.98	**55.02**
MELD	46.41	46.97	47.19	**48.05**
CMU-MOSI	65.07	66.26	**68.14**	66.74
DAIC-WOZ	67.56	65.19	66.59	**68.49**
RAVDESS-Song	57.81	**60.30**	55.09	60.11
CaFE	73.83	72.23	62.46	**76.98**
EmoDB	78.61	**86.51**	77.19	83.41
Avg.	63.44	64.56	62.05	66.02

，而不是 HuBERT 和 WavLM（见表3.2）。例外是 CMU-MOSI 数据集，采用这种初始化的模型相较于使用 WavLM 初始化的模型，性能提高了 4。21%（相对值）。2) 使用 HuBERT 初始化的模型在 RAVDESS-Song 和 Emo-DB 数据集上表现最佳。值得注意的是，对于这两个域外数据集，HuBERT 的表现优于 WavLM（表 3.2）。3) 与语义编码器相比，初始化对声学编码器和通用编码器的影响较小，因为语义编码器需要与文本表征对齐。

3.1.3.5 声学目标的选择

我们进行了一项实验，使用基于 eGeMAPS 的声学编码器训练目标 [36]特征，这些特征由 openSMILE 工具包提取 [27]。CARE 模型中声学编码器的 PASE+目标被 eGeMAPS 特征取代。该模型称为 CARE（eGeMAPS），其性能如图所示。3.2。

关键结论：使用 eGeMAPS 输入特征的基线模型表现不如使用 PASE+ 特征的基线模型，这正如预期，因为 eGeMAPS 是手工设计的。因此，使用 eGeMAPS 的 CARE 平均表现也低于使用 PASE+目标的 CARE。

3.1.3.6 语义目标的选择

我们开展了一项实验，将 Whisper 编码器表征用作语义编码器的监督信号。我们探索了以下几种变体：1）在预训练期间，对 Whisper 表征进行池化，并将其用作语义目标。该模型称为 CARE（Whisper-pool）。2）使用 Whisper 的帧级表示作为目标进行预训练。该模型称为 CARE（Whisper-frame）。3）我们还使用语音与 RoBERTa 词元之间的帧级对齐

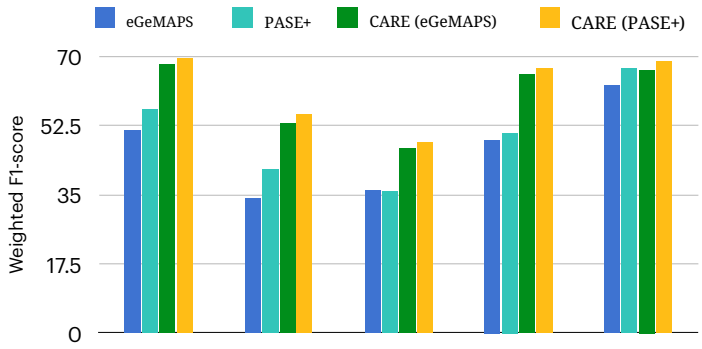


图 3.2：使用不同声学目标时 CARE 的性能。以 eGeMAPS 作为特征的模型采用与 PASE+ 基线相同的方式进行训练。所有数值均为 5 次随机初始化的平均值。

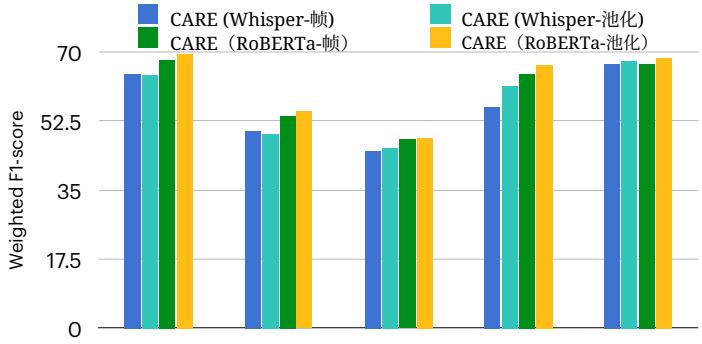


图 3.3：使用不同语义目标时 CARE 的性能。所有数值均为 5 次随机初始化的平均值。

，并将帧级RoBERTa表示用作语义目标。该模型称为 CARE（RoBERTa-frame）。不同变体与所提出的模型 CARE（RoBERTa-pool）的性能对比如图 3.3 中所示。

要点： 1）CARE（RoBERTa-池化）的性能优于使用 Whisper编码表示训练的两种变体。2）使用池化目标训练语义编码器时，系统的性能优于

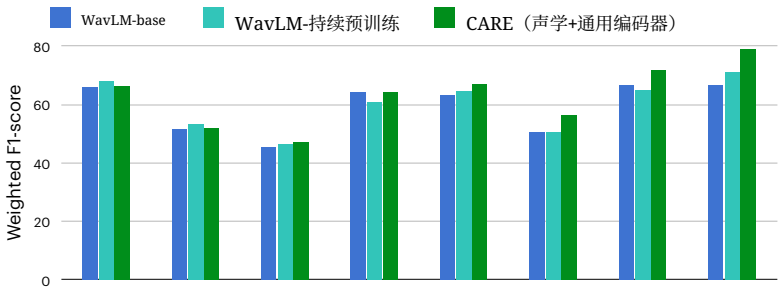


图3.4：WavLM 在 MSP-PODCAST 上持续预训练时的性能比较。图中还展示了 CARE 的声学编码器与通用编码器组合的性能作为参考。此处，RAVDESS 指 RAVDESS-Song数据集。

使用帧级表示训练时的性能。

3.1.3.7 WavLM 的持续预训练

由于所有自监督学习（SSL）模型均在中性数据上训练，因此它们从语音信号中准确识别情感的能力通常有限。因此，在情感数据集上预训练这些自监督学习模型时，其情感识别性能至关重要。为探究预训练设置对所提出的 CARE 的影响，我们使用 MSP-PODCAST 数据集，对公开可用的 WavLM-base模型继续进行预训练。该过程遵循 WavLM 的预训练流程，采用掩码语言建模损失，并额外进行 200，000 步训练（与 CARE 类似）。该实验结果见图 3.4，其中，持续预训练后的 WavLM 模型性能标记为 WavLM-Cont. Pretrain。图中还展示了通用编码器与声学编码器组合的性能，以供比较。

关键要点：持续预训练提升了 WavLM-base 在某些下游任务上的表现，例如 IEMOCAP。然而，除 IEMOCAP 外，WavLM 的表现不如 CARE 的声学-通用编码器组合。由于图 3.4中的所有模型在推理时具有相同的规模（94M），该实验凸显了我们所提出的基于蒸馏的预训练的优势。

3.1.4 语音情感识别预训练总结

图 3.5总结了各种语音情感识别方法的全貌。在当前的建模框架中：模型要么以有限的性能换取效率（位于 x 轴较低端的模型），要么专注于最大化性能，但需要更多的内存和计算资源（通常基于大语言模型）。我们提出的 CARE 系统结合了小型模型的计算效率与大规模系统的高性能，从而实现

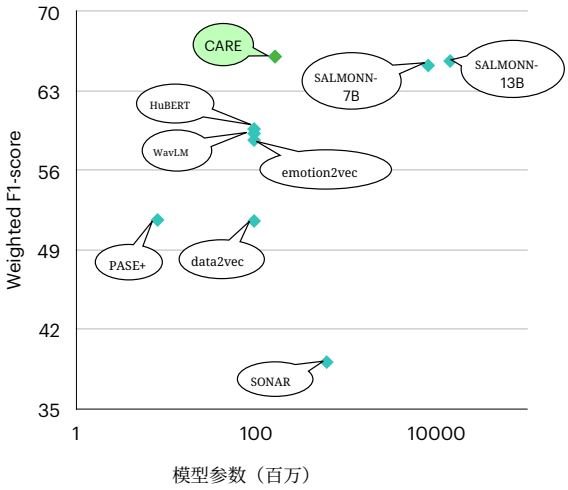


图 3.5：推理模型规模（参数量，单位为百万）与语音情感识别（SER）表现（8 个数据集上的平均加权 F1 分数）的散点图。与现有方案相比，CARE 实现了更优的权衡。更多详情请参阅表 3.2 和3.4 以及相关讨论。

.

效率与性能之间更优的权衡。

3.2 文本情绪识别的预训练

我们提出了一种多模态预训练方法，同时利用文本和语音表示。具体而言，我们引入一种策略，借助大规模语言模型（LLM）利用无监督语音数据，提升文本情绪分类效果。为此，我们首先使用基于 Whisper-large 模型的预训练自动语音识别（ASR） [109] 转录语音。“带噪”的语音转录文本由大语言模型标注，以自动生成语音情感的伪标签。随后，我们使用这些已标注的文本转录稿微调 RoBERTa 文本编码器模型 [74] 以进行情感分类，使其能够捕捉文本数据中细腻的情绪模式。

3.2.1 所提方法

首先从自动语音识别系统（Whisper-large-v3 [109]）中提取情感语音语料库的文本转录。通常，情感语音的词错误率高于中性语音 [110]。利用该语音语料库的自动语音识别生成的转录文本，提示大语言模型（LLM）将转录文本标注为三种类别：“正面”、“负面”或“中性”。对预训练 RoBERTa-large 模型进行微调，以预测伪类别

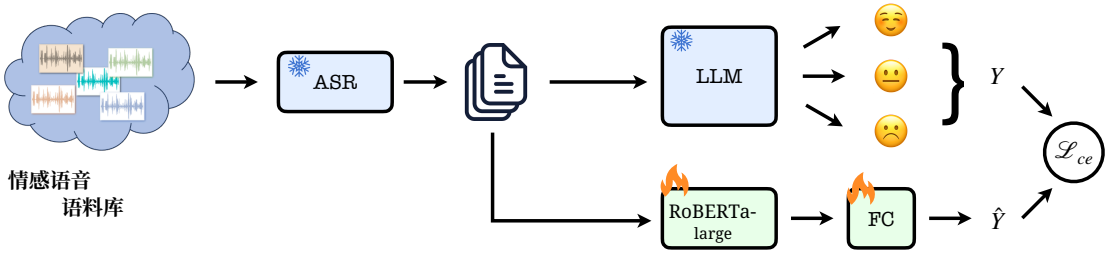


图3.6：使用自动语音识别系统生成预训练数据的转录文本，再由大语言模型（LLM）将其标注为正面、负面或中性情感。这些“银标”标签与文本转录共同构成了 RoBERTa-large 模型的监督训练数据集。

（来自大语言模型的预测结果）。所提出的预训练策略的框图如图 3.6 所示

3.2.2 实验与结果

3.2.2.1 预训练

MSP-PODCAST 语料库 [95] 用于预训练任务。共计使用149,307个语音轮次，合计230小时的情感语音数据。在全部样本中，随机选取80%的数据作为训练集，其余20%作为验证集。Whisper-large-v3模型[1] 用于生成转录文本。这些转录文本使用 `GPT-3.5 Turbo`[2] 模型进行标注。

3.2.2.2 下游任务

三个数据集用于评估我们预训练策略的性能——IEMO-CAP [96]，MELD [98] 和 CMU-MOSI [99]。这些数据集的详细信息见第2.5节。

3.2.2.3 实现细节

一旦Whisper模型生成转录文本，便按如下方式对 `GPT-3.5 Turbo` 模型进行提示：

> 你是一个情感分类机器人。给定 [句子]，，请将其分类为正面、负面或中性情感。请仅输出情感，不要添加其他文本。

[1] https://huggingface.co/openai/whisper-large-v3[2]
https://platform.openai.com/docs/models/gpt-3-5-turbo

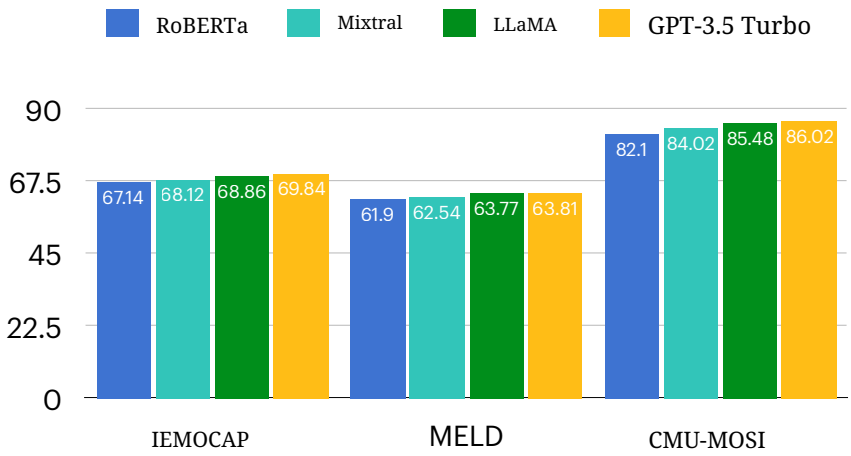


图 3.7：RoBERTa-large 模型在不同数据集上的表现。使用不同的大语言模型，根据语音转录文本生成伪情绪标签。文中也报告了未经任何监督微调的预训练 RoBERTa 的表现。

RoBERTa-large 模型使用大语言模型生成的标签（3 个类别）进行预训练，共训练 10 轮次，学习率为 $1e$-4，批量大小为 32。

3.2.2.4 结果

我们使用三种大语言模型进行下游评估。不同大语言模型标注能力的影响如图3.7所示。图中也列出了预训练 RoBERTa 模型（未接受任何大语言模型指导）的表现，以供参考。我们注意到，使用 `GPT-3.5 Turbo` 提供的标签微调后的 RoBERTa 模型，相较于预训练 RoBERTa 模型，在 IEMOCAP、MELD 和 CMU-MOSI 上分别取得了 8.22%、5.01% 和 21.9% 的相对提升。与 `Mixtral-8x7B-Instruct-v0.1` 相比， `GPT-3.5 Turbo`分别在 IEMOCAP、MELD 和 CMU-MOSI 上实现了 5.4%、3.39% 和 12.51% 的相对提升。CMU-MOSI 数据集上最高的性能提升，可能归因于该数据集采用了二分类任务。

3.2.3 文本模型预训练总结

在本节中，我们提出了一种非常简单的方法，以提升基于文本的模型在情感识别方面的性能。通过利用大语言模型（LLM）的先验知识，我们为情感显著的口语内容标注情感标签。随后，通过有监督预训练，利用这些情感标签对文本模型进行预训练。结果显示，该方法在多个口语情绪识别基准上均能提升性能。

3.3 章节总结

本章探讨了面向口语的情感感知预训练策略。其动机在于，传统的自监督目标主要针对语言重建进行优化，可能无法充分捕捉情感理解所必需的情感线索。我们探讨了如何在预训练过程中显式融入声学与语义信息，以提升模型在不同情感识别任务上的泛化能力。

首先，我们介绍了 CARE，这是一种在预训练期间融合声学与语义监督的语音表征学习框架。通过广泛的实验和消融实验，我们证明 CARE 在计算效率与情感识别性能之间实现了良好平衡，并在多个基准测试中优于轻量级模型和大规模系统。这些结果凸显了在预训练阶段融入情感相关归纳偏置的重要性。

其次，我们提出了一种互补方法，用于对源自口语的文本模型进行情感感知预训练。通过利用大语言模型的先验知识为带转录文本的语音标注情感标签，我们实现了无需人工标注文本语料库即可对文本表征进行有监督预训练。这一策略在多个情感识别基准测试中持续提升了性能，证明了语音驱动监督在语义情感建模中的有效性。

本章介绍的方法共同为口语的声学和语义维度建立了情感感知预训练框架。

第4章

口语多模态情绪识别

本章聚焦于两种具有挑战性的设置下的口语多模态情绪识别：对话情感识别（ERC）以及自然录音条件下的情感识别。ERC 要求模型捕捉对话上下文，同时能够适应涉及多名说话者的互动。在自然条件下录制的语音情感识别还面临其他挑战，例如说话者和录音环境的多样性，以及高度偏斜的情感分布。

在ERC中，情绪不仅必须根据单个话语进行推断，还必须结合其对话上下文，包括说话人之间的交互以及话语之间的时间依赖关系。对这类依赖关系进行建模会带来与上下文聚合和多模态融合相关的挑战。本章研究用于捕捉对话结构的层次化方法，并分析融合声学与语义信息的不同机制。我们进一步提出一种模块化混合专家建模范式，以有效结合这些模态进行ERC。

然而，现有的对话情感识别数据集并未体现自然语音的复杂性。如第2.5节所述，这些数据集主要由在受控实验室环境中录制的表演式情感构成，因此难以推广到自然口语。本章后半部分将介绍我们为改进自然语音情感识别所做的工作，这些工作是我们参与*Interspeech 2025* 自然条件下语音情感识别挑战赛 [1]的一部分。我们评估了多种近期的基于大语言模型的架构，并探讨了针对以下情况设计的损失函数的使用

[1]https://lab-msp.com/MSP-PODCAST_竞赛/IS2025年/

cl类别不平衡，在挑战排行榜上取得了有竞争力的表现 .

4.1 对话情感识别（ER C)

我们首先对对话情感识别（ERC）问题进行形式化定义。随后，我们介绍了两种用于多模态对话情绪识别的建模框架：层次式交叉注意力模型（HCAM），以及模块化混合专家架构。

HCAM框架旨在显式建模对话上下文，同时整合声学与语义信息。该架构采用包含三个阶段的分层训练范式。第一阶段，以不考虑上下文的方式独立训练语音和文本的单模态预测器。第二阶段，通过上下文建模层引入对话上下文，使每种模态都能捕捉话语间的时间依赖关系。最后阶段，使用基于交叉注意力的协同注意力模块融合声学流与文本流，从而生成多模态预测。

尽管HCAM能够有效捕捉上下文和跨模态交互，但它存在特征级融合方法普遍面临的一项关键局限：当一种模态的表现显著优于另一种模态时，多模态融合并不能稳定地超越表现更强的单模态系统。在这种情况下，融合表征会受到较弱模态的主导，从而限制多模态建模的整体收益。

为解决这一局限，我们提出**情感**识别的**语**音-**文**本**专**家**混**合（（MiSTER-E），这是一种模块化架构，在决策层而非特征级进行融合。MiSTER-E采用混合专家（MoE）架构，包含三个独立优化的分支：纯语音上下文专家、纯文本上下文专家和多模态专家。

每个单模态专家都使用时间 Inception 网络和后续的循环层来建模模态特定的对话动态，而多模态专家则通过堆叠的交叉注意力和自注意力模块融合语音和文本表征。所有专家的输出都由门控网络进行整合，该网络在 Softmax 归一化之前计算专家 logits 的加权和。这种对数几率级融合能够实现动态专家选择，使模型根据不同线索的相对可靠性，着重利用单模态或多模态线索。

为进一步促进专家专业化并稳定训练，我们在专家之间加入对比损失和基于KL散度的一致性约束。此外，使用焦点损失监督每个专家，以减轻类别不平衡的影响。实证结果表明，在一种模态占主导地位的场景中，这种模态解耦的混合专家方案具有明显优势，

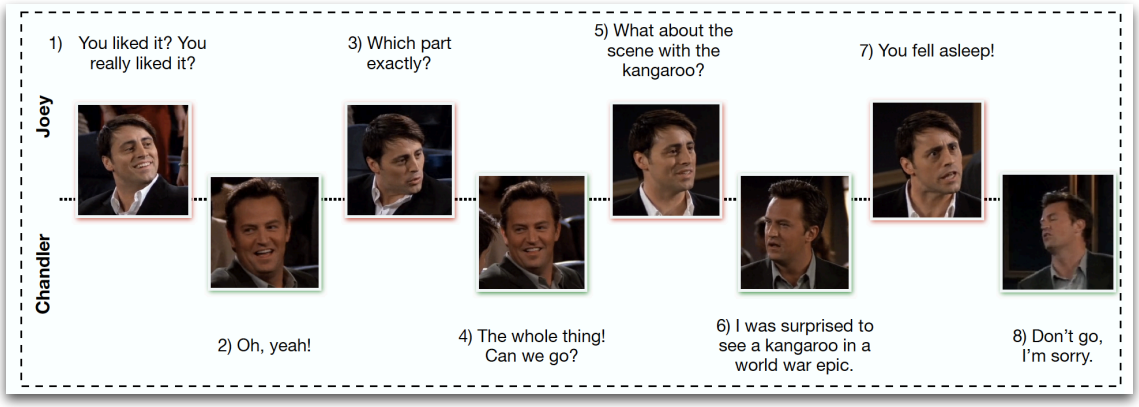


图4.1：情绪识别对话数据集 MELD 中的一段对话示例。该示例展示了多名说话者在对话中互动时，情绪如何变化

.

并始终优于特征级融合方法和单模态基线。

4.1.1 问题描述

设 $\mathcal{D}$ 表示一个包含 P 段对话的数据集 $\mathcal{C} = \{\mathrm{c}_1, \mathrm{c2}、\ldots\ldots、\mathrm{c}_P\}$。每段对话 c_i 由一系列话语组成：$\mathcal{U}_i = \{\mathrm{u}_{i1}、\mathrm{u}_{i2}、\ldots\ldots、\mathrm{u}_{iN_i}\}$，其中 N_i 表示对话 c_i中的话语数量。在本文中，我们考虑每个话语的两种模态：语音和文本。因此，每个话语表示为 $\mathrm{u}_{ik} = \{\mathrm{s}_{ik}、\mathrm{t}_{ik}\}$，其中 $k = 1、\ldots\ldots、N_i$；s_{ik} 表示语音信号，t_{ik} 表示相应的文本内容。每个话语 u_{ik} 都对应一个情绪标签 $y_{ik} \in \mathcal{Y}$，其中 $\mathcal{Y}$ 表示为数据集 $\mathcal{D}$定义的情感类别集合。对话情感识别（ERC）的任务是，对于每段对话 c_i，根据话语序列 $\{\mathrm{u}_{i1}、\mathrm{u}_{i2}、\ldots\ldots、\mathrm{u}_{iN_i}\}$ 预测情感标签序列 $\mathrm{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{iN_i}\}$。这一表述自然对应于序列到序列分类问题，其中话语之间的上下文依赖起着关键作用。

MELD 数据集中的一段对话示例如下 [98]，如图 4.1所示。该对话由两位说话人交替说出的八个话语组成，每个话语都标注了相应的情绪标签。此示例凸显了在 ERC 中同时对话上下文和说话人交互进行建模的重要性。该图改编自 Poria 等人 [14]。

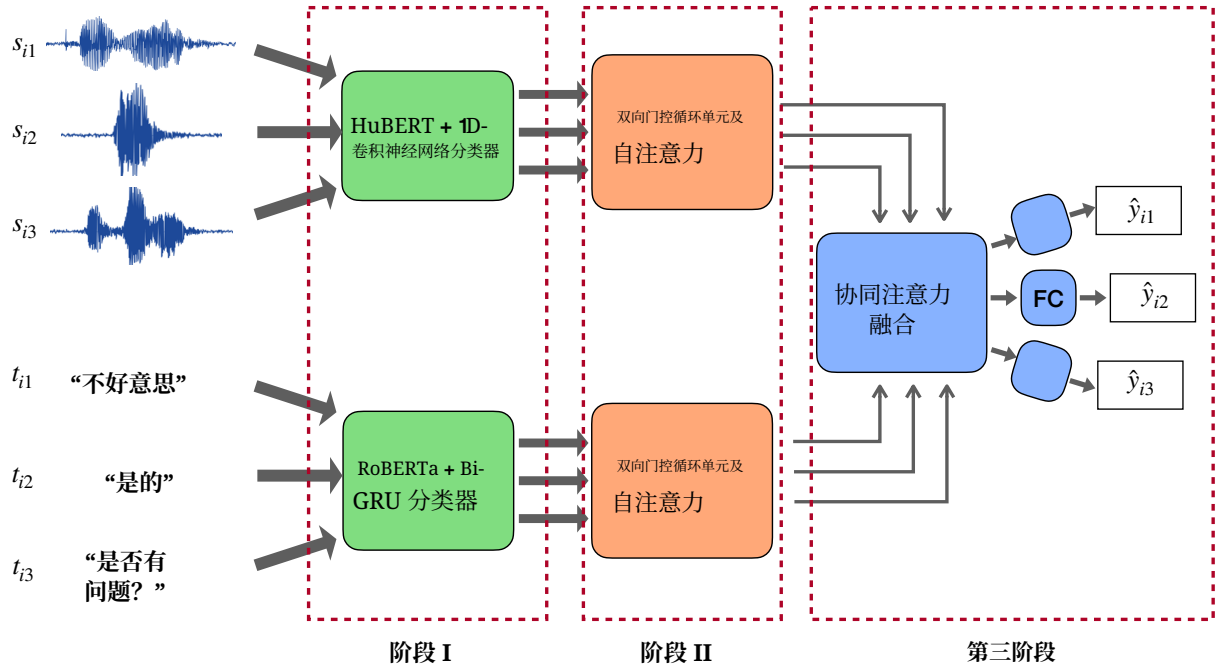


图4.2：所提出模型的框图。图中还标出了训练的三个阶段。

4.1.2 多模态情感识别对话的层次化建模

HCAM 的框图如图 4.2所示。我们的模型包含三个不同的阶段，这些阶段按层次方式进行训练；每个阶段都使用前一阶段的预训练模型参数，且不进行微调。

4.1.2.1 背景

自注意力与交叉注意力：Vaswani 等人提出的自注意力机制 [68],处理一个长度为 N、维度为 d_k 的序列。该序列被转换为三个矩阵，$Q \in \mathcal{R}^{N \times d_k}$、$K \in \mathcal{R}^{N \times d_k}$和 $V \in \mathcal{R}^{N \times d_k}$。自注意力用于计算查询表示（记为 Q）与键表示（记为K）之间的相似度。该相似度矩阵经 softmax 函数转换为概率分布后，用于对值表示（记为 V）进行加权求和。最后，将查询矩阵（Q）与该加权和相加，再经过一个层归一化模块。

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4.1}$$

$$Q^{'} = Q + Attention(Q, K, V) \tag{4.2}$$

$$O = LayerNorm(Q^{'}) \tag{4.3}$$

交叉注意力网络与自注意力模块类似，但有一个关键区别：查询矩阵和键/值矩阵由不同模态的表征构成。假设查询矩阵来自语音模态（记为 Q_S），键/值矩阵来自文本模态（记为 K_T/V_T），则交叉注意力网络的操作如下：

$$Atten(Q_S, K_T, V_T) = softmax(\frac{Q_A K_T^T}{\sqrt{d_k}})V_T \tag{4.4}$$

$$Q_S^{'} = Q_S + Atten(Q_S, K_T, V_T) \tag{4.5}$$

$$O_S = LayerNorm(Q_S^{'}). \tag{4.6}$$

4.1.2.2 所提方法

HCAM 第一阶段：在第一阶段，我们从语音和文本中训练话语级嵌入提取器。这些模型被训练用于对单个话语进行分类，而不考虑话语间对话上下文。本阶段训练的模型根据语音信号或文本转录中的线索，将单个话语分类到相应的情感类别。预训练特征提取器通常是计算需求很大的模型，这限制了每次迭代可处理的话语数量。单个对话中包含大量话语，阻碍了以往工作对这些特征提取器进行微调，因为它们将对话作为整体进行处理。我们的层次化建模使我们能够在本训练阶段微调来自预训练特征提取器的嵌入，从而改进对每个单独话语的情绪分类。

- **语音嵌入提取器：**语音被输入到 HuBERT large 模型 [10]。受 Pepino 等人提出的策略启发， [50], 我们对 HuBERT 网络中的 Transformer层进行微调，同时保持较低的卷积层不变。来自 HuBERT 模型的所有 Transformer层的隐藏层输出在帧级求和，并送入一维 CNN 网络。最后，对卷积神经网络的嵌入在话语级进行平均池化，从而为给定话语生成语音嵌入。在一维 CNN 网络输出端获得的嵌入将用于上下文 GRU 层。

- **文本嵌入提取器：**文本数据的嵌入提取器采用与语音特征提取类似的架构。我们通过预训练 RoBERTa 模型 [74]获取词嵌入，并拼接该模型最后四个隐藏层的表示。双向门控循环单元层能够融入话语内

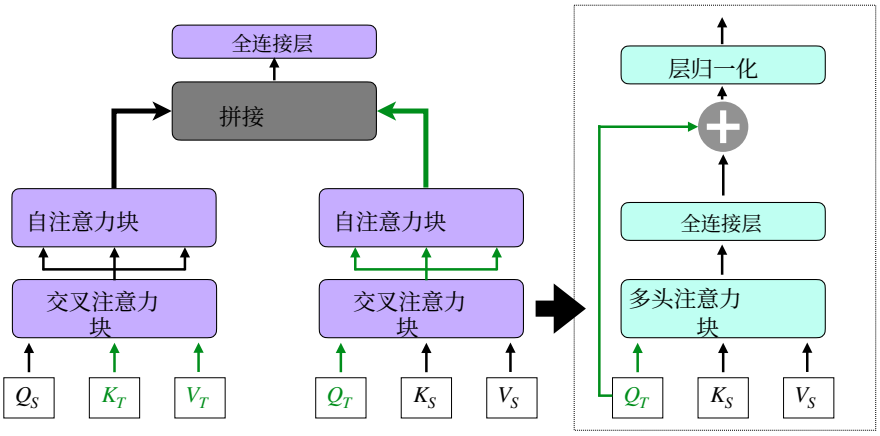


图4.3：所提出模型中使用的协同注意力网络。它由两个子块组成——交叉注意力块和自注意力块。

上下文，从而生成话语级嵌入。与语音嵌入提取不同，RoBERTa 和双向门控循环单元模型的所有层都会进行微调。

HCAM第二阶段： 在此阶段，我们使用前一建模阶段获得的话语级嵌入，并通过采用带有自注意力机制的双向门控循环单元（双向门控循环单元）架构引入话语间上下文。阶段 I 中针对每个单独话语训练的模型所提取的表征，将进一步结合对话上下文信息加以增强。

- **话语间上下文 GRU：** 我们提出了一个简单的块，称为上下文 GRU，它会考虑对话中所有话语的信息。虽然双向 GRU 本身能够整合上下文信息，但对于话语数量较多的对话，它可能无法捕捉长期依赖关系。因此，我们采用自注意力，以有效地建模长期上下文。自注意力块的输出由带有 ReLU 激活的逐位置前馈层处理。

HCAM第三阶段：模型的第三阶段包括对不同模态的嵌入进行有效融合。协同注意力块的网络架构如图4.3所示，由两个子模块组成，即交叉注意力和自注意力。在协同注意力网络中，我们利用模态之间的交叉注意力。语音与文本之间的交叉注意力有两种实现方式：一种是从语音数据中得到查询表示，而键和值表示则

从文本数据中得到。另一种方式则将语音和文本的角色互换。通过在交叉注意力块的输出中添加自注意力块，进一步丰富交叉模态嵌入。如图4.3所示，两个交叉注意力分支经过拼接，形成多模态表示。

最终表征融合了语音和文本中的信息，并被输入逐位置前馈层。

4.1.2.3 损失函数

监督式对比损失：我们探讨 Khosla 等人提出的监督式对比损失函数 [124]，，该函数鼓励来自同一情绪类别的话语表示彼此相似。设批量大小为 $\mathcal{B}$，其中，从多个对话中提取的话语级表征记为 $\{x_1, x_2, \ldots, x_{\mathcal{B}}\}$。此损失公式中使用的特征均经过归一化，即 $||x_i|| = 1 \quad \forall i = \{1, 2, \ldots, \mathcal{B}\}$。相应的标签记为 $\{y_1, y_2, \ldots, y_{\mathcal{B}}\}$。对于样本 x_j，小批次中的正样本集合记为 P_j： $\{i \in \mathcal{B} \quad s.t. \quad y_i = y_j\}$。监督式对比损失为：

$$L^{sup-con} = \sum_{j \in \mathcal{B}} \frac{-1}{|P_j|} \sum_{p \in P_j} log \frac{exp(x_j^T x_p / \tau)}{\sum\limits_{a \in \mathcal{B}} exp(x_j^T x_a / \tau)}, \tag{4.7}$$

其中， τ 是一个超参数，表示该损失的温度。通常，对比损失用于表征学习任务。对比损失依赖真实标签，因此也可用于监督设置。除了交叉熵损失之外，该损失还使系统关注难分类样本。然而，与表征学习任务不同，ERC 数据集规模不足以仅通过对比损失对话语进行分类。因此，我们将监督式对比损失与交叉熵损失结合起来，训练情绪识别模块。

组合损失：我们将交叉熵损失与监督式对比损失进行凸组合。用于训练模型所有阶段的最终损失如下：

$$L^{tot} = \beta CE(y, \hat{y}) + (1 - \beta) L^{sup-con}, \tag{4.8}$$

其中， β 是取值范围为 $[0, 1]$的超参数。此处， y 和ˆy 分别表示真实标签和模型预测的概率。

表4.1：不同数据集上的加权 F1 分数结果。

模态（训练阶段）	IEMOCAP		MELD	CMU-MOSI
	四分类	六分类		
语音（阶段 I）	64。3%	55。6%	48.2%	64。4%
语音（阶段 II）	78。7%	65.7%	50.1%	67.4%
文本（第一阶段）	68.4%	53.8%	63。3%	84.3%
文本（第二阶段）	81.4%	64。4%	65.6%	85。4%
语音 + 文本（第三阶段）	85.9%	70.5%	65。8%	85。8%

4.1.2.4 实验与结果

数据集： 我们在三个广泛使用的数据集上评估了本研究：IEMOCAP [96], MELD [98]和 CMU-MOSI [99]。这些数据集的详细信息见第 2.5节。

实现细节： 模型采用 Adam优化器进行训练，学习率为 $1e-5$，在 MELD 和 IEMOCAP 数据集上的批量大小为 32，而在 CMU-MOSI 数据集上则将批量大小减至 8。本文报告的所有实验均采用 5 种随机权重初始化方案。以下所有实验均报告随机初始化所得性能的平均值。我们将模型的不同阶段训练 100 个训练轮次，因为我们发现验证性能在此范围内趋于饱和。此外，我们在所有实现中都采用 L2范数为 0。25 的梯度裁剪。

结果： 我们先报告所提出模型在各个单独模态上的性能，随后报告模型在多模态设置下的性能。三个数据集上的关键结果如表 4.1 所示。根据这些结果，我们有以下观察：

1. 在 IEMOCAP 数据集上，语音和文本模态的表现相对相近；而在 MELD 和 CMU-MOSI 数据集上，文本转录表现出更强的模态优势。

2. 上下文添加框架对这三个数据集均有效。所提出的结合自注意力的上下文GRU架构虽然简单，却能提升两种模态的性能。对于语音，与第一阶段性能相比，IEMOCAP 四分类、IEMOCAP 六分类、MELD 和 CMU-MOSI 数据集的相对提升分别为 40.3%、22.7%、3.7% 和 8.4%。类似地，对于文本模态，我们观察到相对提升分别为 41.1%、22.9%、6.3% 和 7%。在这三个数据集中，IEMOCAP 的每段对话包含的话语数量最多，而 MELD 最少。随着数据集中话语数量的增加，

表4.2：采用不同训练范式时系统的加权 F1 分数

训练范式	IEMOCAP		MELD	CMU-MOSI
	四分类	六分类		
层次式	85。9%	70。5%	65.8%	85。8%
非层次式	82.3%	68.5%	65%	84.4%

上下文建模会变得更加有效

层次化建模的影响：为了了解我们层次化建模方法的优势，我们修改了训练范式，将训练的阶段 II 和 III 合并。训练范式修改后的结果如表所示4.2。我们注意到，在全部四种数据集设置下，在不改变模型架构的情况下，采用所提出的层次化建模都取得了更好的性能。这些实验凸显了 HCAM 框架所提出的各阶段采用课程式设计的优势。

4.1.2.5 总结与局限性

层次交叉注意力模型（HCAM）通过将单模态表征学习、上下文建模和多模态融合解耦，为对话中的多模态情感识别提供了一个结构化框架。HCAM 以分阶段的方式引入对话上下文，并采用交叉注意力机制进行语音-文本交互，从而有效捕捉时间依赖关系和跨模态相关性。

然而，这种架构也暴露出一个重要的局限性。当不同模态之间存在显著的性能差异时——例如语音或文本中的一方所包含的信息明显更丰富——融合表征并不总是优于表现最强的单模态系统（参见表4.1中的 MELD 和 CMU-MOSI 性能）。在这种情况下，多模态融合可能无法带来互补增益，甚至会削弱模态特定优势。

这一观察结果表明，我们需要一种更加灵活的融合策略，使模型能够进行动态模态选择，而不是强制执行统一的特征级集成。在下一小节中，我们将通过引入一种在决策层进行融合的模块化专家混合框架来解决这一局限，使模型能够根据对话上下文，自适应地利用单模态和多模态专家。

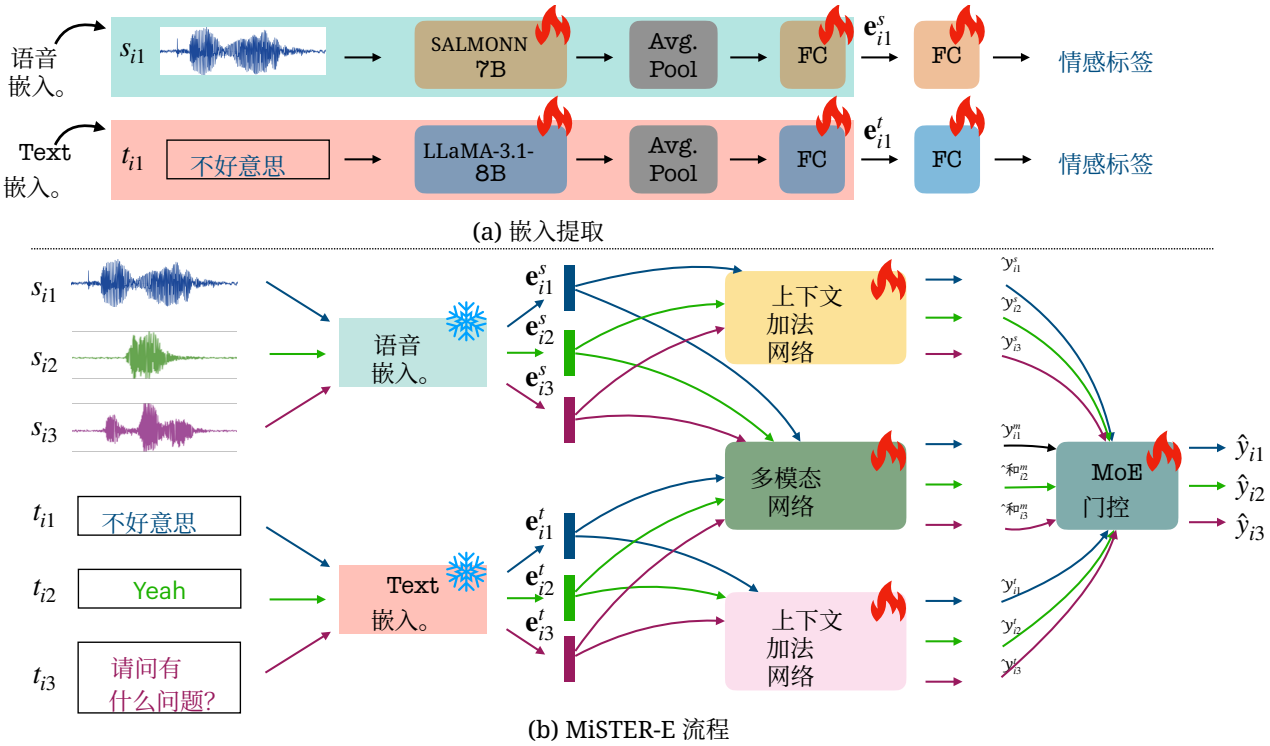


图 4.4：(a) 单模态特征提取模块的训练；(b) MiSTER-E 的完整流程。在训练流程的其余部分时，语音和文本嵌入模块保持冻结。针对两种模态分别训练两个上下文添加网络，同时训练一个多模态网络。最后，训练一个混合专家门控网络，以预测每个话语的情绪类别。

4.1.3 ERC 的混合专家建模

本节将介绍 MiSTER-E [125], ，这是一种在对数几率（决策）层而非特征层强制组合专家的模块化架构。MiSTER-E 采用混合专家（MoE）形式，由三个分别进行优化的分支组成：纯语音上下文专家、纯文本上下文专家和多模态专家。单模态系统使用时序Inception网络，随后连接循环层，以捕捉模态特定的对话动态；而多模态专家则通过交叉注意力和自注意力层融合语音与文本特征。这些专家的输出通过门控机制进行组合，该机制在 Softmax归一化之前计算加权和。这种对数几率级融合实现了动态专家选择，使模型能够结合互补的多模态线索。为了进一步稳定专家专业化过程，我们引入对比损失和基于KL的一致性损失，同时使用焦点损失对每个专家进行监督，以解决类别不平衡问题。特别是，当任一模态（语音或文本）相较于另一模态具有明显优势时

符号	描述
$\mathcal{D}$	对话情绪识别（ERC）数据集
P	数据集中对话的数量
$\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2、\ldots\ldots、\mathbf{c}_P\}$	$\mathcal{D}$ 中所有对话的集合
$\mathbf{c}_i$	第 i个对话
$\mathcal{U}_i = （\mathbf{u}_{i1}、\mathbf{u}_{i2}、\ldots\ldots、\mathbf{u}^{iN_i})$	对话$\mathbf{c}_i$中的话语序列
N_i	对话$\mathbf{c}_i$中的话语数量
$\mathbf{u}_{ik}$	对话$\mathbf{c}_i$中的第 k个话语
$\mathbf{s}_{ik}$	话语$\mathbf{u}_{ik}$的语音模态
$\mathbf{t}_{ik}$	话语$\mathbf{u}_{ik}$的文本模态
y_{ik}	与话语$\mathbf{u}_{ik}$对应的情绪标签
$\mathcal{Y}$	数据集中的情感类别集合
$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iN_i})$	对话$\mathbf{c}_i$的情感标签序列
$\mathbf{e}^s_{ik}, \mathbf{e}^t_{ik}$	话语$\mathbf{u}_{ik}$的语音与文本嵌入
$\hat{\mathbf{y}}^s_{ik}, \hat{\mathbf{y}}^t_{ik}, \hat{\mathbf{y}}^m_{ik}$	语音、文本和多模态专家的softmax前logits
$\hat{\mathbf{y}}_{ik}$	经过混合专家门控后的最终融合logit
$\boldsymbol{\beta}_{ik}$	话语 $\mathbf{u}_{ik}$ 的混合专家权重
$\hat{\mathbf{p}}^s_{ik}, \hat{\mathbf{p}}^t_{ik}, \hat{\mathbf{p}}^m_{ik}$	语音、文本和多模态专家的类别概率分布
$\mathcal{L}^i_{\text{CAN}}$	单模态上下文添加网络的焦点损失
$\mathcal{L}^i$	多模态分类与对比损失
$\mathcal{L}^{多}_{\text{moe}}$	带 KL散度一致性正则化的 MoE 门控损失
$\mathcal{L}_{\text{tot}}$	整体训练目标

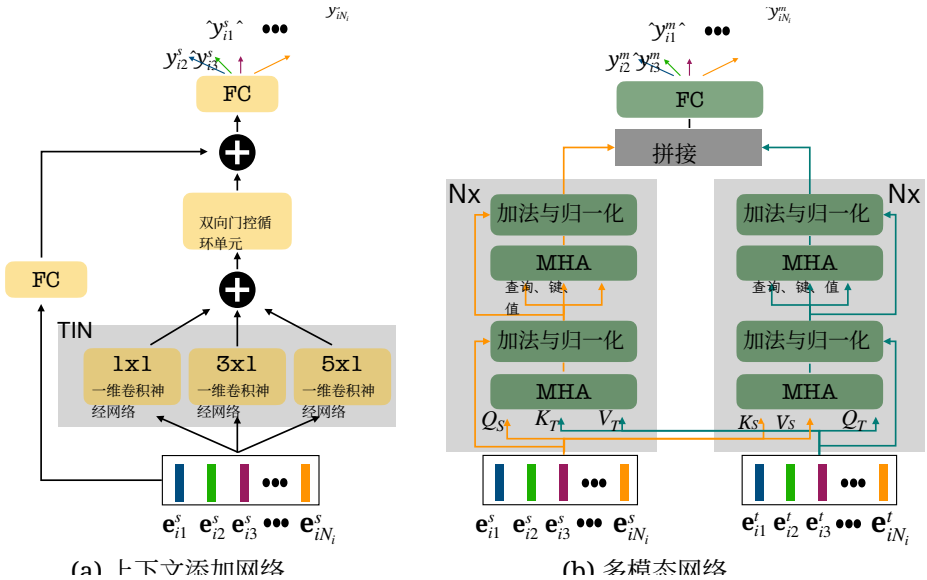
表 4.3：问题形式化和模型描述中使用的符号。

与另一种模态（文本或语音）相比，我们采用模态解耦后再基于 MoE 进行组合，相较于以往其他工作，展现出独特的融合优势。我们所提方法的框图如图 4.4 所示。此外，表 4.1.1 节的问题形式化和模型描述中使用的不同符号汇总见表 4.3。

4.1.3.1 所提方法

单模态特征提取：尽管大语言模型（LLM）擅长生成文本，但其在情感识别中的应用仍然有限。以往的研究通常在自回归模式下使用它们，或将其用作冻结的特征提取器 [126]。

在这项工作中，我们对大语言模型——具体而言，`LLaMA-3.1-8B`——进行微调，使其充当文本编码器，而非生成器。每段话语转录文本 $\mathbf{t}_{ik}$ 都会经过分词，并输入大语言模型处理。随后，对词元级隐藏状态进行池化，并将其传入一个通过任务特定监督训练的两层前馈分类器。为了保留预训练知识，同时实现高效适配，我们采用 LoRA [127]对权重进行微调。我们将由此得到的



(a) 上下文添加网络　　(b) 多模态网络

图 4.5：MiSTER-E 中使用的 (a) 上下文添加网络（用于语音模态）和 (b) 多模态网络。两个块的输入均来自单模态特征提取模块。 TIN 表示时间 Inception 网络， MHA 表示多头注意力。

文本嵌入表示为：

$$\mathbf{e}_{ik}^{t} = \texttt{Text-Embed}(\mathbf{t}_{ik}) \tag{4.9}$$

其中，$\mathbf{e}_{ik}^{t}$ 是从分类器的第一个全连接层提取的。
对于语音，我们采用类似的方法，使用 SALMONN-7B [115]，这是一种语音大语言模型（SLLM），由语音编码器、Q-Former [128], 和大语言模型主干网络组成。对于 ERC，我们通过 LoRA 更新 Q-Former、大语言模型和分类头，对该模型进行微调。这使系统能够学习情感显著的声学模式，同时保留大语言模型主干网络的语义特征。给定语音信号 $\mathbf{s}_{ik}$，我们提取其表征如下：

$$\mathbf{e}_{ik}^{s} = \texttt{Speech-Embed}(\mathbf{s}_{ik}) \tag{4.10}$$

其中，嵌入取自语音分类头的第一个全连接层。

对话建模： 对于一个对话 $\mathbf{c}_i$，文本和语音嵌入序列分别表示为 $\mathbf{E_i^t} = (\mathbf{e}_{i1}^{t}, \ldots, \mathbf{e}_{iN_i}^{t})$ 和 $\mathbf{E_i^s} = (\mathbf{e}_{i1}^{s}, \ldots, \mathbf{e}_{iN_i}^{s})$。如图 4.4 所示，经过微调的语音和文本嵌入提取器被冻结。现在我们讨论用于在 ERC 中建模对话上下文的块。

- **上下文添加网络：**为了让话语表示具备上下文感知能力，我们

提出了一种上下文添加网络（CAN），利用对话上下文增强文本和语音嵌入。其核心是时间 Inception 网络（TIN），该网络采用卷积核尺寸为 1、3 和 5 的一维卷积来捕捉短程依赖，从而模拟话语局部邻域中不同的感受野。TIN 的设计受到最初为图像分类开发的 Inception 架构 [129] 的启发，能够在对话中实现多尺度时间上下文建模。

然而，对话中的情绪信号往往会在更长的时间跨度内演变。为捕捉此类全局依赖，我们添加一个双向 GRU 层，使模型能够整合整个对话跨度内的信息。通过残差连接将原始嵌入与其上下文增强版本相连，从而在保留基础大语言模型特征的语义基础的同时进行加性优化。最后，使用一个全连接层将每个话语 $\mathbf{u}_{ik}$ 映射到相应的情绪类别 y_{ik}。我们分别为语音和文本模态训练两个相似的网络。这些操作表示为：

$$\hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{s}} = (\hat{y}_{i1}^{s}, \ldots, \hat{y}_{iN_i}^{s}) = \texttt{CAN}(\mathbf{E}_{\mathbf{i}}^{\mathbf{s}}) \tag{4.11}$$

$$\hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{t}} = (\hat{y}_{i1}^{t}, \ldots, \hat{y}_{iN_i}^{t}) = \texttt{CAN}(\mathbf{E}_{\mathbf{i}}^{\mathbf{t}}) \tag{4.12}$$

语音侧的示意图 CAN 见图4.5（a）。

- **多模态网络：**为整合不同模态的信息，我们设计了一个基于语音与文本之间交叉注意力的融合模块。该机制使模型能够将文本中的语义线索与音频中的情感信号对齐。首先，将语音与文本嵌入 $\mathbf{E_{si}}$ 和 $\mathbf{E_{ti}}$分别投影到查询、键和值空间。随后双向应用交叉注意力：

$$\mathbf{M}_{\mathbf{i}}^{\mathbf{t}\to\mathbf{s}} = \texttt{LN}(\texttt{FC}(\mathbf{E}_{\mathbf{i}}^{\mathbf{s}}) + \texttt{MHA}(\mathbf{Q}_{\mathbf{i}}^{\mathbf{s}}, \mathbf{K}_{\mathbf{i}}^{\mathbf{t}}, \mathbf{V}_{\mathbf{i}}^{\mathbf{t}})) \tag{4.13}$$

$$\mathbf{M}_{\mathbf{i}}^{\mathbf{s}\to\mathbf{t}} = \texttt{LN}(\texttt{FC}(\mathbf{E}_{\mathbf{i}}^{\mathbf{t}}) + \texttt{MHA}(\mathbf{Q}_{\mathbf{i}}^{\mathbf{t}}, \mathbf{K}_{\mathbf{i}}^{\mathbf{s}}, \mathbf{V}_{\mathbf{i}}^{\mathbf{s}})) \tag{4.14}$$

其中， MHA 和 LN 分别表示多头注意力和层归一化。虽然这有助于实现跨模态对齐，但却忽略了对 ERC 至关重要的对话上下文线索。为了解决这一问题，我们在对齐表征上应用模态特定自注意力层，以捕捉整个对话中的时间上下文：

$$\mathbf{M}_{\mathbf{i}}^{\mathbf{s}} = \texttt{Self-attn.}(\mathbf{M}_{\mathbf{i}}^{\mathbf{t}\to\mathbf{s}}) \tag{4.15}$$

$$\mathbf{M_i^t} = \texttt{Self-attn.}(\mathbf{M_i^{s\to t}}) \tag{4.16}$$

将输出 $\mathbf{M_i^s}$ 和 $\mathbf{M_i^t}$ 拼接后，输入全连接层：

(FC)

$$\hat{\mathbf{y}}_\mathbf{i}^\mathbf{m} = (\hat{y}_{i1}^m, \hat{y}_{i2}^m, \ldots, \hat{y}_{iN_i}^m) = \texttt{FC}([\mathbf{M_i^s}; \mathbf{M_i^t}]) \tag{4.17}$$

概述如图4.5(b)所示。

混合专家门控：我们针对每条话语 u_{ik}从三个专家获取softmax前logits：语音专家（$\hat{}\mathbf{y}_{\mathbf{ik}}^\mathbf{s}$）、文本专家（$\hat{}\mathbf{y}_{\mathbf{ik}}^\mathbf{t}$）以及多模态专家（$\hat{}\mathbf{y}_{\mathbf{ik}}^\mathbf{m}$）。为组合这些逻辑值，我们使用一种学习如何对其加权的门控机制，动态融合各专家的逻辑值。为计算这些权重，我们将逻辑值拼接后送入全连接层（FC）：

$$\mathbf{g}_{ik} = \texttt{FC}([\hat{\mathbf{y}}_{\mathbf{ik}}^\mathbf{s}; \hat{\mathbf{y}}_{\mathbf{ik}}^\mathbf{t}; \hat{\mathbf{y}}_{\mathbf{ik}}^\mathbf{m}]) \tag{4.18}$$

门控网络输出$\mathbf{g}_{ik} \in \mathbb{R}^3$，经Softmax运算变换后得到自适应混合权重 $\boldsymbol{\beta}_{ik}$。

最终逻辑值计算为专家预测的加权和：

$$\hat{\mathbf{y}}_{\mathbf{ik}} = \beta_{ik}^s \cdot \hat{\mathbf{y}}_{\mathbf{ik}}^\mathbf{s} + \beta_{ik}^t \cdot \hat{\mathbf{y}}_{\mathbf{ik}}^\mathbf{t} + \beta_{ik}^m \cdot \hat{\mathbf{y}}_{\mathbf{ik}}^\mathbf{m} \tag{4.19}$$

对$\hat{}\mathbf{y}_{\mathbf{ik}}$ 应用Softmax运算，以获得话语 u_{ik}的最终预测分布。该门控网络以端到端方式训练，使模型能够通过强调最可靠的专家来灵活整合各模态。

4.1.3.2 模型训练

损失函数：ERC 通常具有严重的类别不平衡问题，罕见的情感类别往往会被错误分类 [12]。为解决这一问题，我们采用焦点损失 [130], ，根据每个训练样本的复杂程度调整其贡献，降低分类正确样本的相对损失，同时重点关注难分类的样本（可能属于少数类）。在 MiSTER-E 中，该损失用于训练文本和语音的单模态嵌入提取器，以及各自的上下文添加网络（CAN）。语音和文本 CAN 网络的损失定义如下：

$$\mathcal{L}_{\text{CAN}}^i = \sum_{k=1}^{N_i} \texttt{FL}(\hat{y}_{ik}^s, y_{ik}) + \sum_{k=1}^{N_i} \texttt{FL}(\hat{y}_{ik}^t, y_{ik}) \tag{4.20}$$

这里， $\mathtt{FL}(\cdot)$ 表示焦点损失函数。设逻辑值为 $\hat{y}_{ik}^{\cdot}$ 的样本的预测概率为 $\hat{p}_{ik}^{\cdot}$，则焦点损失定义为：

$$\mathtt{FL}(\hat{y}_{ik}^{\cdot}, y_{ik}) = -\sum_{c=1}^{y} y_{ik}^{c}(1-\hat{p}_{ik}^{\cdot c})^{\gamma}\log(\hat{p}_{ik}^{\cdot c}) \tag{4.21}$$

其中，对于真实类别， $y_{ik}^{c}=1$ 取该值，否则为 0。$\hat{p}_{ik}^{\cdot c}$ 是样本中类别 c 的预测概率，其逻辑值为 $\hat{y}_{ik}^{\cdot}$ （$\cdot$ 分别替换为 s、 t 或 m），而 γ 是与焦点损失相关的超参数。

多模态对比损失： 在许多情况下，语音和文本能够提供关于说话者情绪的互补信号。为利用这一点，我们引入监督式对比损失，根据情感标签构建联合表示空间，使跨模态中具有相同情感意图的话语彼此靠近。考虑以下多模态语音和文本表征，分别记为 $\mathbf{M_i^s}=\{\mathbf{m_{i1}^s},\mathbf{m_{i2}^s},\mathbf{m_{i3}^s},\ldots,\mathbf{m_{iN_i}^s}\}$ 和 $\mathbf{M_i^t}=\{\mathbf{m_{i1}^t},\mathbf{m_{i2}^t},\mathbf{m_{i3}^t},\ldots,\mathbf{m_{iN_i}^t}\}$。将这些嵌入组成批次，得到，

$$\mathbf{Z_i}=\{\tilde{\mathbf{m}}_{\mathbf{i1}}^{\mathbf{s}},\ldots,\tilde{\mathbf{m}}_{\mathbf{iN_i}}^{\mathbf{s}},\tilde{\mathbf{m}}_{\mathbf{i1}}^{\mathbf{t}},\ldots,\tilde{\mathbf{m}}_{\mathbf{iN_i}}^{\mathbf{t}}\} \tag{4.22}$$

其中， $\tilde{\mathbf{m}}_{\mathbf{ik}}^{\cdot}=\frac{\mathbf{m}_{\cdot\mathbf{ik}}}{\|\mathbf{m}_{\cdot\mathbf{ik}}\|}$ 表示归一化嵌入。令 $\mathbf{z}_a\in\mathbf{Z_i}$ 表示 $a\in\{1,2,\ldots,2N_i\}$，并令 y_a 为与 $\mathbf{z}_a$相关联的情绪标签。对比损失定义如下：

$$\mathcal{L}_{\mathrm{con}}^{\mathrm{i}}=\sum_{a=1}^{2N_i}\frac{-1}{|P(a)|}\sum_{p\in P(a)}\log\frac{\exp\left(\mathbf{z}_a^{\top}\mathbf{z}_p/\tau\right)}{\sum\limits_{\substack{q=1\\q\neq a}}^{2N_i}\exp\left(\mathbf{z}_a^{\top}\mathbf{z}_q/\tau\right)} \tag{4.23}$$

其中 τ是温度， $P(a)=\{p\in\{1,2,\ldots,2N_i\}\backslash\{a\}|y_p=y_a\}$ 是锚点 $\mathbf{z}_a$ 的正样本集合。该目标将语音和文本中具有相同情绪类别的话语拉近，同时将其他样本推开，从而引导模型学习情感连贯的模态不变表示。

总多模态损失：用于训练多模态模型的最终损失结合了分类与对比目标：

$$\mathcal{L}_{\mathrm{multi}}^{i}=\sum_{k=1}^{N_i}\mathtt{FL}(\hat{y}_{ik}^{m},y_{ik})+\lambda\mathcal{L}_{\mathrm{con}}^{i} \tag{4.24}$$

其中，$\hat{y}_{ik}^{m}$ 表示多模态预测， λ 控制对比损失的贡献。

MoE 门控损失： 为了训练专家混合（MoE）门控网络，我们结合了两个目标：(i) 用于情绪分类的焦点损失，以及 (ii) 用于促进专家预测一致性的正则化项。具体而言，我们使用 Kullback-Leibler (KL) 散度，约束三个专家（仅语音、仅文本和多模态）的预测分布保持相似。MoE 层的总损失为：

$$\mathcal{L}^{i}_{\text{moe}} = \sum_{k=1}^{N_i} \text{FL}(\hat{y}_{ik}, y_{ik}) + \alpha \cdot \mathcal{L}^{i}_{\text{KL}} \tag{4.25}$$

其中 $\hat{}y_{ik}$ 定义于式 4.19， α 控制一致性正则化的强度，且 KL 项由下式给出：

$$\mathcal{L}^{i}_{\text{KL}} = \sum_{k=1}^{N_i} \Big[\text{KL}(\hat{p}^{m}_{ik} || \hat{p}^{s}_{ik}) + \text{KL}(\hat{p}^{m}_{ik} || \hat{p}^{t}_{ik}) \Big] \tag{4.26}$$

此处，$\hat{}p_{ik}$ 表示通过对相应专家 logits $\hat{}y_{ik}$应用 softmax 函数得到的类别概率分布。KL 一致性损失促使单模态语音和文本专家产生与多模态专家对齐的预测分布。由于多模态专家联合建模两种模态，我们将其输出分布视为更强的监督信号，因此以这种方式采用有向 KL 散度。

总损失：上下文添加网络、多模态网络和 MoE 门控层联合训练。总损失为：

$$\mathcal{L}_{\text{tot}} = \sum_{i=1}^{P} \Big[\mathcal{L}^{i}_{\text{CAN}} + \mathcal{L}^{i}_{\text{moe}} + \mathcal{L}^{i}_{\text{multi}} \Big] \tag{4.27}$$

4.1.3.3 实验

数据集：与先前方法相同，我们在 IEMOCAP、MELD 和 CMU-MOSI 上进行实验。在本实验中，我们仅评估 IEMOCAP 中的六分类问题，因为该设定在近期工作中已被更广泛使用。

实现细节： 两个单模态特征提取器均使用 LoRA 进行训练，秩设为 8，缩放参数设为 32，*dropout*设为 0.1。两个模型的批量大小均设为 8，学习率设为 $1e-5$，并采用焦点损失。全连接层（第 4.1.3.1 节）中的隐藏维度设为 2048。对于 MiSTER-E 流程的其余部分，IEMOCAP 和 MOSI 的批量大小设为 8，MELD 的批量大小设为 32。所有数据集的学习率均设为 $1e-5$。

。对于 MELD 和 MOSI，λ（式4.24）设为 1；对于 IEMOCAP，则设为 2。一致性正则化项 α（式4.25）在 IEMOCAP 和 MOSI 上设为 0.1，而在 MELD 上保持为 $1e-3$（第4.1.3.1 节）。我们报告测试数据上的加权 F1 分数作为性能指标，结果取 3 次随机初始化。

我们的所有实验均使用 NVIDIA RTX A6000 GPU 卡，并配合 Pytorch 2.7.0 与 CUDA 12.6 运行。我们模型中的可训练参数数量共计 97M（大语言模型/小型语言模型的 LoRA 训练各 8M，对话建模为 81M）。训练大语言模型每个训练轮次约需 10 分钟，而小型语言模型每个训练轮次约需 20 分钟。特征编码器之后的模型其余部分进行 100 个训练轮次大约需要 10 分钟。我们在下文进一步说明两个数据集的超参数选择：

- 在 CAN 网络（第4.1.3.1节）中使用的双向门控循环单元的隐藏维度为 512，我们对 IEMOCAP、MOSI 使用 3 层，对 MELD 使用 2 层。我们在每个全连接层之间使用 0.2 的 dropout 进行正则化。

- 对于多模态融合网络，我们使用 4 层、隐藏维度为 120 以及 4 个注意力头。我们在多模态融合网络中使用 0.5 的 dropout 正则化。两个数据集均使用参数相同的融合网络。用于融合网络的对比损失（公式4.23）的温度参数在 IEMOCAP、MOSI 上保持为 1，在 MELD 上保持为 0.05。

- 训练上下文添加网络、多模态网络和混合专家门控网络时，我们对两个数据集均采用范数为 1.0 的梯度裁剪，并训练 100 个训练轮次。训练大语言模型和小型语言模型特征编码器时，我们在两个数据集上均训练 50 个训练轮次 [1]。

与先前工作的比较：表4.4 展示了所提方法与若干基线方法的比较。竞争方法涵盖了多种建模范式，包括基于图的框架 [131, 132]，状态空间表述 [133]，以及基于专家混合（MoE）的模型 [134]。可以看出，所提方法在全部三个数据集上均取得了新的最先进性能，加权 F1 分数在 IEMOCAP 上达到**70.9%**，在 MELD 上达到**69.5%**，在 CMU-MOSI 上达到**87.9%**。

此外，值得指出的是，对于 IEMOCAP，由于缺少标准化验证集，以往研究往往直接在测试集上进行模型选择。这种做法可能会人为抬高性能指标，并降低可复现性。为缓解这一问题，本研究中所有实验（针对有官方实现的实验）均采用我们自行划分的训练集、验证集和测试集

[1] MiSTER-E 的代码可在以下地址获取：https://github.com/iiscleap/MiSTER-E。

方法	IEMOCAP（6类别）	MELD（7类别）	MOSI（2类别）
SMIN [97]##	70。5%	63。7%	79.7%
MultiEmo [135]#	66。9%$^{\pm 2.0}$	65.3%$^{\pm 0.5}$	83。7%$^{\pm 0.3}$
HCAM [123] (第 4.1.2.2节)	70。5%	65。8%	85.8%
DF-ERC [136]	69.5%	64.5%	–
DER-GCN [131]	64.7%	62.6%	–
ELR-GNN [132]	64。4%	63.2%	–
Mamba类模型 [133]	70。2%	65。6%	–
TELME [77]	69.3%	67。2%	–
CFN-ESA [78]	68.7%	67。2%	–
MMGAT-EMO [134]#	65。5%$^{\pm 0.6}$	66.1%$^{\pm 0.4}$	84.3%$^{\pm 0.4}$
MiSTER-E	**70.9**%$^{\pm 0.2}$	**69.5**%$^{\pm 0.3}$	**87.9**%$^{\pm 0.2}$

表 4.4：在 IEMOCAP、MELD 和 MOSI 数据集上使用加权 F1 分数对比不同方法。所有报告结果均针对语音-文本系统。## 我们报告了训练 SMIN 时未使用外部情感数据的结果。 # 我们在自己的实验设置下运行了作者提供的公开实现。上标结果表示在适用情况下，3 次随机初始化所得结果的均值和标准差。

进行。因此，表 4.4 中报告的部分 IEMOCAP 结果可能与原始发表的结果有所不同。

模块化至关重要吗? ：为评估架构模块化的重要性，我们设计了一个去除基于专家的结构的模型单体变体。在此基线中，各模态的上下文表示会直接传入融合模块，绕过模态专属专家与多模态专家之间的分离。这一改动导致性能显著下降：在 IEMOCAP 数据集上，加权 F1 分数从 70.9% 降至 67.8%；在 MELD 上则从 69.5% 降至 67.9%。这些结果凸显了一个关键见解：仅仅在统一流中融合上下文表示是不够的。相反，像 MiSTER-E 中那样明确地将上下文建模与跨模态融合解耦，能够使各个组件专门化，从而更有效地为最终预测作出贡献。

大语言模型嵌入是万能良方吗? ：细心的读者可能会问，我们的方法带来的收益是否仅仅源于使用了大语言模型/小型语言模型特征。为了将架构设计的影响与输入表示的影响区分开来，我们开展了受控实验，使用与所提出方法相同的大语言模型和小型语言模型特征，重新训练了若干已有的 ERC 模型。由于这些基线系统最初针对不同类型的特征进行了优化，我们为每个模型自行进行超参数调优，以确保比较公平（见表 4.5）。研究结果表明，尽管所有基线系统都在一定程度上受益于改进后的表征，MiSTER-E 的表现仍优于此前发表的基线系统。这凸显了一个重要发现：我们的框架所取得的收益并非

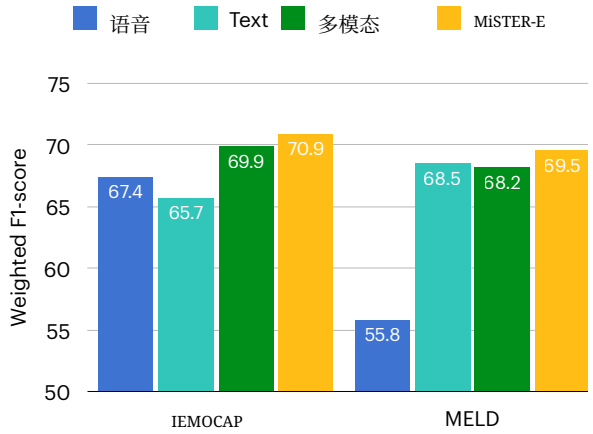
方法	# 参数	文本特征	语音特征	IEMOCAP		MELD
MultiEmo [135]	≈ 450M	RoBERTa	OpenSMILE	66.9%		65.3%
+ 大语言模型/小型语言模型	≈ 14B	LLaMA	SALMONN	69.5%		68.2%
HCAM [123]	≈ 750M	RoBERTa	wav2vec	70。5%		65。8%
+ 大语言模型/小型语言模型	≈ 14B	LLaMA	SALMONN	70。6%		68。4%
MMGAT-EMO [134]	≈ 450M	EmoBERTa	OpenSMILE	65.5%		66.1%
+ 大语言模型/小型语言模型	≈ 14B	LLaMA	SALMONN	68.6%		67.7%
MiSTER-E（无 LLM/SLLM）	≈ 450M	RoBERTa	OpenSMILE	69.5%		66。4%
MiSTER-E 无 LLM/SLLM	≈ 750M	RoBERTa	wav2vec	**71.1**%		68。4%
MiSTER-E	≈ 14B	LLaMA	SALMONN	70。9%		**69.5**%

表4.5：对部分基线方法在使用与不使用大语言模型特征重新设计后的比较。结果表明，该方案即使对于OpenSMILE和wav2vec等非大语言模型表征也同样适用。

仅仅归功于强大的嵌入，而是源自模块化架构设计——尤其是将上下文建模与跨模态融合分离，并采用专家级监督和门控。

为进一步评估MiSTER-E的性能提升是源于架构设计而非编码器规模，我们构建了MiSTER-E的两个变体，将大语言模型/小型语言模型编码器替换为更常规的预训练表征：文本使用RoBERTa，语音使用 {OpenSMILE、wav2vec} 特征。尽管模型容量大幅降低，MiSTER-E的两个变体均持续优于所有使用相同特征表征的基线方法（表4.5）。值得注意的是，在IEMOCAP上，结合RoBERTa与wav2vec特征的MiSTER-E取得了与相应大语言模型/小型语言模型配置相当、且在某些情况下更高的性能，表明在模型容量降低时泛化能力有所提升。这些实验表明，本方法的优势并不依赖于任何特定的预训练嵌入集合，而是源于本文所提出的建模设计。

决策级 MoE 门控是否至关重要? ：为了评估决策级专家混合（MoE）门控策略的重要性，我们通过比较 MiSTER-E 与两种架构变体开展消融实验。在第一种变体中，我们将融合点移至特征空间，即在分类之前组合特定模态和多模态表示，而不是融合它们的预测结果。这一设计用于检验更早进行整合是否足够。在第二种变体中，我们完全取消专家级监督。在这种设置下，模型仅使用焦点损失，基于最终门控预测进行训练，不再使用音频、文本和多模态各自的损失信号



(a) MiSTER-E在IEMOCAP和MELD上随混合专家门控策略变化的性能以及

模态	模型	IEMOCAP	MELD
Text	Gemma-7B	48.2%	58。1%
	Qwen2.5-7B	53.1%	65.5%
	LLaMA-3.1-8B	**55.3**%	**67.1**%
Speech	SALMONN-7B	**59.7**%	54.3%
	Qwen2-Audio-7B	59.2%	**54.9**%
	SALMONN-13B	58。7%	53。2%

(b) 不同 SLLM 和大语言模型在两个数据集上的表现。

图 4.6：IEMOCAP 和 MELD 上的模型性能比较。(a) MiSTER-E 中不同混合专家门控策略的影响。(b) 两个数据集上单模态语音大语言模型和大语言模型的性能。

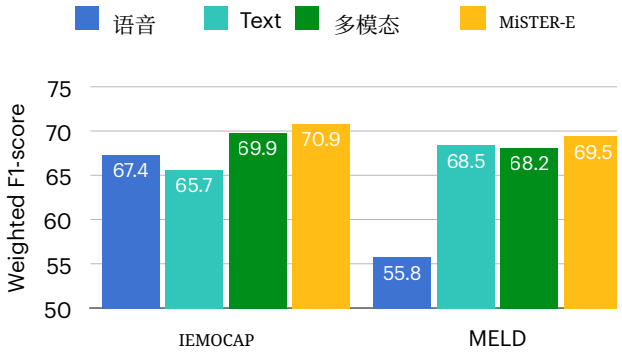
专家。如图所示 4.6a，这两种变体都会导致性能下降：`feat-MoE` 导致 IEMOCAP 上的结果下降 2。6%，MELD 上下降 0。6%；而 `No-Loss-MoE`也会导致类似的性能下降。这些发现进一步印证了 MiSTER-E 背后的两项关键设计原则：（1）决策级融合比早期特征融合更有效，这可能是因为它能够保留模态特定判别线索；（2）专家特定监督对于促使每位专家专注于各自的模态或融合模式至关重要。

其他大语言模型/小型语言模型的表现 我们使用其他语音和语言大型模型（分别为语音大语言模型和大语言模型）对语音和文本进行编码。所有模型均采用与 `LLaMA-3.1-8B` 和 `SALMONN - 7B`相同的超参数进行训练。对于文本模态，我们测试了 `Qwen2.5-7B` 和 `Gemma-7B`；对于语音，则使用 `Qwen2-Audio-7B` 和`SALMONN-13B`。

如表 4.6b所示，各模型的语音编码器表现相当，其中 `SALMONN-7B` 在 IEMOCAP 上取得最佳结果，而 `Qwen2-Audio-7B` 在 MELD 上表现最佳。较大的 `SALMONN-13B` 表现不及其较小的对应版本。对于文本，`LLaMA-3.1-8B` 在两个数据集上的表现都优于 `Gemma-7B` 和 `Qwen2.5-7B` 。



(a) 专家权重分布



(b) 专家在两个数据集上的表现

图 4.7: （a）不同数据集上各专家权重的分布；（b）两个数据集上各专家与 MiSTER-E 的性能

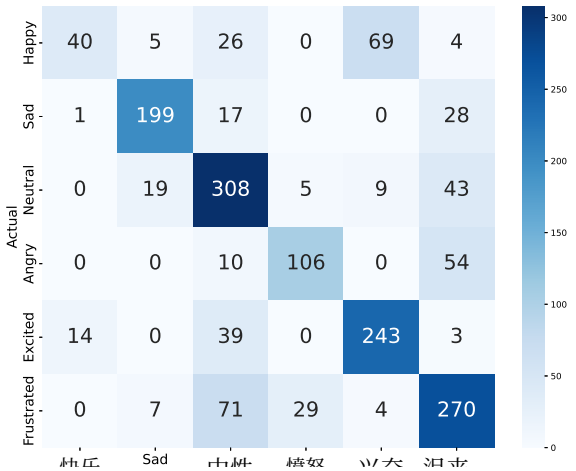
4.1.3.4 讨论

专家行为与模态不平衡： 我们分析 MoE 门控机制在 IEMOCAP 和 MELD 上的行为，以了解模型如何适应数据集特定的模态特征。如图4.7(a)所示，门控网络表现出明显的数据集依赖偏好：在 IEMOCAP 上为多模态专家分配较高权重，而在 MELD 上则偏向文本专家。这种行为反映了两种数据集中各模态的相对可靠性。

这些趋势与图4.7(b) 中各个专家的性能表现一致。在 MELD 上，语音专家的表现远逊于文本专家；由于不同预测能力的模态难以对齐，多模态专家的表现略逊于文本专家。相比之下，在语音和文本模态表现相近的 IEMOCAP 上，多模态专家能够有效融合两种信息，性能比最佳单模态专家高出 2.5%。在两个数据集上，MiSTER-E 的表现始终优于表现最好的单个专家约 1%，这表明自适应的逐话语专家加权具有优势。

专家一致性正则化的效果与这些观察结果相符。对于 IEMO-CAP，适中的一致性强度（$\alpha = 0.1$）可将性能从 70.4% 提升至 70.9%，而更强的正则化则会降低性能。对于模态不平衡较为明显的 MELD，较小的取值（$\alpha = 10^{-3}$）可带来 0.1% 的小幅提升，而较大的取值则会显著降低性能。

MiSTER-E 的逐类性能表中展示了我们提出的方法在两个数据集上的逐类性能4.8b。我们注意到，对于 IEMOCAP，表现最差的



（a）我们的模型在 IEMO-CAP 数据集上的混淆矩阵。

IEMOCAP		MELD	
类别	F1	类别	F1
愤怒	68。4%	愤怒	62.5%
兴奋	77.9%	厌恶	42。9%
沮丧	69。0%	Fear	30。6%
快乐	40.2%	Joy	65.4%
中性	80.2%	中性	81.3%
Sad	83.8%	Sad	50。3%
-	-	惊讶	61.5%

(b) IEMOCAP 和 MELD 各类别的 F1分数。

图 4.8：(a) IEMOCAP 的混淆矩阵。(b) IEMOCAP 和 MELD 上各类别的 F1分数，展示了不同情绪之间的差异。

类别是“快乐”——这可能是因为它是出现频率最低的类别，而且最容易与“兴奋”类别混淆（图4.8a）。同样，对于 MELD，模型在“恐惧”类别上的表现最差，这也是出现频率最低的类别。不过，我们注意到，对于这两个数据集，模型在大多数情感类别上的表现都相当不错，因此整体性能有所提升。

对比损失的重要性：在 MELD 上，引入对比目标后，性能从 68.2% ($\lambda = 0$) 提升至 $\lambda = 1$ 时的 69.5%；而在 IEMOCAP 上，性能则从 70.2% 提升至 $\lambda = 2$ 时的 70.9%。这些提升表明，与单独使用焦点损失相比，对比学习有助于对齐与情绪相关的语音和文本表征。然而，进一步增大 λ 会导致两个数据集上的性能下降，这表明对比损失是一种有益的辅助信号，而非独立的监督目标。

案例研究：我们在图4.9中展示了一个来自 MELD 数据集的案例研究。以下是我们希望从该示例中着重指出的几点：

- 遇到情绪变化时，MiSTER-E 仍能正常工作。例如，这段对话在 7 个话语中包含 6 种情绪。尽管模型错误地将第三个话语预测为中性（而非喜悦），但它仍能预测出情绪从惊讶到悲伤再到愤怒的变化。

- 另一个 i 有趣的一点是第五个话语的输出。没有任何专家预测 t

为悲伤类别，但 MoE 混合专家策略正确地将该话语标记为悲伤。这表明了混合专家门控策略在我们提出的方法中的效用，也使其有别于静态集成技术。

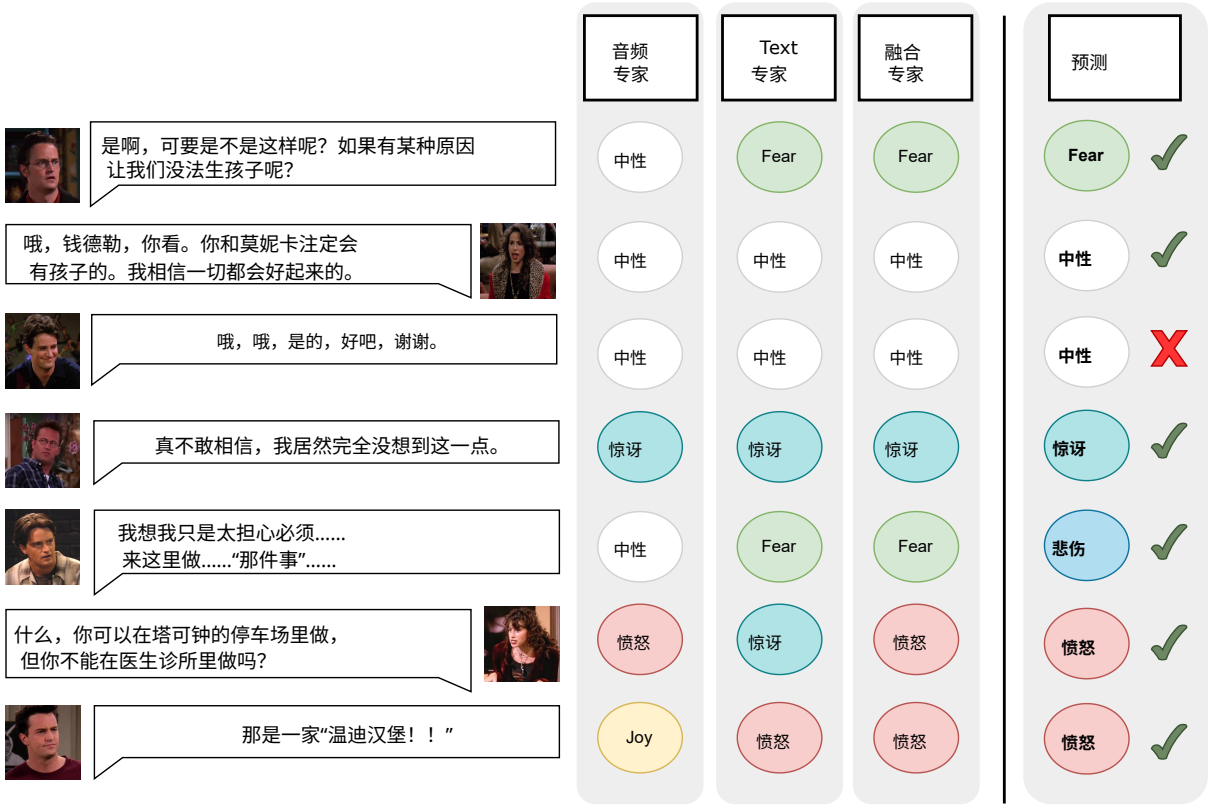


图 4.9：MELD 中的一个案例研究，展示了对话过程中发生的情绪变化。在这段对话的 7 个话语中，MiSTER-E 正确预测了 6 个。

4.1.3.5 MiSTER-E 的总结与局限

我们提出了 MiSTER-E，这是一种用于 ERC 的模块化框架，将上下文建模与多模态融合明确分离。我们利用语音和文本的基于大语言模型的表征，通过时间起始模块和双向门控循环单元对对话中的上下文进行建模，并通过基于注意力的网络进行模态融合。混合专家（MoE）门控机制能够自适应地整合上下文感知专家与多模态专家的决策。MiSTER-E 在 I EMOCAP、MELD 和 CMU-MOSI 上取得了新的最先进性能，证明了模块化上下文融合和决策级门控的有效性。尽管 MiSTER-E 取得了强劲

g p在多个 p基准测试中的表现，若干局限性

仍存在以下局限性：(i) 尽管采用了参数高效微调，大型LLM/SLLM编码器仍会带来显著的计算和内存开销，可能限制其在资源受限或实时环境中的适用性。我们在表 4.5 中部分回应了这一问题：结果表明，即使使用基于非LLM的特征，所提方法也有优势；(ii) 我们的评估主要基于由脚本化或半脚本化对话（如电视节目和表演式对话）构成的基准数据集，而模型在领域偏移至自然发生的真实世界对话环境时的表现仍有待研究；(iii) 众所周知，情感推断模型容易受到与人口统计因素、语言使用和标注主观性相关的数据集偏差影响，而本研究未进行任何偏差/公平性分析；(iv) 在本研究中，我们始终有意避免显式的说话人身份建模，但利用说话人元数据可能有助于改善人际动态的建模。

4.1.4 ERC 建模总结

在本节中，我们讨论了对话中的多模态情感识别（ERC）问题，其中情绪推断需要对多名说话者之间的对话上下文和多模态线索进行建模。针对这一任务，我们提出了两种技术：层次化建模和专家混合模型。总体而言，这些贡献通过强调模块化、对模态不平衡的鲁棒性以及灵活的多模态整合，推动了 ERC 建模的发展。然而，现有的 ERC 基准主要由表演式或半受控对话语音构成，这限制了其反映真实录音中所遇到的变化性的能力。

在下一节中，我们将超越对话场景，研究自然录音条件下的口语情感识别，此时，自发语音、环境噪声、说话人多样性以及高度偏斜的情感分布等挑战将成为核心问题。

4.2 Em自然条件下的情感识别

尽管语音情感识别已取得了长足进展，但现有大多数系统都是在受控或半受控条件下收集的数据集上开发和评估的，其中通常包含表演式情感和干净录音环境。相比之下，现实世界中的口语具有高度的可变性、自发性和噪声，情感表达在不同说话人之间往往细微且不均衡，录音条件也各不相同。这些因素使得自然语音中的情感识别相比实验室式基准更具挑战性。

情感识别　自然条件下的识别面临若干挑战，而这些挑战尚未得到　t

现有语音情感识别框架的充分应对。这些挑战包括说话人之间较大的差异、丰富多样的声学环境、自发语音模式，以及高度偏斜的情感分布，其中中性或表达微弱的情绪占主导。此外，在表演式数据集上训练的模型往往无法泛化到这类设置，导致其部署于现实应用时性能大幅下降。

为系统研究这类现实场景中的情感识别，本文采用 Interspeech 2025 自然条件下语音情感识别挑战赛作为研究基础，该挑战赛基于 MSP-PODCAST 语料库 [95]。该挑战赛提供了一个大规模、真实世界的基准，其中包含从多种播客来源收集并在真实条件下标注的自发语音。我们将这一设置用作严格的测试平台，以评估现代语音与语言模型在自然条件约束下进行情感识别的能力。

4.2.1 所提方法

下面介绍“Abhinaya”（在梵语中意为表达艺术）系统 [137], 的具体细节。该语音情感识别模型是作为此次挑战赛的一部分开发的。Abhinaya 系统通过集成仅语音、仅文本和语音-文本联合模型，直接应对挑战赛的关键问题——类别不平衡、多模态建模和数据变异性。我们没有从零开始构建这些组件，而是采用自监督学习模型（WavLM-Large [11]）和语音大语言模型（SLLMs）（SALMONN [115]）作为语音特征编码器。考虑到多模态信息的潜在优势，我们使用 Whisper-large-v3 自动语音识别（ASR）系统转录语音 [109] ，并使用大语言模型处理转录文本。此外，我们为多模态模型采用了一种方法：在输入端拼接语音和文本，并使用语音大语言模型对其进行联合处理。为缓解类别不平衡问题，我们探索了多种损失函数，用于微调大语言模型和语音大语言模型以实现语音情感识别。Abhinaya 系统的示意图 [1] 见图4.10。

4.2.1.1 纯语音模型

Wav-LM自监督学习（S1）：该子系统使用的语音基础模型是 WavLM-large 模型 [11]。卷积特征提取器被冻结，Transformer 层则经过微调。由于 WavLM 会生成帧级表示，我们采用注意力池化 [138] ，在进行情绪分类之前聚合这些特征。注意力池化层会输出特征的加权均值和加权标准差，其中权重通过注意力机制学习得到。

SALMONN 小型语言模型 (S2)：SALMONN 语音大语言模型 (SLLM) [115] 由 -

[1]代码：https://github.com/iiscleap/ABHINAYA

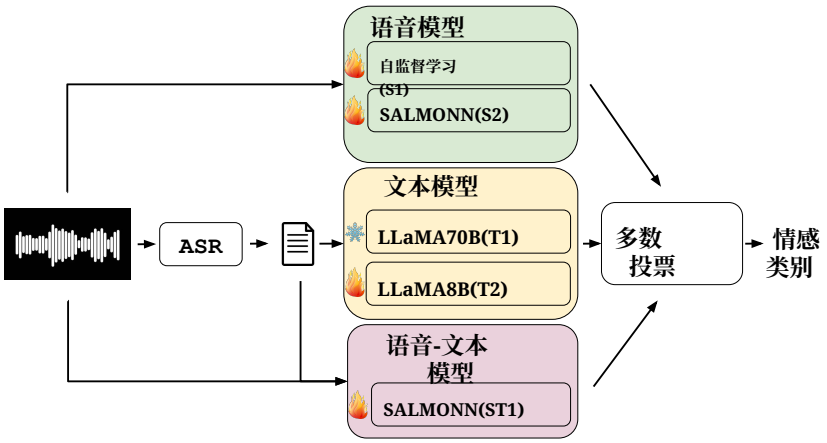


图4.10：Abhinaya 情感语音识别系统不同组件的示意图。我们使用三类模型——仅语音（S1、S2）、仅文本（T1、T2）和语音-文本（ST1）。只有T1用于零样本设置。模型使用的文本是由 Whisper 生成的自动语音识别转录文本

.

语音编码器（Whisper 编码器） [109] + BEATS 编码器 [139]，以及基于文本的大语言模型（LLaMA [140]）组成。语音编码器和 LLaMA 的权重保持冻结状态，而仅训练 Q-Former [128] 和 LLaMA 模块中的低秩适配器 (LoRA)，用于音频和语音任务 [115]。SALMONN 模型有两种规格，分别为 70亿参数和 13B，具体取决于后端文本大语言模型的规模。

在我们的设置中，我们使用 SALMONN 模型的所有层，但不包括最终的文本词元层。最终 LLaMA Transformer 层的表征经过注意力统计池化层，然后通过 softmax 分类头对模型进行微调，以完成情感分类任务。在微调过程中，我们只更新 Q-Former 和 LoRA 权重，以及分类头。

4.2.1.2 仅文本模型

在自然条件挑战赛中，参考文本转录仅提供给训练集和验证子集，而评估数据缺少真实转录文本。为确保一致性，我们在训练和测试期间使用 Whisper-Large v3 [109] 生成的自动语音识别转录文本进行模型开发。我们探索了两种仅文本模型的方案。**零样本 LLaMA（T1）：**在此任务中，我们在零样本设置下使用 `LLaMA-3.3-70B-Instruct` 模型 [141] 。模型接受提示后，将情绪分类为预定义类别。**微调较小型文本大语言模型（T2）：**先前研究 [142, 143] 表明，大语言模型在零样本场景下通常难以进行细粒度情感识别。为此，我们对其进行微调

该 LLaMA-3.1-8B 模型 [141]，使用 LoRA 进行情感识别。与 SALMONN 的微调设置类似，我们从 LLaMA 中提取最后一层表征，应用注意力统计池化，并将话语级嵌入输入情感分类头。

4.2.1.3 语音-文本联合模型（ST1）

由于语音大语言模型（SLLMs）主要是针对语音序列对大语言模型进行适配，我们探索其在语音文本联合建模中的潜力。为此，我们使用 SALMONN-70亿参数模型。除语音序列外，我们还附加了由 Whisper 生成的文本转录，然后使用 LoRA 对 SALMONN 模型进行微调。与仅使用语音或文本进行微调时一样，我们提取最后一层的表征，应用注意力统计池化，并训练分类头。

4.2.1.4 损失函数

该数据集类别极不平衡：训练数据中多数类（“中性”）的频次几乎是少数类（“恐惧”）的 26 倍。为解决这一问题，我们探索了多种损失函数：

- **加权交叉熵（WCE）：** 设数据点总数为 N，类别数为 C，分配给特定类别 $c \in [1, 2, \ldots, C]$ 的权重为 $w^c = \frac{N}{N_c \times C}$，其中训练数据中属于类别 c 的样本数为 N_c。样本 i 的交叉熵损失 l_i 定义如下：

$$l_i = -\sum_{c=1}^{C} t_{ic} w^c \log(p_{ic}) \quad ; \quad \mathbb{E} = \frac{1}{N} \sum_{i=1}^{N} l_i \tag{4.28}$$

  其中 p_{ic} 是样本 i属于类别 c的预测概率。对于样本 i的真实类别 c， t_{ic} 设为**1**，否则保持为**0**。损失 $\mathbb{E}$ 定义为加权交叉熵损失（加权交叉熵）。

- **加权焦点损失（WFL）：** 为了进一步强调难分类样本，我们探索焦点损失 [130], ，它根据模型的置信度调整权重。加权焦点损失（WFL）定义为：

$$l_i = -\sum_{c=1}^{C} t_{ic} w^c (1 - p_{ic})^{\gamma} \log(p_{ic}), \tag{4.29}$$

  且总损失 $\mathbb{E}$ 与公式4.28中的相同。在此， $\gamma \geq 0$ 是一个超参数。

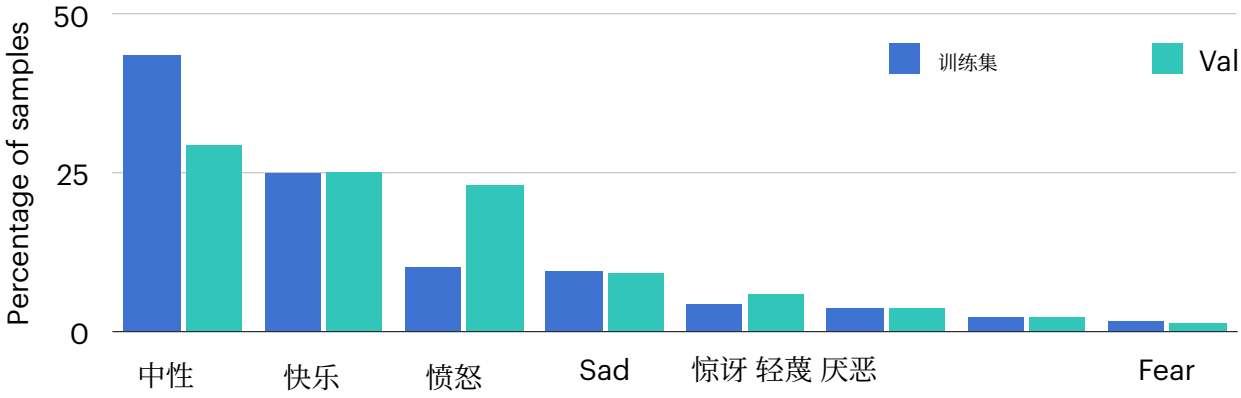


图4.1 1：训练集和验证集中 8 种情感类别的分布 a

.

• **向量缩放（VS）损失：**向量缩放（VS）损失使用依类别而定的温度缩放和偏差校正来调整预测。它对预Softmax模型预测 z_{ic} 进行如下修改：

$$\hat{z}_{ic} = (\frac{N_c}{N_{max}})^{\gamma} z_{ic} + \tau \log(\frac{N_c}{N}) \tag{4.30}$$

其中， $\gamma \geq 0$ 和 $\tau \geq 0$ 为超参数，而 N_{max} 为最高频类别中的样本数。加权损失中使用 Softmax $\hat{p}_{ic} = softmax(\hat{z}_{ic})$：

$$l_i = -\sum_{c=1}^{C} t_{ic} w^c \log(\hat{p}_{ic}) \tag{4.31}$$

其中， w^c 和 t_{ic} 的定义与前文相同。总损失（$\mathbb{E}$）与公式 4.28相同。

4.2.1.5 实验

数据集： MSP-PODCAST [95] 数据集是本次挑战的基础。训练集、验证集和测试集划分分别包含 84260、31961 和 3200 个语音文件。该数据集包含 8 个主要情绪类别的标注：“快乐”“愤怒”“悲伤”“中性”“惊讶”“恐惧”“轻蔑”和“厌恶”。此外，还包含另外两个标签——“其他”（表示预定义类别之外的情绪）和“X”（表示标注者未达成共识的情况），但这两类不纳入情感建模。图中展示了这 8 个类别在训练集和验证集中的分布。4.11。在模型选择时，我们构建了一个平衡验证集，其中 8 个情绪类别的话语数量相同，均为 326 条。挑战赛未提供测试数据标签，模型性能只能通过排行榜查看。与存在明显类别不平衡的训练分布不同，测试分布中的 8 个情绪类别是平衡的。

表 4.6：不同模型在平衡验证集和测试数据上的性能（%）。某些模型在测试数据上的性能未予报告，因为这些模型未提交至排行榜。 $*$ 表示组织者在官方竞赛截止日期之后评估的测试得分。

模型	模态	Loss	验证集 F1（宏平均）	测试集 F1（宏平均）
基线 [144]	语音	WCE	-	32.93
S1(WavLM-large-317M)	语音	WFL	34.43	33
S2 （SALMONN-13B）	语音	WFL	37.68	35.34
T1 (LLaMA-3.3-70B)	Text	-	32.78	-
T2 (LLaMA-3.1-8B)	Text	VS	33.68	-
ST1 (SALMONN-7B)	语音-文本	VS	35.43	-
ABHINAYA（第4名）	语音-文本 -		**43.01**	41.81
ABHINAYA* （最先进水平）	语音-文本 -		42.31	**44.02**

实现细节： 除采用零样本方式的 T1 外，所有模型均使用 AdamW 优化器进行微调 [119]，学习率为 $1e$-5。对于采用 LoRA（S2、T2 和 ST1）的模型，LoRA 参数 [127] 设为 $r = 8$ 和 $\alpha = 32$，丢弃率为 0。1。由于各模型的内存需求不同，其训练配置也有所差异：

- 对于 S1 和 T2，批量大小设为 8，并进行 4 步梯度累积。
- 对于基于 SALMONN 的模型（S2 和 ST1），批量大小设为 4，并进行 8 步梯度累积。

在纯语音模型（S1 和 S2）以及语音-文本模型（ST1）中，音频样本的处理时长上限为 10 秒。所有四个模型均训练 20 轮，并选取验证集宏平均 F1 分数最高的检查点进行评估。关于损失函数，使用 WFL 训练的模型（式 4.29）采用 $\gamma = 2$，而使用 VS 损失（式 4.30）训练的模型，其超参数为 $\gamma = 0.3$ 和 $\tau = 1$。

获得五个模型的输出后，对这五个分类器采用多数投票，并由于 S2 的验证性能更优，将其作为平局裁决者。

Abhinaya 的性能：表 4.6 展示了验证集和测试集的结果。主要发现总结如下：

- **微调后的 SALMONN-13B (S2) 实现了最佳的单系统性能**：请注意，其零样本验证性能仅为 18。63%，因此微调

显著提升了性能。

- **微调小模型与大型模型的零样本评估**：微调后的 LLaMA-3.1-8B（T2）模型优于零样本评估的 LLaMA-3.3-70B（T1）模型。

- **集成可提升性能**：五个模型的多数投票集成在测试集上的表现比最佳单模型（S2）相对提升了 24.56%。

- **显著优于挑战赛基线**：我们的系统相较于基线实现了 33.68% 的相对提升，展现了其在自然语境语音情感识别方面的有效性。

- **挑战赛后的评估证实其达到最先进性能**：在挑战赛提交截止时，ST1 的微调仅完成了 3 个训练轮次。我们继续对该模型进行微调，直至完成全部 20 个训练轮次；随后，组织者使用更新后的模型对 Abhinaya 系统进行了挑战赛后的评估。最终系统性能达到 44.02%，超过了挑战赛中已发表的最佳结果，从而在这项任务上取得了最先进水平的结果。

仅文本模型中不同大语言模型的评估：对于 T1 零样本设置，我们使用相同的提示结构，评估多个大语言模型（LLM）执行情感识别任务的表现。验证集上的性能对比见图4.12。主要观察结果如下：

- LLaMA-3.3-70B-Instruct 的表现最佳，优于此处考察的所有其他大语言模型。

- LLaMA-3.1-8B-Instruct 模型的零样本性能仅为 28.47%，但使用 VS 损失进行微调后，性能提升至 33.68%（表 4.6）。

- 其他“更大”的大语言模型在使用自动语音识别转录文本进行情感识别任务时，表现并不理想。

损失函数的选择：表4.7分析了不同损失函数的影响。主要观察结果如下：

- 加权焦点损失对纯语音模型（S1、S2）的提升最大，表明它能有效处理语音情感识别中的类别不平衡问题。相比之下，VS 损失对基于文本的模型（T2）和语音-文本模型（ST1）更有效。

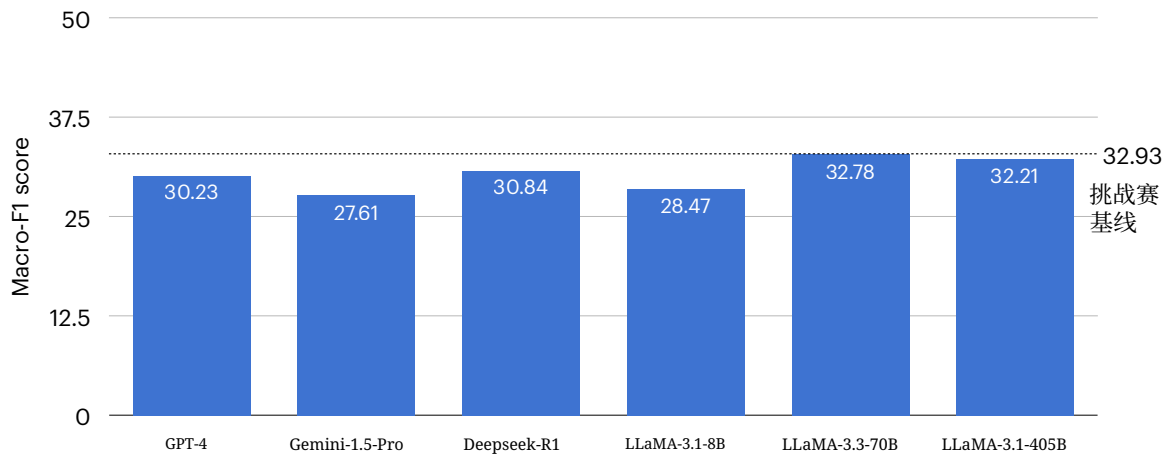


图4.12：不同大语言模型的验证宏F1分数（%） [141, 145, 8, 146] 在零样本设置下使用自动语音识别转录文本进行评估。所考虑的 LLaMA 模型为 Instruct 版本。图中也显示了基线性能。

表 4.7：采用不同损失函数的模型性能。所有结果均为平衡验证数据上的宏F1分数（%）。**粗体** 表示各模型的最佳损失函数。

模型	WCE	WFL	VS
S1	33.07	**34.43**	32.12
S2	36.34	**37.68**	33.17
T2	29.79	30.12	**33.68**
ST1	33.92	34.73	**35.43**

- T2 的零样本性能更强（28.47%，而 S2 为 18.63%），这表明情感类别之间的初始区分度更高。我们假设，这种初始区分有助于 VS 损失发挥作用，因为它会按类别调整逻辑值；相比之下，初始表征区分度较低的语音模型则能从样本级损失加权中获益更多。

语音-文本与纯语音 SLLM：为评估联合语音-文本微调的影响，我们使用语音输入对 SALMONN-7B 模型进行微调。该设置与模型 S2 相同，但将 SALMONN-13B 模型替换为更小的 7B 版本，以便与同时纳入语音和文本的 ST1 模型进行直接比较。仅语音模型的验证集宏平均 F1 达到 33.87%，而 ST1 达到 35.43%，表明多模态微调具有优势。然而，由于输入序列更长，语音文本模型会产生更高的计算成本。

表4.8：Abhinaya及不同模型组件在平衡验证数据上的按类别宏F1分数（%）。**粗体**表示每个情绪类别的最佳模型。

情感类别	S1	S2	T1	T2	ST1	ABHINAYA
愤怒	49.74	50.60	32.48	37.13	50.73	**51.49**
轻蔑	20.02	13.74	22.37	22.56	**25.62**	22.87
厌恶	18.91	31.53	33.33	**37.31**	32.08	36.13
Fear	12.22	16.79	**29.22**	27.74	29.09	26.41
快乐	55.93	**62.94**	40.00	43.33	50.88	62.83
中性	33.13	37.06	31.24	20.45	16.37	**40.29**
Sad	50.60	51.47	40.69	47.47	46.86	**55.62**
惊讶	34.94	37.31	32.93	33.48	31.83	**42.81**

分类别性能分析：表4.8 给出了 Abhinaya 及其组成部分各类别的宏平均F1分数。语音模型（S1、S2）在出现频率最低的类别上表现较差（见图 4.11）：“恐惧”、“轻蔑”、“厌恶”。相比之下，微调文本模型（T2）和语音文本模型（ST1）在这些类别上的表现有所提升。有趣的是，尽管整体性能较弱（表4.6），零样本文本模型（T1）在这三个稀有类别上仍优于基于语音的模型，甚至在出现频率最低的类别“恐惧”上优于集成模型。语音文本模型（ST1）在各类别上的表现通常介于语音模型和文本模型之间（“中性”除外），而集成模型在八个情感类别中的四个上取得了最佳分数。

与其他组合进行多数投票：我们使用多数投票评估模型组合（表4.9）。表现最佳的三模型组合——S2（语音）、T2（文本）和 ST1（语音-文本）——在验证集上取得了 40.36% 的宏平均 F1 分数，在测试集上取得了 39.74%（组合 I）。加入 S1 后，验证性能绝对提升 0.63 %（组合 III），而集成全部五个模型则使其绝对提升近 1.5%（ABHINAYA）。由于 T1、T2 和 ST1 在少数类上的表现更好，集成全部模型相比仅集成 S1、S2、T2 和 ST1，能提升这些代表性不足类别上的性能。移除 ST1 后，组合 IV 的性能绝对下降近 2%，这清楚地表明了 ST1 的影响。我们还评估了参数量最少的三个模型（S1、T2 和 ST1）的多数投票效果（组合 II），并使用 ST1 进行平局决胜。验证集 F1 分数降至 39.07%，凸显了语音情感识别任务需要更大型模型。

表 4.9：Abhinaya 中不同模型组合通过多数投票得到的结果（%）。

模型集成	S1	S2	T1	T2	ST1	验证集. F1（宏平均）	测试集 F1（宏平均）
组合 I	✗	✓	✗	✓	✓	40.36	39.74
组合 II	✓	✗	✗	✓	✓	39.07	–
组合 III	✓	✓	✗	✓	✓	40.99	–
组合 IV	✓	✓	✓	✓	✗	40.48	–
ABHINAYA	✓	✓	✓	✓	✓	42.31	44.02

4.2.2 自然条件下的语音情感识别总结

本节介绍 Abhinaya，这是一个为 Interspeech自然语音情感2025挑战赛开发的语音情感识别（SER）系统。我们的方法结合了五个利用大语言模型（LLM）和语音大语言模型（SLLM）的模型——两个基于语音的模型、两个基于文本的模型，以及一个多模态模型。为了解决类别不平衡问题，我们探索了多种损失函数，并展示了它们对模型性能的影响。我们的模型集成在该任务上取得了已发表成果中的最佳表现。实验分析突显了不同组件的重要性，以及适用于微调各个模型的损失函数。

4.3 章节总结

本章探讨了口语多模态情感识别面临的两项挑战：对话情感识别，以及自然录音条件下的情感识别。在 ERC 设置中，我们强调了对对话上下文和模态特定动态进行建模的重要性，并提出了层次式和模块化架构，使声学与语义线索能够有效融合，同时保持对模态不平衡的鲁棒性。

我们进一步将研究拓展至自然语音领域。在这一领域，自发情感表达、环境变异性和严重的类别不平衡给现有语音情感识别系统带来了重大挑战。我们利用大规模真实世界基准，在这些条件下评估了当代语音与语言模型，并分析了它们的局限性和优势。

本章的研究共同表明，稳健的情感识别需要对声学与语义信息进行有原则的建模。

第5章

语音到语音系统中的情感风格迁移

情感理解是构建具有情感智能的语音系统的必要条件，但并非充分条件。除了识别口语中的情绪之外，许多应用还需要能够明确控制语音中情绪的表达方式，同时保留底层语言内容和说话人身份。这种可控生成对于富表现力语音合成、个性化对话代理以及情感识别系统的数据增强至关重要。

在口语中，情感表达通过与语言结构紧密耦合的细微声学变化来实现。因此，生成富表现力语音需要将与情感相关的声学模式从语义内容和说话人特有特征中解耦。这一问题在语音到语音场景中尤为具有挑战性，因为没有文本监督可用，且同一内容在不同情绪下的配对录音也十分稀缺。

在本章中，我们研究情感风格迁移 （EST）问题，其目标是将参考语音信号中的情感说话风格迁移到源语音信号，同时保留源语音的语言内容和说话人身份。与表现性语音转换不同，EST要求情感的操控独立于说话人身份，因此它对声学与语义解耦提出了更为严格的检验。

我们探索一种无文本且非平行的设置，在训练过程中既没有转录文本，也没有平行的情感录音。通过将情感风格迁移构建为解耦问题，我们提出一个框架，能够在说话人和话语之间实现情感风格的零样本迁移。大量客观评估和主观评估表明

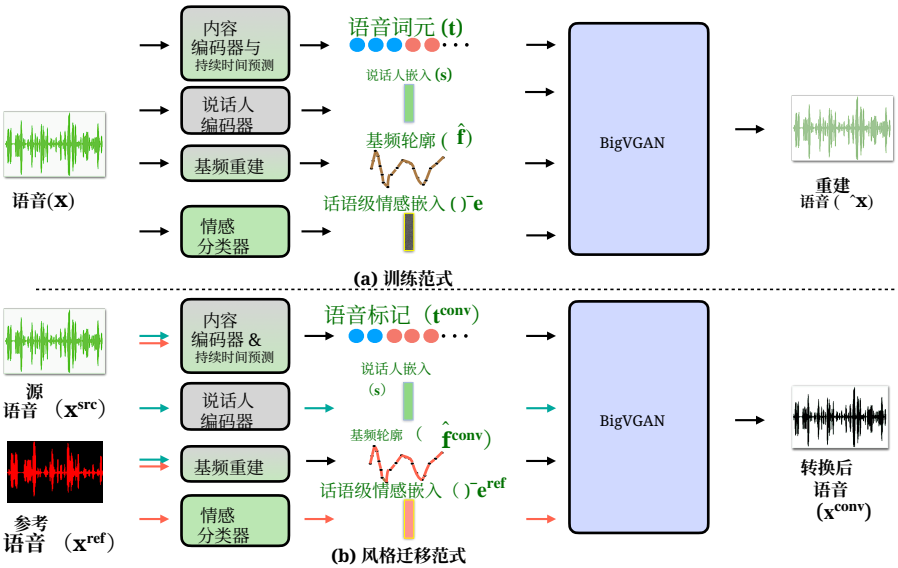


图 5.1： (a) S2S-ZEST 训练和 (b) 风格迁移范式的概览。情感风格因素以不同于其余部分的颜色标示。在风格迁移过程中，源语音标记会经过内容编码器和说话人编码器，而时长预测器、基频重建模块和情感分类器模块则接收来自参考语音的输入。

所提方法的有效性，并进一步展示如何将情感风格迁移用作数据增强策略，以提升下游情感识别性能 [147]。

5.1 所提方法

5.1.1 总体概览

将输入语音录音表示为 $\boldsymbol{x} = \{x_1, ..., x_N\}$，其中 N 表示录音中的样本数。经过加窗操作，每帧包含 N_w 个样本，我们得到 T 帧（对于不重叠帧，$T = N/N_w$）。使用内容编码器将加窗后的语音信号转换为词元序列（详见第 5.1.2 节）。分词器生成 $\{t_1, ...t_T\}$ 个离散标记，其中 $t_i \in \{1...V\}$，是聚类索引，V 是分词器的总词汇表大小。说话人信息由话语级嵌入 $\mathbf{s}$ 表示，该嵌入由预训练说话人编码器提取（第 5.1.3 节）。信号中的情感信息由情感分类器提取，该分类器提供帧级嵌入（$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..\mathbf{e}_T\}$）和话语级池化输出（$\bar{\mathbf{e}}$）。情感嵌入提取详见第 5.1.4.1 节。

使用 b　　基本组件（词元序列、说话人嵌入和情感嵌入），我们训练　　，

- 基频轮廓重建模块，用于重建帧级F0序列，$\hat{\mathbf{f}} = \{f_1, ...f_T\}$，其所属的 x，详见第 5.1.4.2节。

- 词元时长预测器，用于生成去重词元 $\mathbf{d} = d_1$，...，$d_{T'}$的时长 t_1，$..t_{T'}$。此处，T'表示去重后的词元数量[1]。详情见第 5.1.4.3节。

因此，编码器分析流程从语音信号中提取五个组成部分：i）去重词元，ii）词元时长，iii）说话人嵌入，iv）情感嵌入，以及 v）基频轮廓。以这些组成部分为输入，我们使用自编码损失训练 BigVGAN 模型，以重建语音信号（第 5.1.5节）。图 5.1 （a）展示了 S2S-ZEST 训练流程的简要概览。

风格迁移期间（如图 5.1 (b) 所示），生成并去重源语音标记以获得唯一词元，t_1，..，$t_{T'}$。时长预测器的输入为源语音$\hat{}$内容和参考语音风格信息，据此生成词元时长，$\mathbf{d}^{conv}$。说话人嵌入s从源语音中提取，而情感嵌入$\mathbf{E}^{ref}$则从参考语音中提取。基频轮廓$\hat{\mathbf{f}}^{conv}$由源语音内容、说话人嵌入以及参考情感嵌入构建。将这些表征输入合成器，以生成风格转换语音。情感风格迁移过程详见第 5.1.6 节。

5.1.2 内容编码器

为了编码不含说话风格或说话人信息的语音内容，Polyak 先前的研究 [148] 展示了离散化自监督语音嵌入的优势。另一方面，Niekerk 等人[149]通过继续对 HuBERT 主干网络进行预训练，改进了 HuBERT-base 模型，使其能够预测帧级软嵌入。然而，连续的 soft-HuBERT 特征难以控制语速。我们使用预训练的 soft-HuBERT 嵌入 [149]训练 k 均值聚类模型，以生成语音离散标记。这样既保留了 soft-HuBERT 在可懂度方面的优势，又能在合成过程中控制语音时长。将内容编码器的输出记为$\mathbf{t} = \{t_1, t_2, \ldots, t_T\}$，则有$\mathbf{t} = \texttt{k-means}(\mathbf{H})$，其中 soft-HuBERT 模型嵌入记为$\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_T\}$，即$\mathbf{H} = \texttt{soft-HuBERT}(\boldsymbol{x})$

[1]例如，如果词元为 $\{1, 1, 1, 41, 41, 1, 1, 5, 5, 5, 5\}$，则去重后的词元序列为 $\{1, 41, 1, 5\}$，其持续时间为 $\{3, 2, 2, 5\}$。

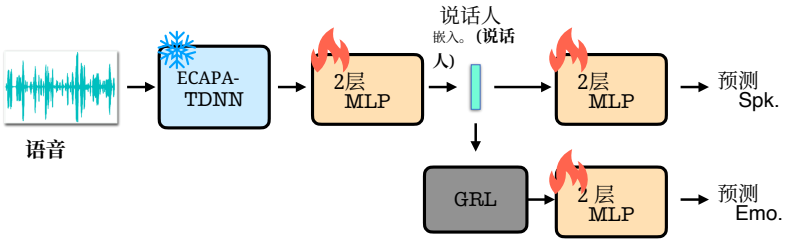


图 5.2：用于提取说话人嵌入的模型。GRL 是梯度反转层的缩写。蓝色块在训练期间保持冻结。

5.1.3 说话人编码器

为了提取说话人嵌入，我们采用预训练的说话人验证模型。具体而言，我们使用 ECAPA-TDNN [150], ，它通过对帧级表示进行话语级池化来生成 x-vector 嵌入。先前的研究 [151, 152] 表明，x向量也会编码情感信息，从而干扰风格迁移。受 Li 等人启发，为缓解这一问题[153], ，我们在 x-vector 模型中加入两个全连接层，并应用情感对抗损失 [154]。S2S-ZEST 中的说话人嵌入采用以下损失函数进行训练：

$$\mathrm{L}_{tot-spk} = \mathrm{L}_{ce}^{spk} - \lambda_{adv}^{emo}\mathrm{L}_{ce}^{emo} \tag{5.1}$$

其中 L_{ce}^{spk}是说话人分类损失，L_{ce}^{emo}是情感分类损失。这一点如图 5.2所示。

5.1.4 情感风格因素

5.1.4.1 情感分类器

我们使用预训练的 HuBERT-base模型构建情感分类器，该模型由卷积特征提取器和其后的 12 个 Transformer层组成。为了进行情绪分类，我们对 Transformer层进行微调。情感嵌入是最后一个 Transformer 层的输出，记为 $\mathbf{E} = \texttt{Emo-embed(x)}$，其中 $\mathbf{E} = \{\mathbf{e}_1, ...\mathbf{e}_T\}, \mathbf{e}_i \in \mathrm{R}^D$，其中 $D = 768$。帧级嵌入矩阵 $\mathbf{E}$ 沿时间维度进行平均池化，以获得话语级情感嵌入 $\bar{\mathbf{e}}$。我们在 $\bar{\mathbf{e}}$ 上训练一个 Softmax分类头，以预测情绪类别。我们还要求情感嵌入与说话人属性解耦。为此，我们采用说话人对抗损失（类似于说话人

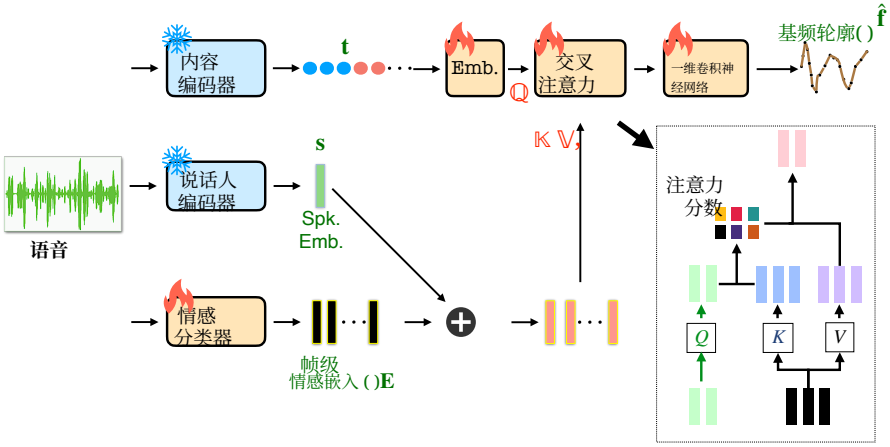


图 5.3：基频轮廓重建模块——说话人嵌入（s）与帧级情感嵌入 E 相加，形成键值序列，而源语音词元嵌入（C）则形成查询序列。交叉注意力模块的帧级输出经过使用一维卷积神经网络的逐位置前馈网络处理，以重建ˆ基频轮廓（f）。图中还展开了交叉注意力模块的架构以供参考。

嵌入提取器）。损失函数表示为：

$$\mathrm{L}_{tot-emo} = \mathrm{L}_{ce}^{emo} - \lambda_{adv}^{spk}\mathrm{L}_{ce}^{spk} \tag{5.2}$$

其中，L_{ce}^{emo}和 L_{ce}^{spk}分别是用于情感分类和说话人分类的交叉熵损失函数。

5.1.4.2 基频轮廓重建

我们提出了一种分解式方法，用于重建受内容、说话人和情感信息影响的基频轮廓。可学习嵌入层根据词元序列提供内容表示 C t。由于局部音高变化会影响情感表达的丰富程度，[155, 24], 我们使用帧级情感嵌入 E 以及内容表示 C 和说话人嵌入 s。基于交叉注意力和一维 CNN 网络的基频重建模块可表示为：

$$\hat{\mathbf{f}} = \texttt{1D-CNN}(\texttt{Attn}(\mathbf{C},\ \mathbf{s}+\mathbf{E},\ \mathbf{s}+\mathbf{E})) \tag{5.3}$$

在上述方程中，查询表示Q来自内容嵌入，而说话人嵌入s 与帧级情感嵌入E 相加，共同构成键（K）和值（V）表征。在所提出的表述中，内容嵌入

在交叉注意力模型中用作查询，而情感嵌入和说话人嵌入分别用作键和值。该设计使模型能够根据目标情感的韵律模式和源说话人身份来调节基频生成，同时将时间对齐锚定到语音源内容。目标基频轮廓$\mathbf{f} = \{f_1, f_2, f_3, \ldots, f_T\}$使用YAAPT算法提取 [156],，基频重建模块的损失定义为:

$$\mathrm{L}^{f0} = ||\mathbf{f} - \hat{\mathbf{f}}||_{L_1} \tag{5.4}$$

YAAPT算法识别出的清音帧赋值为零，并纳入损失计算。这样，我们便可以根据一段语音信号中的内容和说话人，以及另一段语音信号中的情感，推导出基频轮廓。音高重建的框图如图 5.3所示。

5.1.4.3 时长预测

自监督学习（SSL）特征$\mathbf{H}$的离散化通常会导致词元重复 [157, 158, 159, 160]。为解决这一问题，我们对内容编码器输出的词元进行去重，$\mathbf{t}' = \texttt{de-dup}(\mathbf{t})$。在风格迁移过程中，我们需要根据说话人身份和参考语音的情感 [158, 157]，确定每个去重后源词元的时长（语速）。为实现这一目标，我们训练时长预测器模块 $\mathtt{D}_{\mathtt{pred}}$ 如下:

$$\hat{\mathbf{d}} = \mathtt{D}_{\mathtt{pred}}(\mathbf{t}', \mathbf{s}, \bar{\mathbf{e}}) \tag{5.5}$$

该 $\mathtt{D}_{\mathtt{pred}}$ 模块由一个可学习的词元到嵌入层组成（输入为 $\mathbf{t}'$）。词元嵌入与话语级说话人嵌入$\mathbf{s}$和情感嵌入$\bar{\mathbf{e}}$拼接后，输入 $1-D$卷积神经网络层，以预测去重词元的时长。用 $\hat{}$ $\hat{}\,\hat{}$ $\hat{}\,\hat{}$ 分别表示目标时长和预测时长 $\mathbf{d} = \{d_1, d_2, \ldots, d_{T'}\}$ 与 $\mathbf{d} = \{d_1, d_2, d_3, \ldots, d_{T'}\}$，时长预测器的损失定义如下:

$$\mathrm{L}^{dur}_{mse} = \frac{1}{T'}||\mathbf{d} - \hat{\mathbf{d}}||^2_{L_2} \tag{5.6}$$

5.1.4.4 训练范式

时长预测器、情感分类器和基频轮廓重建模块共同进行训练。总损失为:

$$\mathrm{L}_{all} = \lambda_e \mathrm{L}_{tot-emo} + \lambda_f \mathrm{L}^{f0} + \lambda_d \mathrm{L}^{dur}_{mse} \tag{5.7}$$

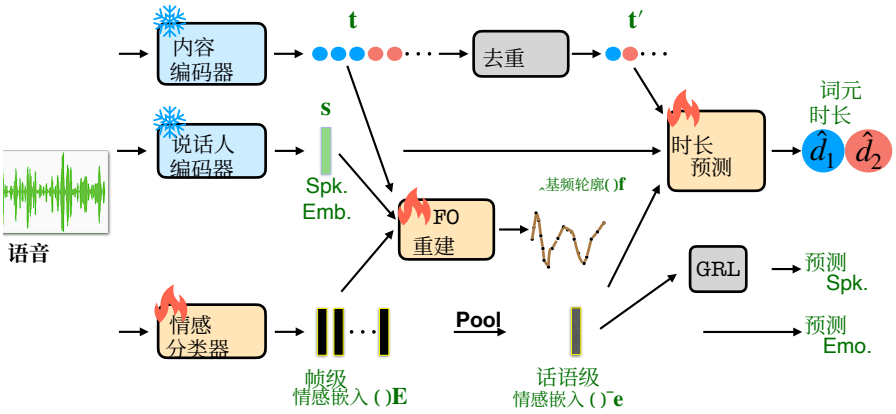


图 5.4：分析阶段从语音中提取的不同因素。情感分类器采用说话人对抗损失进行训练。帧级嵌入（$\mathbf{E}$）、说话人嵌入（$\mathbf{s}$）和语音标记（$\mathbf{t}$）用于重建基频轮廓 $\hat{}\mathbf{f}$。此外，话语级情感嵌入（$\bar{}\mathbf{e}$）与去重词元 $\mathbf{t}' = \{t_1, ..., t_{T'}\}$ 一同用于预测每个词元的时长（$\hat{}\mathbf{d}$）。所有蓝色块均保持冻结状态，黄色块则进行训练。灰色块不包含任何可学习参数。

其中，λ_e、λ_f 和 λ_d 为加权系数。
这些组件的提取过程如图5.4所示。

5.1.5 语音重建

词元序列$\mathbf{t}$通过可学习嵌入层进行向量化。另一方面，语音信号的重建基$\hat{}$频轮廓$\mathbf{f}$ 通过可学习的 CNN-BiLSTM 网络进行向量化。随后，将这些表示与向量化后的词元序列、说话人嵌入$\mathbf{s}$以及话语级情感嵌入 $\bar{}\mathbf{e}$ 拼接后输入 BigVGAN [161] 模型进行语音合成。

BigVGAN 网络被提出作为一种通用声码器，可生成高保真原始语音波形 [161]。研究还表明，该模型无需微调，就能很好地泛化到各种分布外场景。BigVGAN 模型使用带有残差连接的卷积层，将输入序列N_w 倍上采样。该网络称为生成器，它采用 $Snake$ 激活函数（$f_\alpha(x) = x + \frac{1}{\alpha}\sin^2(\alpha x)$），以在语音信号中引入周期性 [161]。判别器网络包含两种类型：1）多周期判别器（MPD），通过关注语音的周期结构，将真实语音样本与生成的语音样本区分开；2）多分辨率判别器（MRD），其中多个判别器通过处理不同分辨率的频谱图进行训练，以区分真实语音样本与生成的语音样本。我们使用文中提到的相同损失函数和超参数

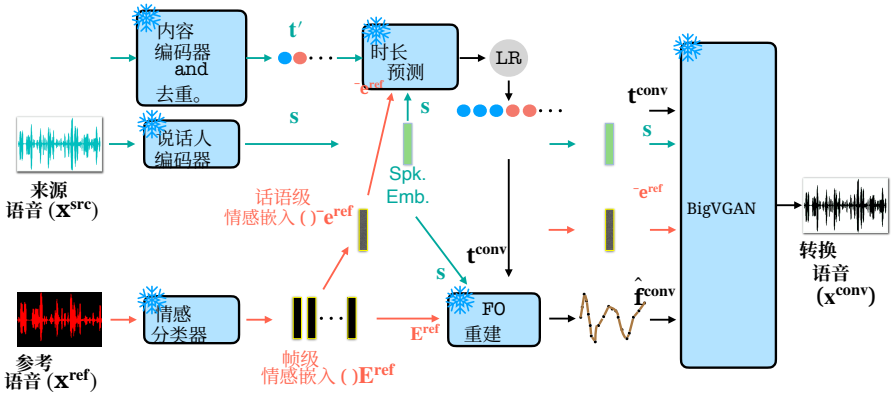


图 5.5：情感风格迁移——从参考语音中提取帧级（$\mathbf{E}^{ref}$）和话语级（$\bar{\mathbf{e}}^{ref}$）嵌入。使用源词元 $\mathbf{t}'$、说话人向量 $\mathbf{s}$ 和情感嵌入 $\bar{\mathbf{e}}^{ref}$进行时长预测。利用这些预测时长$\hat{\mathbf{d}}^{conv}$ 生成重复词元序列$\mathbf{t}^{conv}$。基于该词元序列、$\mathbf{E}^{ref}\hat{}$和说话人嵌入 $\mathbf{s}$，预测 F_0 轮廓$\mathbf{f}^{conv}$。最后，将词元序列、说话人嵌入和情感嵌入，以及预测的基频轮廓输入 BigVGAN 模型，生成转换语音。

在 Lee 等人 [161]中。因此，重建的语音信号 $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \hat{x}_3, ..., \hat{x}_N\}$ 由 BigVGAN 得出，形式如下：$\hat{\mathbf{x}} = \texttt{BigVGAN}(\mathbf{t}, \mathbf{s}\ \bar{\mathbf{e}}, \hat{\mathbf{f}})$

5.1.6 情感风格迁移

情感风格迁移包括以下步骤：

- **情感嵌入**：从参考语音$\mathbf{x_{ref}}$中提取帧级和话语级情感嵌入，分别记为$\mathbf{E_{ref}}$ 和 $\bar{\mathbf{e}}_{\mathbf{ref}}$。

- **词元序列**：我们从源语音中提取去重后的词元序列$\mathbf{t}'$ 以及说话人嵌入$\mathbf{s}$ 。接下来，我们使用情感嵌入 $\bar{\mathbf{e}}_{\mathbf{ref}}$ （类似于公式5.5）来预测每个去重后词元的时长：

$$\hat{\mathbf{d}}^{conv} = \texttt{D}_{\texttt{pred}}(\mathbf{t}', \mathbf{s}, \bar{\mathbf{e}}^{\mathbf{ref}}) \tag{5.8}$$

序列$\mathbf{t}'$ 中的词元使用$\hat{\mathbf{d}}^{conv}$进行重复，

$$\mathbf{t}^{conv} = \texttt{dup}(\mathbf{t}', \hat{\mathbf{d}}^{conv}). \tag{5.9}$$

- **基频 Cont　我们的**：词元序列$\mathbf{t}^{conv}$ 被向量化以形成$\mathbf{C}^{conv}$ ，并与　h

嵌入$\mathbf{s}$，$\mathbf{E}^{\mathbf{ref}}$，以预测基频轮廓（类似于式 5.3），如下：

$$\hat{\mathbf{f}}^{\mathbf{conv}} = \texttt{Attn}(\mathbf{C}^{conv},\ \mathbf{s}+\mathbf{E}^{\mathbf{ref}},\ \mathbf{s}+\mathbf{E}^{\mathbf{ref}}) \tag{5.10}$$

- **语音合成**：最后，转换语音为，

$$\mathbf{x}^{\mathbf{conv}} = \texttt{BigVGAN}(\mathbf{t}^{\mathbf{conv}}, \mathbf{s}, \bar{\mathbf{e}}^{\mathbf{ref}}, \hat{\mathbf{f}}^{\mathbf{conv}}) \tag{5.11}$$

这些步骤在图 5.5 中标出。

5.2 实验与结果

5.2.1 数据集与预训练

对于内容编码器，我们使用 Niekerk 等人提出的 soft-HuBERT 基础模型 [149]，，该模型在 960 小时的 LibriSpeech 数据集上预训练 [162]。我们在随机选取的 10% 训练数据集上训练一个 k 均值聚类模型，由此构成语音分词器。

情感嵌入提取自在情感语音数据库（ESD） [105]上训练的情感分类器模型。ESD 数据库包含由 10 名英语母语说话人和 10 名汉语母语说话人说出的 350 条平行语句，涵盖 5 种情感类别（中性、快乐、愤怒、悲伤和惊讶）。我们仅使用 ESD 数据集的英语训练子集来构建情感分类器模型。我们遵循该数据集预定义的训练集-验证集-测试集划分，每种情感下每位说话人使用 300 条话语进行训练。由此共得到 15000 条训练语句，并以 2500 条未见语句（每种情感下每位说话人 50 条）用于验证。

说话人编码器采用 ECAPA-TDNN 模型初始化 [150]，，该模型在 VoxCeleb 数据集中的 2794 小时数据和 7363 位说话人上进行了预训练。我们使用情感对抗损失（公式 5.1 ）在 ESD 数据集上微调说话人编码器模型。

除非另有说明，所有模块均在 ESD 的性别均衡的英语训练分区上进行训练。为分析训练数据多样性对泛化的影响，我们还使用原始 ESD 数据以及额外增加的 100 小时 Librispeech 数据训练了一个变体（记为 S2S-ZEST-diverse）。对于该变体，语音分词器、说话人编码器和 BigVGAN 模型均在增强数据集上进行训练。该变体仅用于研究说话人多样性增加时的规模扩展行为。

表 5.1：情感风格迁移的各种评估设置。

设置	描述	源数据集	参考数据集	匹配说话人	匹配文本	已见情感	已见说话人	样本数
SSST	相同说话人、相同文本	ESD	ESD	✓	✓	✓	✓	1200
SSDT	相同说话人、不同文本	ESD	ESD	✓	✗	✓	✓	1160
DSST	不同说话人相同文本	ESD	ESD	✗	✓	✓	✓	10800
DSDT	不同说话人不同文本	ESD	ESD	✗	✗	✓	✓	10440
UTE	未见目标情感	ESD	CREMA-D	✗	✗	✗	✓	1000
USS	未见源说话人	TIMIT	ESD	✗	✗	✓	✗	800
USUE	未见说话人未见情感TIMIT		CREMA-D	✗	✗	✗	✗	2000

5.2.2 实现

内容编码器使用 soft-HuBERT 表征，并通过 k均值聚类算法进行量化，聚类数量为 $K = 100$ 个 [148]。说话人编码器使用 $\lambda_{adv}^{emo} = 10$ （公式 5.1）进行训练，共训练 10 个训练轮次，批量大小为 32。情感分类器使用 $\lambda_{adv}^{spk} = 1$（公式 5.2）进行训练，并通过公式 5.7中的总损失函数，与音高重建网络和时长预测器进行联合优化。加权系数设为 $\lambda_e = 1000$、$\lambda_f = 1$和 $\lambda_d = 10$，其取值根据验证损失确定。该联合训练过程进行 200 个训练轮次，批量大小为 32，学习率为 $1e-4$。推理过程中，词元的预测时长被限制在原始源词元时长的 ± 40 % 范围内。该约束作为后处理步骤应用，以防止可能导致不自然节奏或不稳定音高重建的极端时长偏差。我们观察到，如果不施加此约束，偶发的异常预测可能导致片段过度拉伸或压缩。 ± 40% 的边界是通过经验确定的，以平衡表现力变化与语音自然度。音高重建模块中的交叉注意力模型具有 4 个注意力头和 256 的隐藏维度，而时长预测器由一个卷积核大小为 3 、隐藏维度为 256 的一维卷积神经网络组成。最后，语音重建模块 BigVGAN [161]，使用批量大小为 16、学习率为$1e-4$进行训练。所有模型均使用 AdamW 优化器 [119]进行优化。对于 S2S-ZEST-diverse 变体，采用相同的架构配置和超参数，并在第 5.2.1节所述的增强数据集上进行训练。代码和样本已开源。[1]。

5.2.3 评估设置

评估设置大致分为两种，

- **已见**：源语音包含来自 ESD 测试集的未见内容，但说话人和情绪存在于训练数据中。

[1] https://github.com/iiscleap/A2A-ZEST

- **未见**：源说话人或参考语音情感，或两者，均未在训练期间出现。

不同的测试评估设置概述于表5.1。对于源语音和参考语音均取自 ESD 的四种评估设置 [105]，源语音始终为中性，参考语音属于四种情感类别之一：“快乐”、“愤怒”、“悲伤”和“惊讶”。在 SSDT 和 DSDT 设置中，我们从 ESD 中随机选取 10 条中性话语（每位说话人一条）作为源语音。

对于未见评估设置，未见情绪使用 CREMA-D，未见源说话人使用 TIMIT。在未见目标情绪（UTE）设置中，我们从 CREMA-D 的“恐惧”和“厌恶”类别中各选取 50 条话语作为参考语音，并与来自 ESD 的 10 条中性话语（每位说话人一条）配对，从而得到 1000 条评估话语。在未见源说话人（USS）设置中，将来自 TIMIT 的 100 条话语用作源语音，每条分别与来自 ESD 的 8 条情绪话语（每种情绪 2 条，不含中性）配对，共得到 800 条评估话语。在未见说话人未见情绪（USUE）设置中，我们将同样的 100 条 TIMIT 源话语与 CREMA-D 中“恐惧”和“厌恶”类别各 10 条话语作为参考语音进行配对，共计 2000 条评估话语。

5.2.4 客观评估指标

- **相对于参考语音的情感迁移**：我们使用 emotion2vec [25]提取参考语音和转换语音的嵌入，并计算二者的平均余弦相似度（记为情感相似度）。

- **相对于参考语音的节奏迁移**：我们使用预训练的自动语音识别模型 [49]，将参考语音和转换语音信号与各自对应的文本转录进行强制对齐。
  随后，计算参考语音和转换语音的词语语速。采用参考语速与转换语速之间的皮尔逊相关系数（PCC）作为节奏迁移的衡量指标。这与 Barrault 等人 [164]提出的指标类似。

- **相对于源语音的内容保留**：我们使用自动语音识别（ASR）系统转录转换语音。具体而言，我们采用预训练的 Whisper-large-v3模型 [109] 进行语音识别。使用源语音的真实转录文本计算词错误率（WER）。

- **相对于源语音的说话人保持**：我们报告说话人嵌入的平均余弦相似度

表 5.2：客观评估结果。此处，WER 表示词错误率，情感相似度表示根据 emotion2vec 嵌入计算的情感相似度，Spk.-Sim. 表示根据 Resemblyzer 嵌入计算的说话人相似度。词级皮尔逊相关系数指词语速率皮尔逊相关系数。不同的测试设置见表 5.1。 $*$ 表示这些条目未**以粗体或下划线标示**，因为 StarGANv2-EST 不适用于 7 种测试设置中的 2 种。

类别	方法	指标	SSST	SSDT	DSST	DSDT	UTE	USS	USUE	Avg
情感迁移	StarGANv2-EST [163]	情感相似度 (↑)	0.39	0.37	0.37	0.37	-	0.25	-	0.35
	VEVO [91]		−0.01	−0.03	−0.01	−0.04	−0.01	0.37	−0.02	0.04
	ZEST [22]		0.59	0.62	0.52	0.51	0.53	0.48	0.41	0.52
	S2S-ZEST		**0.69**	**0.71**	**0.58**	**0.56**	**0.64**	**0.59**	**0.51**	**0.61**
	S2S-ZEST-diverse		0.67	0.70	0.55	0.52	0.59	0.57	0.50	0.59
节奏迁移	StarGANv2-EST [163]	词级皮尔逊相关系数 (↑)	0.54	0.03	0.43	0.02	-	0.02	-	0.21
	VEVO [91]		0.65	0.10	**0.59**	**0.16**	0.01	0.08	0.01	0.23
	ZEST [22]		0.53	0.04	0.44	4×10^{-4}	5×10^{-3}	0.02	2×10^{-4}	0.15
	S2S-ZEST		**0.68**	**0.18**	0.58	0.14	**0.07**	**0.11**	**0.06**	**0.26**
	S2S-ZEST-diverse		**0.68**	0.16	0.58	0.12	0.05	0.09	0.04	0.25
内容保留	StarGANv2-EST [163]	词错误率 (↓)	5.72	8.28	6.61	6.02	-	**9.41**	-	7.21*
	VEVO [91]		4.55	7.65	4.72	7.68	6.16	14.36	14.06	8.45
	ZEST [22]		5.11	5.32	5.19	6.11	5.92	16.57	16.42	8.66
	S2S-ZEST		5.58	4.94	5.55	6.48	6.03	13.53	12.43	7.79
	S2S-ZEST-diverse		**3.65**	**3.15**	**3.51**	**4.82**	**5.02**	11.63	**10.84**	**6.09**
说话人保持	StarGANv2-EST [163]	说话人相似度(↑)	0.76	0.76	0.67	0.65	-	0.65	-	0.70
	VEVO [91]		**0.86**	**0.87**	**0.86**	**0.86**	**0.85**	**0.87**	**0.86**	**0.86**
	ZEST [22]		0.74	0.75	0.74	0.72	0.73	0.54	0.53	0.68
	S2S-ZEST		0.74	0.73	0.74	0.72	0.73	0.53	0.53	0.68
	S2S-ZEST-diverse		0.75	0.74	0.76	0.74	0.76	0.61	0.59	0.71

源语音与转换语音之间的嵌入。为获得这些说话人嵌入，我们使用 Resemblyzer 工具[1]。该指标记为说话人相似度。

5.2.5 与基线系统的比较

考虑了三个基线系统：

- **StarGANv2-EST**：Li 等人 [163] 提出了一种用于语音转换的 StarGAN-v2 网络。该网络在判别器之外还训练了一个辅助分类器来识别源说话人，从而实现说话人转换。在我们的实验中，我们对这一架构进行了修改，以便

[1] https://github.com/resemble-ai/Resemblyzer

通过使用辅助网络来识别情绪，而非说话人，实现情感迁移。我们将这个修改后的模型称为 StarGANv2-EST。该模型使用与 S2S-ZEST 相同的数据集进行训练。请注意，此基线无法处理训练期间未见的情绪，因此未在 UTE 和 USUE 测试设置中进行评估。

- **VEVO**：我们采用 Zhang 等人提出的预训练 VEVO 模型。 [91]由于该模型的训练语料库远大于 S2S-ZEST 所用的语料库，因此我们不会针对测试设置对其进行微调。相反，我们使用 VEVO 的风格模仿流程进行比较。

- **ZEST**：我们还与先前的工作进行比较——ZEST [22]。

5.2.6 客观结果

这些客观测试的结果见表 5.2。以下是从这些评估中得出的结论。

- **情感迁移**：S2S-ZEST 在已见和未见设置下均取得了最高的情感相似度。对于已见说话人和情绪，在 SSST 设置下相似度为 0。69，在 SSDT 设置下为 0。71，优于所有基线系统。在未见目标情感（UTE）场景中，S2S-ZEST 的相似度为 0。64，表明它能够有效迁移训练期间未观察到的情绪。对于未见源说话人（USS），该模型的相似度为 0。59，展现了对新说话人的稳健泛化。在最具挑战性的未见说话人且未见情绪（USUE）设置下，S2S-ZEST 达到 0。51，此时说话人和情绪均未见过。相比之下，StarGANv2-EST 和其他先前工作取得的相似度较低，表明它们有效进行情感风格迁移的能力有限。
S2S-ZEST-diverse 的平均情感相似度相近（0。59 对比 0。61）。

- **节奏迁移**：在同文本设置下，S2S-ZEST 在 SSST 中的相关性达到 0.68，在 DSST 中达到 0.58，表现出较强的节奏迁移能力。在不同文本设置下，即 SSDT 和 DSDT 中，相关性分别降至 0.18 和 0.14，反映出文本差异导致对齐度较低。在 USS、UTE 和 USUE 设置下，相关性进一步降至 0.11、0.07 和 0.06。然而，综合所有设置来看，S2S-ZEST 的平均表现（0.26）优于 ZEST（0.15）和 StarGANv2-EST（0.21），且与 VEVO（0.23）的表现相近。S2S-ZEST-diverse 变体的节奏迁移表现也相近。

- **内容保留**：在源语音来自 ESD 的测试设置中（即除 USS 和 USUE 外的所有设置），S2S-ZEST 的词错误率保持稳定。尽管

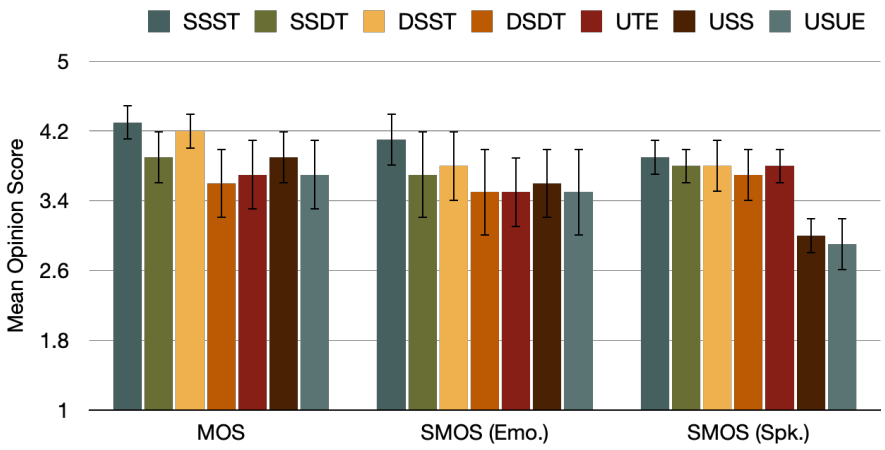


图 5.6：不同测试设置下的主观评估。使用的缩写：MOS——平均意见分，SMOS——相似度平均意见分。图中还显示了 95% 置信区间。

VEVO 使用 6 万小时的语音进行训练，S2S-ZEST 在适用的七种测试设置中有五种（SSDT、DSDT、UTE、USS、USUE）取得了更低的词错误率。虽然 StarGANv2-EST 在部分单项设置中（DSDT 和 USS）取得了最低词错误率，但它不适用于 UTE 和 USUE，因此整体评估受到限制。总体而言，S2S-ZEST 的平均词错误率最低（7.79），表明其在进行情感风格迁移的同时，能够有效保留内容。S2S-ZEST-diverse 变体显著提升了所有测试设置下的内容保留能力，将平均词错误率从 7.79 降至 6.09。值得注意的是，该变体在未见说话人条件下也取得了改善（USS：$13.53 \rightarrow 11.63$，USUE：$12.43 \rightarrow 10.84$），表明其语言稳定性有所增强。

- **说话人身份保留**：在所有测试设置中，VEVO始终取得最高的说话人相似度，范围为 0.85 至 0.87，展现出其保留说话人身份的强大能力。使用额外的说话人多样性进行训练后，S2S-ZEST- diverse 在所有设置中的说话人相似度均有所提升，尤其是在未见过的说话人场景中（USS：$0.53 \rightarrow 0.61$, USUE：$0.53 \rightarrow 0.59$）。这些结果表明，未见过的条件下的说话人保持会受到训练数据多样性的影响。使用更多说话人和多样化的情感语音进行训练，将是我们未来工作的方向之一。

5.2.7 主观测试

5.2.7.1 S2S-ZEST-diverse 的评估

我们使用 Prolific[1] 平台开展听力测试。我们招募了 30 名参与者进行主观评估。我们选取了 44 段录音，其中 4 种测试设置（SSST、SSDT、DSST、DSDT）各选取 8 段录音，UTE、USS 和 USUE 设置各选取 4 段录音。录音以随机顺序播放。我们还向参与者提供了训练示例，以阐明测试的目的。

调查中的所有参与者均被要求根据三项标准对语音文件给出意见分数（范围为 1-5）：i）转换信号与参考信号之间的情感相似度，ii）转换语音的质量，以及 iii）转换信号与源信号之间的说话人相似度。主观评估结果（以平均意见分（MOS）表示）报告于图 5.6。主观测试的主要观察结果如下：

- 在 SSST 测试设置下，S2S-ZEST-diverse 在全部三项标准上的表现均为最佳。这是意料之中的，因为当参考与源具有相同的内容和说话人时，将情感说话风格从参考迁移到源，是所有测试设置中最容易的。

- 当源与参考具有相同内容但说话人不同（DSST）时，主观分数表明，与 SSDT 设置（相同说话人、不同内容）相比，情感迁移更为容易。

- UTE 和 USUE 设置下的 SMOS（情绪）分数与 DSDT 设置相当，这表明，与在不同说话人和内容之间迁移已见情感相比，泛化到未见情绪并没有更大的挑战。

- 与已见说话人条件相比，USS 和 USUE 设置的说话人相似度分数较低，这反映出对于训练期间未见过的说话人而言，保留说话人身份本身就较为困难。多样化变体朝这个方向迈进了一步；使用更多说话人数据训练编码器和声码器，将有助于进一步提升说话人泛化能力。

5.2.7.2 S2S-ZEST-diverse 与基线系统的比较

我们从7种测试设置中的每一种选取4个转换文件，并对三种方法中的每一种（StarGANv2-EST在UTE和USUE测试设置下没有任何输出）加以考虑，并要求

[1] https://www.prolific.co

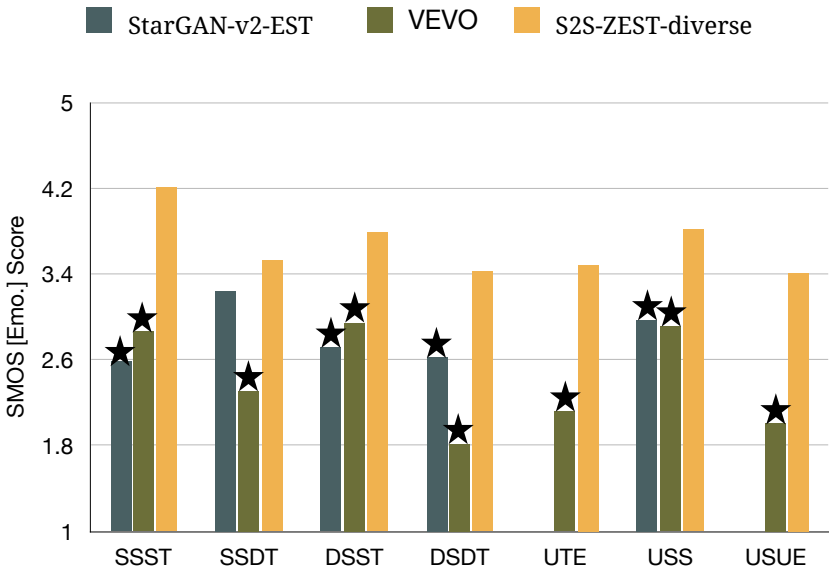


图5.7：不同测试设置下的主观评估。SMOS 表示相似度平均意见分。 ★ 表示基线与 S2S-ZEST 的得分差异具有统计显著性（$p < 0.05$）

30名参与者对参考语音与转换语音之间的情感说话风格相似度进行评分，评分范围为1到5。该主观测试的结果如图5.7。

结果表明，在所有测试设置下，S2S-ZEST-diverse 的表现都显著优于这两种基线系统（SSDT 测试设置下的 StarGANv2-EST 除外）。这体现了 S2S-ZEST-diverse 在情感风格迁移方面优于基线方法，并且与客观评估中观察到的趋势一致。

5.2.8 任意到任意情绪风格迁移

在本实验中，源话语和参考话语具有相同的说话人身份和语言内容，仅情感类别不同。这种受控设置基于 ESD 数据集构建，包含 3600 条测试语句。

我们评估了 4×4 以下源情感与目标情感的所有组合：{愤怒、快乐、悲伤、惊讶}。情感相似度采用第5.2.4节所述指标计算。图5.8 展示了两个情感相似度矩阵：(i) 转换语音与参考情感之间的相似度；(ii) 转换语音与原始源情感之间的相似度。成功的情感迁移应提高与参考情感的相似度，同时降低与源情感的相似度。

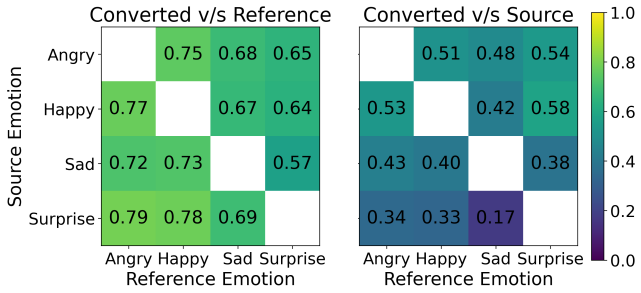


图 5.8：不同情感对之间的平均情感相似度。

表 5.3：中性到情感设置下对加性 MUSAN 噪声的鲁棒性。

条件	情感相似度 ↑	词错误率 (%) ↓
干净	0.67	3.65
20 dB	0.65	4.38
10 dB	0.64	7.74

在所有源情感与目标情感的组合中，转换语音始终与参考情感表现出较高的相似度（例如，惊讶→愤怒：0.79，愤怒→快乐：0.75）。与此同时，对于大多数组合，转换语音与原始源情感的相似度相对较低，这表明模型已摆脱源情感。对于某些转换（例如，悲伤→惊讶），与参考情感的相似度相对较低，反映出在这两种情绪状态之间进行转换更具难度。总体而言，这些结果表明，S2S-ZEST-diverse 支持在富表现力的源条件下进行定向情绪风格迁移，而不仅限于中性源场景。

5.2.9 对加性噪声的鲁棒性

为评估噪声条件下的鲁棒性，我们在 SSST 测试设置（表 5.1）中进行实验，在源语音中加入 20 dB 和 10 dB 信噪比的加性 MUSAN 噪声。参考语音保持干净。情感相似度和词错误率的计算方式如第 5.2.4 节所述。结果汇总于表 5.3。

在中等噪声下，情感相似度基本保持稳定，仅从 0.67（干净）小幅下降至 0.65（20 dB）和 0.64（10 dB）。然而，就可懂度而言，WER 从 3.65%（干净）逐步上升至 4.38%（20 dB）和 7.74%（10 dB），凸显了加性噪声对内容保留的影响。尽管在较低信噪比下识别性能有所下降，情感相似度仍相对稳定，表明即使源信号受到噪声干扰，情绪风格仍能得到有效传递。

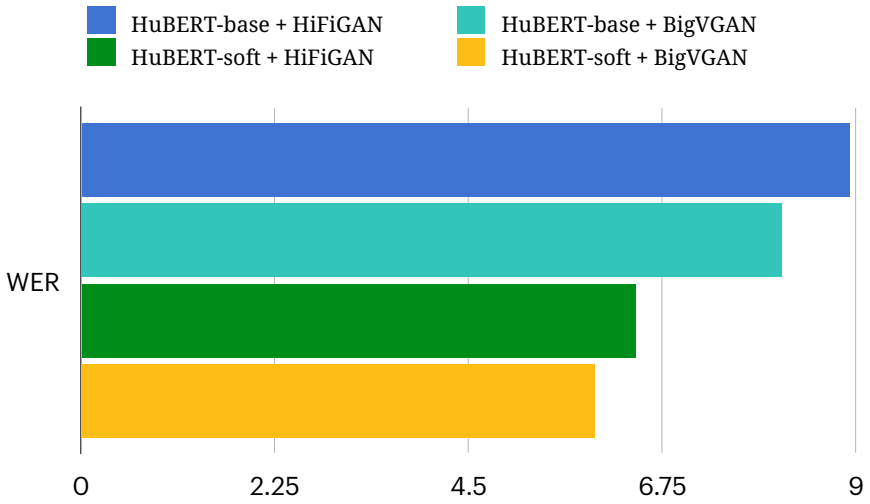


图 5.9：使用语音分词器和合成器的不同组合重新合成的语音的词错误率。

5.3 讨论

本节中的所有实验均采用在 ESD数据集上训练的基础 S2S-ZEST 配置。

5.3.1 语音分词器和合成器的选择

为评估语音分词器和合成器的影响，我们开展了一项实验，将 soft-HuBERT 特征替换为预训练 HuBERT-base模型提取的特征。我们使用由 HuBERT-base 特征生成的词元，对 ESD 测试集中的语音信号进行重构。此外，我们还使用 soft-HuBERT 特征生成的词元训练了一个 HiFi-GAN模型，作为替代语音合成器。采用这些不同配置重构的语音信号的词错误率结果如图 5.9 所示。

要点：1）无论使用哪种合成器，使用 soft-HuBERT 词元重建的语音，其词错误率都低于使用 HuBERT-base 词元重建的语音。先前的研究 [149] 表明，与 HuBERT-base 词元相比，连续的 soft-HuBERT 特征能够提升可懂度。我们的结果显示，即使将这些特征离散化为词元，语音重建的词错误率仍低于使用 HuBERT-base 词元的情况，保留了这一优势。2）无论使用哪种分词器，在参数数量相同（14M）的情况下，BigVGAN-base模型的表现始终优于 HiFi-GAN。这一提升可归因于 BigVGAN 中改进的激活函数。

表 5.4：在 ESD 数据集的四种测试设置下，使用 S2S-ZEST 和不含持续时间预测的 S2S-ZEST 时的词级 WER、情感相似度和皮尔逊相关系数（PCC）。

Test 设置	词错误率 ↓		情感相似度 (%) ↑		词级皮尔逊相关系数 ↑	
	S2S-ZEST	–持续时间预测	S2S-ZEST	–持续时间预测	S2S-ZEST	–持续时间预测
SSST	5.58	**4.76**	**0.69**	0.61	**0.68**	0.51
SSDT	4.94	**4.10**	**0.71**	0.62	**0.18**	0.04
DSST	5.55	**4.85**	**0.58**	0.51	**0.58**	0.41
DSDT	6.48	**5.59**	**0.56**	0.52	**0.14**	5×10^{-4}

5.3.2 时长预测器的影响

为评估时长预测器的影响，我们开展了一项实验：使用源语音词元序列合成转换后的语音，不对时长进行任何修改。实验结果报告于表 5.4。

关键要点：1）使用时长预测器会导致更高的词错误率。然而，它提升了语速相关性，表明更好的节奏表达是以在使用中性语音训练的自动语音识别上识别性能下降为代价的。这与 Maimon 等人的观察一致 [157]。2）在全部4种测试设置中情感相似度的提升表明，引入时长预测器能够实现有效的情感风格迁移。

5.3.3 对抗损失的重要性

细心的读者可能会询问在训练 S2S-ZEST 时两项对抗损失的必要性。我们通过 t-SNE 可视化 [165] 对 x向量 [150] 以及来自 ESD 数据集的两名说话人相应的说话人嵌入进行分析（见图 5.10）。在未进行情感对抗训练时，x向量的 t-SNE图呈现出与情感类别相对应的清晰聚类。为此，我们引入情感对抗损失（式5.1），其中 $\lambda_{adv}^{emo} > 0$，迫使说话人嵌入实现情感解耦。该方法的有效性由图5.10中的说话人嵌入 t-SNE图得以证实。

此外，我们通过实证评估情感分类器中的说话人对抗损失（记为 λ_{adv}^{spk} 见公式5.2）。这些消融实验的结果见表5.5。**主要结论**：1）在说话人嵌入的训练过程中移除情感对抗损失，会导致四种测试设置下的情感准确率指标均显著下降。这证实了 x 向量本质上会将说话人信息与情感信息纠缠在一起，因此解耦至关重要。2）从情感



图 5.10：两位说话人 I 和 II 在使用和不使用对抗损失时的说话人嵌入 t-SNE图。在这些图中，颜色对应于 ESD 中的 5 种情感类别。

表 5.5：在 ESD 数据集的四种测试设置下，使用 S2S-ZEST、移除情感对抗损失的 S2S-ZEST（$\lambda_{adv}^{emo} = 0$）以及移除说话人对抗损失的 S2S-ZEST（$\lambda_{adv}^{spk} = 0$）所得的 WER、情感相似度和说话人相似度

Test 设置	词错误率 ↓			情感准确率（%） ↑			说话人相似度 ↑		
	S2S-ZEST	−情感。Adv.	−说话人。对抗式	S2S-ZEST	−情感。Adv.	−说话人。对抗式	S2S-ZEST	−情感 Adv.	−说话人 Adv.
SSST	**5.58**	5.91	5.64	**0.69**	0.58	0.67	**0.74**	0.74	0.74
SSDT	**4.94**	5.02	5.01	**0.71**	0.47	0.67	0.73	**0.75**	0.74
DSST	**5.55**	5.97	5.73	**0.58**	0.48	0.56	**0.74**	0.74	0.72
DSDT	**6.48**	6.59	6.51	**0.56**	0.32	0.51	**0.72**	0.7	0.71

表 5.6：在不同训练条件下以及有无数据增强时，IEMOCAP 4 类别与 6 类别分类的加权 F1 分数。Orig. 指使用全部训练样本，而 100 和 200 表示每个类别仅使用 100 或 200 个样本来训练模型。无数据增强指未使用增强数据的情况，而使用合成样本的实验则记为 Aug。

训练数据	IEMOCAP-4			IEMOCAP-6		
	原始	200	100	原始	200	100
无数据增强	65.85	59.33	49.76	51.67	43.13	36.41
Aug.	**67.13**	**61.09**	**52.61**	**52.91**	**46.06**	**40.02**

分类器（见式 5.2）会导致测试设置 DSST 和 DSDT 的说话人相似度下降。3）去除任一对抗损失都会提高词错误率，表明说话人-情感解耦提升了转换语音的质量。

5.3.4 在语音情感识别中的应用

S2S-ZEST 的情绪风格迁移能力可用于语音情感识别（SER）中的数据增强。为对此进行评估，我们使用 IEMOCAP 数据集[96]。

该数据集的详细信息见第 2.5 节。对于语音情感识别模型训练，我们使用 WavLM-base 提取特征[11]，并遵循 SUPERB 评测框架[118]。

对于数据增强，我们从 IEMOCAP 训练集划分中选取 100 条中性话语，以及每种情绪类别各 50 条话语。我们将这些情绪话语用作参考语音信号，通过 S2S-ZEST 为每种情绪类别生成总计 5000 条转换后话语。由于转换后的话语涉及未见说话人和参考语音信号，我们根据 Emo.Sim. 分数进行质量筛选（第 5.2.4 节），并依据该分数为每种情绪类别选取排名前 100 的话语。因此，IEMOCAP-4 类别设置使用 300 条话语进行数据增强，而 IEMOCAP-6 类别设置则使用 500 条话语。

为进一步探究数据增强的影响，我们创建了规模较小的 IEMOCAP 数据集版本，其中每种情绪类别分别包含 100 条和 200 条训练话语。我们在表 5.6 中报告了该数据集在所有测试设置下的加权 F1 分数。

要点：1）在完整的 IEMOCAP-4 数据集中，数据增强使性能提升了 3.69%（相对值）。然而，对于 IEMOCAP-4(100)（每个类别仅有 100 个样本），性能提升达到 5.67%，凸显了数据增强在低资源环境中的更大作用。IEMOCAP-6 也呈现类似趋势，提升幅度从 2.57%（完整数据集）跃升至 5.68%（每个类别仅有 100 个样本）。2）尽管 IEMOCAP-6 包含两个未见情感类别（“兴奋”和“恐惧”），S2S-ZEST 在全部三种测试设置下均有所提升，表明 S2S-ZEST 具有良好的泛化能力。

5.4 章节总结

主要内容：本章提出了一种语音到语音情感风格迁移方法 S2S-ZEST，这是一种能够从情感因素和非情感因素中重建语音的自编码框架。内容与说话人嵌入采用预训练模型提取，而情感嵌入提取模型则通过训练获得。此外，我们还开发了一个以这些因素为条件的音高与时长预测流程，并将其与基于 BigVGAN 架构的语音合成模块相结合，以实现语音重建。在风格迁移过程中，内容与说话人嵌入来自源语音，而从参考话语中提取的情感嵌入则决定转换语音的音高和时长。我们通过客观与主观实验，在说话人与文本匹配及不匹配场景下评估 S2S-ZEST，并证明了其对未见说话人和目标情感的泛化能力。

第 6 章

总结与未来拓展

本论文通过联合分析声学与语义信息，探讨了口语情感建模问题，旨在推动情感理解与情感合成的发展。本章首先综述了针对不同问题设置所取得的各项贡献，随后讨论所提方法的局限性，并概述基于本研究所得见解的未来研究方向。

6.1 贡献总结

第 3章：情感感知预训练。本章介绍了针对语音和文本模型的预训练策略，使其更好地契合情感理解。本章的主要贡献总结如下：

- 我们提出了一种用于语音情感识别（SER）的新型自监督框架，该框架采用双编码器，分别对语义信息和声学信息进行建模。该模型称为 CARE（情感的内容与声学表征），能够显式捕捉口语中互补的情感相关线索（第 3.1.1.2节）。

- 我们提出了一种监督预训练流程，利用大语言模型从口语中生成情感标注文本数据，从而在缺少人工标注文本语料库的情况下实现情感感知文本表示学习（第 3.2.1 节）。

- 通过大量实验，我们证明了声学-语义建模在提升语音情感识别方面的优势，以及情感感知监督预训练对于基于文本的情感识别的有效性（第3.1.2.5、3.2.2.4节）。

第4章：口语多模态情绪识别。本章探讨了两种具有挑战性的设置下的口语情感识别问题：对话情感识别（ERC）以及自然录音条件下的情感识别。本章的主要贡献总结如下：

- 针对对话情感识别（ERC），我们提出了层次化和模块化建模框架，能够显式捕捉对话结构和模态特定动态。具体而言：

  – 我们提出了一种层次化建模范式：首先独立学习单模态话语级表示，随后对对话轮次进行上下文建模，并利用交叉注意力进行多模态融合。实验结果表明，显式建模对话层次能够带来持续的性能提升（第 4.1.2.2 节）。

  – 我们提出了 MiSTER-E，这是一种模块化专家混合框架，通过 logit 级专家融合，将模态特定上下文建模与多模态整合解耦。该设计可实现自适应的逐话语专家加权，并有效应对模态不平衡；当跨模态线索不可靠时，单模态专家可以占据主导地位（第 4.1.3.1 节）。

  – 所提出的专家混合模型在基准对话情感识别数据集——IEMOCAP、MELD 和 CMU-MOSI——上取得了最先进的性能，且不依赖说话人身份信息（第 4.1.3.3 节）。

- 针对自然录音条件下的情感识别，我们介绍了*Abhinaya*，这是我们为自然条件下语音情感识别挑战赛开发的系统。

第 5章：语音中的情感风格迁移。本章探讨了语音中的情感风格迁移（EST）问题，重点关注无文本、非平行语音到语音场景。目标是在保留源语音信号的语言内容和说话人身份的同时，将参考语音信号中的情感说话风格迁移到源语音信号中。本章的主要贡献总结如下：

- 我们将情感风格迁移表述为一个解耦表示学习问题，在语音表示中明确分离情感风格、语言内容和说话人身份。这种表述无需依赖文本监督，即可对情感属性进行可控操控（第5.1节）。

- 我们提出了一种适用于非平行设置的零样本情感风格迁移框架。在这种设置下，训练期间无法获得同一说话人以不同情绪说出相同内容的配对录音。所提出的方法能够泛化至未见说话人、未见内容和未见情感组合（第5.2.3节）。

- 我们开展了广泛的客观评估和主观评估，以评估转换语音的情感相似度、说话人保持、内容可懂度和自然度。结果表明，所提方法在保持说话人身份和语言内容的同时，实现了有效的情感迁移（见第 5.2.3节）。

6.2 局限性

我们在未充分探索所有替代方案的情况下，做出了各种假设并采用了多种建模策略。在此，我们重点介绍该工作的若干一般性局限。

- 针对不同问题所提出的架构主要基于经验，并受到已有工作的指导。我们根据性能指标对其进行了提出和分析。最终模型的网络架构和处理过程是基于性能表现确定的。

- 本论文开展的部分工作是在大语言模型（LLM）尚未普及时完成的。随着这些大规模模型不断改进，这类建模范式可能不再具有相关性。不过，其中一些方法在低算力环境下仍可能具有价值。一个有趣的未来方向是研究：当 CARE 作为大语言模型架构中的语音编码器使用时，诸如 CARE 这样的模型是否能够改善语音大语言模型的情感理解。此外，使用语音大语言模型进行语音到语音情感风格迁移也是一个值得探索的研究方向。

- 我们并未为本论文提出的不同模型所预测的情感类别提供任何论证或推理。

针对本论文所考虑的不同问题，所提出技术的局限性现列举如下：

- 尽管 CARE 在所考察的 8 个数据集上的平均表现优于许多大型模型，但它在某些数据集（例如 CMU-MOSI）上的表现仍然不佳。在更大规模的数据集上对模型进行充分训练，或改进语义建模，可能有助于缓解这一问题。

- 尽管 MiSTER-E 在多个基准测试中取得了出色的性能，但仍存在一些局限性：（i）使用大型 LLM/SLLM 编码器会带来不可忽视的计算和内存开销，这可能限制其在资源受限环境或实时环境中的适用性，尽管采用了参数高效微调。我们已在表 4.5中部分回应了这一问题，其中所提方法即使应用于基于非 LLM 的特征也显示出优势；（ii）我们的评估主要在由脚本化或半脚本化对话组成的基准数据集上进行（例如电视节目和表演式对话），而在领域偏移至自发的真实世界对话环境时的表现仍是一个悬而未决的问题；（iii）已知情感推断模型容易受到与人口统计因素、语言使用和标注主观性相关的数据集偏差影响，而我们尚未在本研究中开展任何偏差/公平性分析；（iv）在整个研究过程中，我们有意避免了明确的说话人身份建模，但利用说话人元数据可能有助于改进对人际动态的建模。未来可以进一步解决其中一些局限性。

- 对于自然语音情感识别任务，不同人工标注者对语音文件给出的评分作为数据集的一部分提供（第 4.2.1.5节）。对标注者行为进行建模，或许能更好地对口语中的情绪进行建模。然而，在本论文中，我们关注的是人工标注者给出的多数票。

- 基础版 S2S-ZEST 配置使用 ESD 数据集的英语子集进行训练（包含约 15 小时、来自 10 位说话人的语音），这限制了训练期间的说话人多样性，并影响了模型对未见过的声音的泛化能力。尽管 S2S-ZEST-diverse 变体表明，增加训练数据多样性能够提升说话人相似度和内容保留能力，但其在未见说话人场景中的表现仍不及使用规模大得多的语料库训练的大规模系统。未来的工作将探索使用规模更大、更多样化的情感语音数据集进行训练。

6.3 未来研究方向

本论文所呈现的工作为未来研究开辟了若干颇具前景的方向。

- **情绪分段：**在这篇论文中，当对话中的话语彼此划分清楚时，ERC 问题便得到了解决。更贴近实际的情形是，话语的起止位置事先并未标注。这一问题被称为

表6.1：Google 自动语音识别系统在所用三个数据集上的词错误率（WER）（%）。所用缩写：快乐：快乐，中性：中性，愤怒：愤怒，兴奋：兴奋，沮丧：沮丧，厌恶：厌恶，积极：积极和消极：消极。对于IEMOCAP，这些数字指的是合并后的训练集和验证集。

数据集划分	IEMOCAP						MELD							CMU-MOSI	
	Hap.	Sad	Neu.	Ang.	Exc.	Fru.	Ang.	Sad	Neu.	Fear	Sur.	Dis.	Joy	Pos.	Neg.
训练集	36	44	31.2	23.4	36.5	28	50.1	50.1	50.2	54.8	59	47.5	54.1	37.3	37.9
Val							47.4	49.9	46.4	52.8	58.8	46.6	50.7	34	37.9
Test	34.5	47.5	30.2	27.3	36	30.5	50.5	49.2	48.1	59.5	61.2	49.3	54.5	38.2	40.4

表 6.2：我们的系统在测试时使用自动语音识别转录文本时，在各数据集上的加权 F1 分数。

文本转录使用测试时	IEMOCAP		MELD	CMU-MOSI
	四分类	六分类		
ASR	84.6%	68。1%	50。2%	80.1%
原始	85.9%	70。5%	65.8%	85。8%

情绪分段由 Wang 等人提出， [166], 可以视为 ERC 问题的下一步。

- **在对话情感识别中使用自动语音识别转录文本：**在本论文中，所提供的文本转录被用作口语语义成分，以进行对话情感识别。除此之外，在推理过程中可能需要对语音信号进行转录。为了了解这种变化所带来的影响，我们使用所提供的转录文本训练提出的 HCAM 模型（第 4.1.2.2节），但使用自动语音识别系统生成的转录文本对其进行测试[1]。该自动语音识别系统的词错误率（WER）针对三个数据集分别报告于表 6.1 中。如表所示，情感对话语音的词错误率显著高于其他受控数据集。对于 MELD，自动语音识别性能最低，部分原因在于该数据集中的背景噪声水平较高。在表 6.2中展示了使用自动语音识别转录文本对情感识别的影响，其中可以看到，MELD 等数据集的性能显著下降。这一观察结果引出了一个有趣的未来研究方向——尝试联合进行对话情感识别和对话自动语音识别。

[1]https://cloud.google.com/speech-to-text

- **将零样本 EST 扩展到多语言环境：**我们提出的 EST 系统仅使用英语语音进行评估。一个有趣的未来方向是将这项工作推广到其他语言。更具挑战性的一点是，在以无文本、非平行的方式训练模型的同时，将一种语言的情感风格迁移到另一种语言。为了在保留说话人身份的同时，使用来自不同语言的多样化说话人进行训练，其挑战可能会高于单语言环境。

- **从语音中识别心理健康特征：**最后，提升对口语情感的理解能力，应有助于那些向聊天机器人寻求心理支持的用户。对这些模型进行预训练，以提升其情感理解能力，目标应是让它们更好地服务这一群体。这是一个鲜有探索的领域，目前最大的障碍是缺乏合适的数据集。收集数据并建立适当的基准，是推动这一领域发展的下一项重要工作。

6.4 章节总结

本章总结了论文的主要贡献，并将其置于口语情感建模的更广阔领域中。尽管所提出的方法在多项任务中表现出色，涵盖情感感知预训练、多模态情感识别和情感风格迁移，但也凸显了若干尚待解决的挑战，包括对自动语音识别系统错误的鲁棒性、计算成本，以及实验仅限于英语。本章概述的未来研究方向指向更加真实、包容且具有泛化能力的情感感知语音系统。总的来说，本论文的工作为推进具有情感智能的口语技术奠定了基础，同时也明确了持续取得进展的路径。

附录 A

架构参考组件

A.1 自监督语音模型

A.1.1 PASE+

PASE+ [47]是早期专为语音表征学习设计的自监督模型之一。该模型以多任务方式进行训练，从输入波形的受损版本中预测一系列短期和长期语音描述符，例如梅尔频率倒谱系数、滤波器组能量和韵律特征。通过学习重建多种语音属性，PASE+能够捕捉通用声学表征。尽管它最初用于自动语音识别（ASR）评估，但其自监督训练范式使其能够应用于广泛的下游语音处理任务，包括语音情感识别。

A.1.2 HuBERT

HuBERT（隐藏单元 BERT） [10] 引入了一种用于语音表征学习的掩码预测框架。在这种方法中，首先使用k均值算法对连续语音特征进行聚类，以获得离散伪标签。随后，模型通过 Transformer 编码器进行训练，以预测输入序列中被掩码部分对应的聚类分配。通过迭代优化聚类目标和模型表征，HuBERT 学得了语音嵌入，并在各种语音任务中展现出强大的性能。

A.1.3 WavLM

WavLM [11] 通过引入更加多样的数据增强策略和去噪训练目标，扩展了 HuBERT 框架。除了掩码预测，WavLM 还引入了话语混合和噪声扰动，以提升所学表征的鲁棒性和说话人感知能力。

这些改进使 WavLM 在多项基准测试中取得了优于 HuBERT 的表现，包括语音识别、说话人验证和语音情感识别。

A.1.4 data2vec

data2vec [72] 提出了一种适用于语音、视觉和文本模态的统一自监督学习框架。在语音设置中，学生网络经过训练，用于预测教师网络生成的上下文化潜在表征。教师接收未屏蔽的输入，并通过对学生网络参数进行指数移动平均（EMA）来更新。与 HuBERT 等基于聚类的方法不同，data2vec 直接回归连续隐表征，从而鼓励模型捕捉语音中的全局上下文结构。事实证明，由此生成的表征能够有效泛化到各种下游语音处理任务。

A.2 SONAR

SONAR [121] 是一种大规模多语言、多模态嵌入模型，旨在学习语音与文本之间的对齐表征。与仅专注于声学建模的传统自监督语音模型不同，SONAR 使用平行语音文本数据和对比目标进行训练，从而生成在不同模态和语言之间语义对齐的嵌入。

该模型由面向语音和文本的模态专用编码器组成，可将输入投射到共享嵌入空间中。模型使用大量多语言数据进行训练，并采用翻译和语音文本对齐监督，使学习到的表征能够捕获高层次的语义信息。因此，SONAR 嵌入在跨语言检索、语音到文本对齐和多模态迁移任务中尤其有效。

A.3 Whisper

Whisper [109] 是一个大规模编码器—解码器语音模型，以监督方式训练，用于自动语音识别（ASR）和语音翻译。该模型使用从网络收集的约 680,000 小时多语言、多任务语音数据进行训练，使其能够在不同语言、口音和录音条件下实现稳健的泛化。

从架构上看，Whisper 遵循基于 Transformer 的序列到序列框架。编码器处理从原始语音中提取的对数梅尔频谱特征，而解码器则以自回归方式生成文本词元。除了转录之外，该模型还经过训练，能够

预测语言识别结果和任务标记，使其能够在统一框架内处理多种语音处理任务。

A.4 SALMONN

SALMONN [115]是一种语音大语言模型（SLLM），它将预训练语音编码器（Whisper-large-v2）和音频编码器（BEATs）与大语言模型相结合，从而在指令遵循框架内实现语音理解。其架构包含一个投影层，用于将声学表征与预训练大语言模型的嵌入空间对齐。组合后的模型通过指令微调进行训练，因此能够处理语音输入，并针对多种语音理解任务生成基于文本的响应。与以重建目标为优化方向的传统自监督语音模型不同，SALMONN 经过训练，可将语音表征与大语言模型的高层语言与推理能力对齐。

参考文献

[1] Pawe-lTarnowski, MarcinKo-lodziej, Andrzej Majkowski, 和 Remigiusz J Rak。使用面部表情进行情感识别。*Procedia Computer Science*，108：1175–1184，2017年。1[2] Klaus R Scherer。情绪的声音交流：研究范式综述。*Speech communication*，40(1-2)：227–256，2003年。1，10

[3] Costanza Navarretta。交际性身体行为中的个体性。载于*Cognitive Behavioural Systems: COST 2102* 国际培训学校，德累斯顿，德国，*2011*年*2*月*21*日至*26*日，修订版精选论文集，第417–423页。Springer，2012年。1

[4] R Benjamin Knapp, Jonghwa Kim, andElisabeth Andr´e。生理信号及其在人机交互中增强情感识别方面的应用。载于情感导向系统：*Humaine* 手册，第133–159页。Springer，2010年。1

[5] Jason Phang、Michael Lampe、Lama Ahmad、Sandhini Agarwal、Cathy Mengying Fang、Auren R Liu、Valdemar Danry、Eunhae Lee、Samantha WT Chan、Pat Pataranutaporn 等人。《探究 ChatGPT 上的情感化使用与情绪健康》。*arXiv* 预印本 *arXiv:2504.03888*，2025年。1

[6] Aditya Nrusimha Vaidyam、Hannah Wisniewski、John David Halamka、Matcheri S Kasha- van 和 John Blake Torous。《心理健康领域中的聊天机器人与对话代理：精神医学领域综述》。*The Canadian Journal of Psychiatry*，64(7):456–464，2019年。1

[7] Xi Zheng、Zhuoyang Li、Xinning Gui 和 Yuhan Luo。《定制情感支持：个体如何构建并与大语言模型驱动的聊天机器人互动》。载于 *2025*年*CHI*计算系统人因大会论文集，第1–20页，2025年。1

[8] Gemini 团队、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 等。Gemini：一个

能力极强的多模态模型家族。*arXiv* 预印本 *arXiv:2312.11805*，2023年。xiv，1，73

[9] Steffen Schneider、Alexei Baevski、Ronan Collobert 和 Michael Auli。wav2vec：用于语音识别的无监督预训练。载于*Interspeech 2019* 会议论文集，第3465–3469 页，2019年。3，10

[10]Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：通过隐藏单元的掩码预测进行自监督语音表示学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29：3451–3460，2021年。3，10，23，30，35，47，105

[11] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等。WavLM：面向全栈语音处理的大规模自监督预训练。*IEEE* 信号处理精选主题期刊，16(6)：1505–1518，2022年。3，10，11，23，30，32，67，97，105

[12] Soujanya Poria、Navonil Majumder、Rada Mihalcea 和 Eduard Hovy。对话中的情感识别：研究挑战、数据集与最新进展。*IEEE Access*，7:100943–100953，2019年。3，56

[13] Soujanya Poria、Erik Cambria、Devamanyu Hazarika、Navonil Majumder、Amir Zadeh 和 Louis-Philippe Morency。用户生成视频中的上下文依赖型情感分析。载于第*55*届计算语言学协会年会论文集（第*1*卷：长论文），第873–883页，2017年。3，13

[14] Soujanya Poria、Erik Cambria、Rajiv Bajpai 和 Amir Hussain。情感计算综述：从单模态分析到多模态融合。*Information Fusion*，37:98–125，2017年。3，45

[15] Seunghyun Yoon、Seokhyun Byun 和 Kyomin Jung。使用音频和文本进行多模态语音情感识别。载于 *2018*年 *IEEE* 口语语言技术研讨会（*SLT*），第 112–118 页。IEEE，2018年。 3

[16] Ying Shen、Huiyu Yang 和 Lin Lin。自动抑郁检测：情感音频文本语料库和基于GRU/BiLSTM的模型。载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议*(ICASSP)*，第 6247–6251 页。IEEE，2022年。3

[17] Wen Wu、Chao Zhang 和 Philip CWoodland。语音抑郁检测中的自监督表征。载于 *ICASSP 2023-2023 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 1–5 页。IEEE，2023年。3，19，29

[18] Xin Wang、Hong Chen、Si'ao Tang、Zihao Wu 和 Wenwu Zhu。解耦表征学习。*IEEE*模式分析与机器智能汇刊，46(12):9677–9696，2024年。3

[19] Yoohwan Kwon、Soo-Whan Chung 和 Hong-Goo Kang。解耦框架中说话人表征的类内变化缩减。载于*Interspeech 2020*论文集，第3231–3235页，2020年。3

[20] David Chan 和 Shalini Ghosh。用于自动语音识别的内容-上下文因子化表征。载于 *Interspeech 2022*论文集，第61–65页，2022年。3

[21] Berrak Sisman、Junichi Yamagishi、Simon King 和 Haizhou Li。语音转换及其挑战概述：从统计建模到深度学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29:132–157，2020年。3

[22]Soumya Dutta 和 Sriram Ganapathy。基于说话人解耦的零样本音频到音频情感迁移。载于 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 10371–10375 页。IEEE，2024年。3,16, 88,89

[23] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深层双向变换器预训练。载于 *2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长论文和短论文），第 4171–4186 页，2019年。4,11, 12,14,24

[24] Weidong Chen、Xiaofen Xing、Peihao Chen 和 Xiangmin Xu。Vesper：用于语音情感识别的紧凑且高效的预训练模型。 *IEEE*情感计算汇刊，15(3):1711–1724，2024年。4, 11,25,81

[25] Ziyang Ma、Zhisheng Zheng、Jiaxin Ye、Jinchao Li、Zhifu Gao、Shiliang Zhang 和 Xie Chen。emotion2vec：用于语音情感表征的自监督预训练。载于计算语言学协会会议论文集：*ACL 2024*，第 15747–15760 页，2024年。 4，12，30，32，87

[26] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。神经网络中的知识蒸馏。 *arXiv* 预印本 *arXiv:1503.02531*，2015年。 4[27] Florian Eyben、Martin W¨ollmer 和 Bj¨orn Schuller。OpenSMILE：慕尼黑推出的多功能快速开源音频特征提取器。载于第 *18* 届 *ACM*国际多媒体会议论文集，第 1459–1462 页，2010年。 5，10，36[28] GustavoL´ opez、Luis Quesada 和 Luis A Guerrero。Alexa、Siri、Cortana 与 Google Assistant：基于语音的自然用户界面对比。载于应用人因与工效学国际会议，第 241–250 页。Springer，2017年。 5[29] Yujia Wang、Wenguan Wang、Wei Liang 和 Lap-Fai Yu。漫画引导的语音合成。*ACM*图形学汇刊 *(TOG)*，38(6):1–14，2019年。 5

[30] Matthew Marge、Carol Espy-Wilson、Nigel G Ward、Abeer Alwan、Yoav Artzi、Mohit Bansal、Gil Blankenship、Joyce Chai、Hal Daum´e III、Debadeepta Dey 等。与机器人的口语交互：对未来研究的建议。 *Computer Speech& Language*，71:101255，2022年。 5

[31] Tianxin Xie、Yan Rong、Pengfei Zhang、Wenwu Wang 和 Li Liu。大语言模型时代的可控语音合成：系统综述。载于 *2025*年自然语言处理实证方法会议论文集，第 764–791 页，2025年。 5

[32] Philip Lieberman与Sheldon B Michaels。与语音的情感内容相关的基频和包络幅度的若干方面。《美国声学学会杂志》，34(7):922–927，1962年。 10[33] Valery Petrushin。语音情感：识别及其在呼叫中心中的应用。载于《工程中的人工神经网络会议论文集》，第710卷，第22页，1999年。 10，25[34] Bj¨orn Schuller、Stefan Steidl和Anton Batliner。INTERSPEECH 2009情感挑战赛。载于*Interspeech 2009*，第312–315页，2009年。 10[35] Bj¨orn Schuller、Stefan Steidl、Anton Batliner、Alessandro Vinciarelli、Klaus Scherer、Fabien Ringeval、Mohamed Chetouani、Felix Weninger、Florian Eyben、Erik Marchi等。INTERSPEECH 2013计算副语言学挑战：社会信号、冲突、情感、自闭症。载于 *Proc. Interspeech 2013*，第148–152页，2013年。 10

[36] Florian Eyben、Klaus R Scherer、Bj¨orn W Schuller、Johan Sundberg、Elisabeth Andr´e、Carlos Busso、Laurence Y Devillers、Julien Epps、Petri Laukka、Shrikanth S Narayanan等。面向语音研究与情感计算的日内瓦极简声学参数集（GeMAPS）。*IEEE*情感计算汇刊，7(2):190–202，2015年。10，36

[37] Zafi Sherhan Syed、Julien Schroeter、Kirill A Sidorov 和 A David Marshall。计算副语言学：根据语音声学自动评估情感、心境和行为状态。载于 *Interspeech*，第 511–515 页，2018年。10

[38] Andreas Triantafyllopoulos、Gil Keren、Johannes Wagner、Ingmar Steiner 和 Bj¨orn W Schuller。使用用于语音增强的深度残差网络实现鲁棒的语音情感识别。载于 *Proc. Interspeech 2019*，第 1691–1695 页，2019年。10

[39] Fasih Haider、Senja Pollak、Pierre Albert 和 Saturnino Luz。低资源环境中的情感识别：自动特征选择方法评估。*ComputerSpeech & Language*，65:101119，2021年。10

[40] Seong-Gyun Leem、Daniel Fulford、Jukka-Pekka Onnela、David Gard 和 Carlos Busso。在梯形网络上分离情绪嵌入与重构嵌入，以提升噪声条件下语音情感识别的鲁棒性。*Proc. Interspeech 2021*，第 2871–2875 页，2021年。10

[41] Tara N Sainath、Brian Kingsbury、Abdel-rahman Mohamed 和 Bhuvana Ramabhadran。在*2013*年 *IEEE* 自动语音识别与理解研讨会上发表，页码 297–302。IEEE，2013年。10

[42] Tara N Sainath、Ron J Weiss、Andrew W Senior、Kevin W Wilson 和 Oriol Vinyals。在*Interspeech*上发表，页码 1–5。德国德累斯顿，2015年。10

[43] Yedid Hoshen、Ron J Weiss 和 Kevin W Wilson。基于原始多通道波形的语音声学建模。发表于*2015*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*），页码 4624–4628。IEEE，2015年。10

[44] Mirco Ravanelli 与 Yoshua Bengio。使用 SincNet 从原始波形进行说话人识别。见 *2018*年 *IEEE* 口语语言技术研讨会（*SLT*），第 1021–1028 页。IEEE，2018年。10

[45] Neil Zeghidour、Olivier Teboul、F´elix De Chaumont Quitry 与 Marco Tagliasacchi。LEAF：一种用于音频分类的可学习前端。 *arXiv* 预印本 *arXiv:2101.08596*，2021年。 10[46] Aaron van den Oord、Yazhe Li 与 Oriol Vinyals。基于对比预测编码的表征学习。 *arXiv* 预印本 *arXiv:1807.03748*，2018年。 10[47] Santiago Pascual、Mirco Ravanelli、Joan Serr`a、Antonio Bonafonte 与 Yoshua Bengio。从多个自监督任务中学习与问题无关的语音表示。见*Interspeech 2019* 会议论文集，第 161–165 页，2019年。 10，105[48] Mirco Ravanelli、Jianyuan Zhong、Santiago Pascual、Pawel Swietojanski、Joao Monteiro、Jan Trmal 与 Yoshua Bengio。面向鲁棒语音识别的多任务自监督学习。见*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 6989–6993 页。IEEE，2020年。 10， 24，33

[49] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：用于自监督学习语音表征的框架。神经信息处理系统进展，33：12449–12460，2020年。10，14，23，87

[50] Leonardo Pepino、Pablo Riera 和 Luciana Ferrer。使用 wav2vec 2.0 嵌入进行语音情感识别。载于*Interspeech 2021*，第 3400–3404 页，2021年。10，47

[51]Yingzhi Wang、Abdelmoumene Boumadane 和 Abdelwahab Heba。用于语音情感识别、说话人验证和口语语言理解的微调 wav2vec 2.0/HuBERT 基准。*arXiv* 预印本 *arXiv:2111.02735*，2021年。10

[52] Edmilson Morais、Ron Hoory、Weizhong Zhu、Itai Gat、Matheus Damasceno和Hagai Aronowitz。使用自监督特征的语音情感识别。载于*ICASSP 2022-2022*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第6922–6926页。IEEE，2022年。10

[53] Daria Diatlova、Anton Udalov、Vitalii Shutov 和 Egor Spirin。将 WavLM 应用于语音情感识别。载于 《*Odyssey 2024* 论文集》，第 303–308 页，2024年。 10[54] Christiane Fellbaum。 *WordNet*：电子词汇数据库。MIT出版社，1998年。 10[55] Carlo Strapparava、Alessandro Valitutti 等。WordNet Affect：WordNet 的情感扩展。载于 *LREC*，第 4 卷，第 40 页。葡萄牙里斯本，2004年。 10

[56] Raoul Biagioni。 *SenticNet* 情感词典：探索多词概念中的语义丰富性，第 4 卷。Springer，2016年。 10

[57] Cecilia Ovesdotter Alm、Dan Roth 和 Richard Sproat。文本中的情绪：用于基于文本的情绪预测的机器学习。载于人类语言技术会议暨自然语言处理实证方法会议论文集，第579–586页，2005年。11[58] Soumaya Chaffar 和 Diana Inkpen。使用异构数据集进行文本情绪分析。载于加拿大人工智能会议，第62–67页。Springer，2011年。11

[59] Flor Miriam Plaza-del Arco、Salud Mar´ıaJim´enez-Zafra、M Teresa Mart´ın-Valdivia 和 L Alfonso Urena Lopez。Sinai 在 SemEval-2018 任务1中的工作：推文情感识别。载于第*12*届语义评测国际研讨会论文集，第128–132页，2018年。11

[60] Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。向量空间中词表示的高效估计。*arXiv* 预印本 *arXiv:1301.3781*，2013年。11

[61] Jeffrey Pennington、Richard Socher 和 Christopher D Manning。GloVe：词表示全局向量。发表于*2014*年自然语言处理经验方法会议（*EMNLP*）论文集，第 1532–1543 页，2014年。 11

[62] Marco Polignano、Pierpaolo Basile、Marco De Gemmis 和 Giovanni Semeraro。使用双向长短期记忆网络、卷积神经网络和自注意力，对文本情绪检测中的词嵌入进行比较。发表于第*27*届用户建模、自适应与个性化会议附录论文集，第 63–68 页，2019年。 11

[63] Piyush Gupta、Inika Roy、Gunnika Batra 和 Arun Kumar Dubey。使用 GloVe嵌入解读文本中的情绪。发表于*2021*年计算、通信与智能系统国际会议（*ICCCIS*），第 36–40 页。IEEE，2021年。 11

[64] Elangel Neilea Shaday、Ventje Jeremias Lewi Engel 和 Hery Heryanto。比较 Word2Vec、GloVe 和 FastText 在歌曲歌词情绪分类中应用双向长短期记忆网络的方法。*Procedia Computer Science*，245：137–146，2024年。11

[65] Diksha Shukla 和 Sanjay K Dwivedi。用于亚马逊产品评论数据集情绪检测的词嵌入。载于 数据科学与应用国际会议，第 441–453 页。Springer，2024年。11

[66] Matthew E. Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。深度上下文化词表示。载于 Marilyn Walker、Heng Ji 和 Amanda Stent 编，*2018*年计算语言学协会北美分会会议论文集：人类语言技术，第 *1* 卷（长论文），第 2227–2237 页，新奥尔良，路易斯安那州，2018年6月。计算语言学协会。11

[67] Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。通过生成式预训练提升语言理解。2018年。11

[68] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez， Lukasz Kaiser 和 Illia Polosukhin。注意力机制就是你所需要的一切。神经信息处理系统进展，30，2017年。11，46

[69]Abdullah Al Maruf、Fahima Khanam、Md Mahmudul Haque、Zakaria Masud Jiyad、Muhammad Firoz Mridha 和 Zeyar Aung。基于文本的情绪检测的挑战与机遇：一项综述。*IEEE Access*，12：18416–18450，2024年。11

[70] Francisca Adoma Acheampong、Henry Nunoo-Mensah 和 Wenyu Chen。基于文本的情绪检测的 Transformer模型：基于BERT的方法综述。《人工智能评论》，54(8)：5789–5829，2021年。11

[71] Sheetal Kusal、Shruti Patil、Jyoti Choudrie、Ketan Kotecha、Deepali Vora 和 Ilias Pappas。自然语言处理应用的系统综述，以及未来的挑战，特别关注基于文本的情绪检测。《人工智能评论》，56(12):15129–15215，2023年。11

[72] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli。data2vec：一个用于语音、视觉和语言自监督学习的通用框架。载于国际机器学习会议，第1298–1312页。PMLR，2022年。12、30、35、106

[73] Hao Tian、Can Gao、Xinyan Xiao、Hao Liu、Bolei He、Hua Wu、Haifeng Wang 和 Feng Wu。Skep：用于情感分析的情感知识增强预训练。载于计算语言学协会第*58*届年会论文集，第4067–4076页，2020年。12

[74] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*，2019年。12，24，25，33，39，47

[75] Tiberiu Sosea 和 Cornelia Caragea。EMLM：面向情感相关任务的新预训练目标。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*2*卷：短论文），第286–293页，2021年。12

[76] Sepp Hochreiter 和J¨urgenSchmidhuber。长短期记忆。*Neural Computation*，9(8):1735–1780，1997年。13

[77] Taeyang Yun、Hyunkuk Lim、Jeonghwan Lee 和 Min Song。TELME：用于对话情感识别的教师引导多模态融合网络。载于*2024*年计算语言学协会北美分会年会论文集：人类语言技术（第*1*卷：长论文），第82–95页，2024年。13，60

[78] Jiang Li、Xiaoping Wang、Yingjian Liu 和 Zhigang Zeng。CFN-ESA：一种用于对话情感识别、具备情感转变感知能力的跨模态融合网络。*IEEE*情感计算汇刊，15(4):1919–1933，2024年。13，60

[79] Soumya Dutta 和 Sriram Ganapathy。用于情感识别的多模态Transformer，采用可学习的前端和自注意力机制。载于 *ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 6917–6921 页。IEEE，2022年。13

[80] Chul Min Lee 和 Shrikanth S Narayanan。探索检测口语对话中的情绪。*IEEE*语音与音频处理汇刊，13(2):293–303，2005年。14

[81] Qin Jin、Chengxin Li、Shizhe Chen 和 Huimin Wu。结合声学和词汇特征的语音情感识别。载于*2015*年 *IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 4749–4753 页。IEEE，2015年。14

[82] Haoqin Sun、Shiwan Zhao、Xuechen Wang、Wenjia Zeng、Yong Chen 和 Yong Qin。面向多模态情感识别的细粒度解耦表征学习。载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 11051–11055 页。IEEE，2024年。14

[83] Dongchan Min、Dong Bok Lee、Eunho Yang 和 Sung Ju Hwang。Meta-stylespeech：多说话人自适应文本转语音生成。载于机器学习国际会议，第 7748–7759 页。PMLR，2021年。 15，16

[84] Rongjie Huang、Yi Ren、Jinglin Liu、Chenye Cui 和 Zhou Zhao。GenerSpeech：迈向可泛化的域外文本转语音风格迁移。神经信息处理系统进展，35：10970–10983，2022年。15

[85] Yinghao Aaron Li、Cong Han、Vinay Raghavan、Gavin Mischler 和 Nima Mesgarani。StyleTTS 2：通过风格扩散、对抗训练和大型语音语言模型，实现人类水平的文本转语音。神经信息处理系统进展，36：19594–19621，2023年。15

[86] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Om- mer。利用潜在扩散模型实现高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10684–10695页，2022年。15

[87] Georgios Rizos、Alice Baird、Max Elliott 和 Bj¨orn Schuller。用于情感语音转换的 StarGAN：通过端到端情感识别的数据增强进行验证。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 3502–3506 页。IEEE，2020年。15

[88] Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim 和 Jaegul Choo。StarGAN：用于多域图像到图像转换的统一生成对抗网络。载于*IEEE*计算机视觉与模式识别会议论文集，第 8789–8797 页，2018年。15

[89] Gallil Maimon 和 Yossi Adi。在波形域中使用离散自监督单元进行说话风格转换。载于计算语言学协会 *EMNLP 2023* 论文集：*EMNLP 2023*，第 8048–8061 页，2023年。16

[90] Jixun Yao、Yuguang Yang、Yi Lei、Ziqian Ning、Yanni Hu、Yu Pan、Jingjing Yin、Hongbin Zhou、Heng Lu 和 Lei Xie。Promptvc：由自然语言提示驱动的潜在空间灵活风格化语音转换。载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议*(ICASSP)*，第 10571–10575 页。IEEE，2024年。 16

[91] Xueyao Zhang、Xiaohui Zhang、Kainan Peng、Zhenyu Tang、Vimal Manohar、Yingru Liu、Jeff Hwang、Dangna Li、Yuhao Wang、Julian Chan 等。Vevo：通过自监督解耦实现可控的零样本语音模仿。载于第十三届国际学习表征会议，2025年。16，88，89

[92] Jixun Yao、Yang Yuguang、Yu Pan、Ziqian Ning、Jianhao Ye、Hongbin Zhou 和 Lei Xie。Stablevc：基于条件流匹配的风格可控零样本语音转换。载于*AAAI*人工智能会议论文集，第39卷，第25669–25677页，2025年。16，17

[93] Zeqian Ju、Yuancheng Wang、Kai Shen、Xu Tan、Detai Xin、Dongchao Yang、Eric Liu、Yichong Leng、Kaitao Song、Siliang Tang 等。NaturalSpeech 3：采用因子化编解码器和扩散模型的零样本语音合成。载于国际机器学习会议，第22605–22623页。PMLR，2024年。16

[94] Jinsung Yoon、Wooyeol Jeong、Jio Gim 和 Young-Joo Suh。Maestro-evc：由参考音频和显式韵律引导的可控情感语音转换。*arXiv*预印本 *arXiv:2508.06890*，2025年。17

[95]Carlos Busso、Reza Lotfian、Kusha Sridhar、Ali N Salman、Wei-Cheng Lin、Lucas Goncalves、Srinivas Parthasarathy、Abinay Reddy Naini、Seong-Gyun Leem、Luz Martinez-Lucas 等。MSP-PODCAST 语料库。*arXiv* 预印本 *arXiv:2509.09791*，2025年。17，21，28，40，67，70

[96] Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan。IEMOCAP：交互式情感双人动作捕捉数据库。语言资源与评估，42(4):335–359，2008。17，21，28，33，40，50，96

[97] Zheng Lian、Bin Liu 和 Jianhua Tao。SMIN：用于对话情感识别的半监督多模态交互网络。*IEEE* 情感计算汇刊，14(3)：2415–2429，2022年。18，33，60

[98] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea。MELD：用于对话情感识别的多模态多方数据集。发表于计算语言学协会第*57*届年会论文集，第527–536页，2019年。18，21，28，33，40，45，50

[99] Amir Zadeh、Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency。MOSI：在线观点视频中情感强度与主观性分析的多模态语料库。*arXiv*预印本 *arXiv:1606.06259*，2016年。 18，21，28，33，40，50

[100] Jonathan Gratch、Ron Artstein、Gale M Lucas、Giota Stratou、Stefan Scherer、Angela Nazarian、Rachel Wood、Jill Boberg、David DeVault、Stacy Marsella 等。人机访谈的痛苦分析访谈语料库。载于 *LREC*，第 14 卷，第 3123–3128 页。雷克雅未克，2014 年。 18，21、28，33

[101] Steven R Livingstone 和 Frank A Russo。Ryerson情感语音与歌曲视听数据库（RAVDESS）：北美英语中面部与声音表达的动态多模态数据集。 *PloS one*，13(5)：e0196391，2018年。 19，21、28，33

[102] Philippe Gournay、Olivier Lahaie 和 Roch Lefebvre。加拿大法语情感语音数据集。载于第九届 *ACM*多媒体系统会议论文集，第 399–402 页，2018年。 19，21、28，33

[103] Felix Burkhardt、Astrid Paeschke、Miriam Rolfes、Walter F Sendlmeier、Benjamin Weiss 等人。德语情感语音数据库。载于 *Interspeech*，第 5 卷，1517–1520 页，2005年。 19，21，28，33

[104] Carlos Busso、Srinivas Parthasarathy、Alec Burmania、Mohammed AbdelWahab、Najmeh Sadoughi 和 Emily Mower Provost。MSP-IMPROV：用于研究情感感知的双人互动表演语料库。*IEEE*情感计算汇刊，8(1)：67–80，2016年。 19，21，31

[105] Kun Zhou、Berrak Sisman、Rui Liu 和 Haizhou Li。用于语音转换的已见与未见情感风格迁移：一种新的情感语音数据集。载于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，920–924 页。IEEE，2021年。 20， 21， 85， 87

[106] Houwei Cao、David G Cooper、Michael K Keutmann、Ruben C Gur、Ani Nenkova 和 Ragini Verma。《Crema-d：众包情感多模态演员数据集》。 *IEEE*情感计算汇刊，5(4):377–390，2014。 20，21

[107] John S Garofolo、Lori F Lamel、William M Fisher、Jonathan G Fiscus 和 David S Pallett。美国国防高级研究计划局 TIMIT 声学-音素连续语音语料库光盘只读存储器。美国国家标准与技术研究院语音光盘 1-1.1。美国国家航空航天局 *STI/Recon* 技术报告 *n*，93:27403，1993。20，21

[108] Soumya Dutta 和 Sriram Ganapathy。利用内容和声学表征进行语音情感识别。*IEEE* 音频、语音与语言处理汇刊，2025年。24

[109] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现稳健的语音识别。载于《国际机器学习会议》，第 28492–28518 页。PMLR，2023年。24, 32, 33, 39, 67, 68, 87, 106[110] Yuanchao Li 等人。自动语音识别与情感语音：对语音识别和情感识别相互影响的词级研究。arXiv 预印本 arXiv:2305.16065，2023年。24, 39[111] Shuai Fan 等人。面向情感分析的情感感知词级与句级预训练。载于 EMNLP，第 4984–4994 页，2022年。25[112] Varun Krishna 和 Sriram Ganapathy。通过解耦迈向语音表征学习的新前沿。arXiv 预印本 arXiv:2407.02543，2024年。25[113] Wenyi Yu 等人。连接语音编码器与大语言模型以实现自动语音识别。载于 ICASSP，第 12637–12641 页。IEEE，2024年。26[114] Yassir Fathullah 等人。利用具备语音识别能力的大语言模型进行提示。载于 ICASSP，第 13351–13355 页。IEEE，2024年。26[115] Changli Tang 等人。SALMONN：迈向大语言模型的通用听觉能力。arXiv 预印本 arXiv:2310.13289，2023年。26, 30, 31, 32, 54, 67, 68, 107[116] Yingting Li 等人。在语音理解的 sure 基准上评估参数高效迁移学习方法。载于 ICASSP，第 1–5 页。IEEE，2023年。26[117] Kwangyoun Kim 等人。用于语音识别的卷积增强的参数高效微调。载于 Interspeech，第 2830–2834 页，2024年。26[118] Shu wen Yang 等人。SUPERB：语音处理通用性能基准。载于 Interspeech，第 1194–1198 页，2021年。27, 29, 97[119] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于 ICLR。OpenReview.net，2019年。29, 71, 86[120] Vinod Nair 和 Geoffrey E Hinton。修正线性单元改进受限玻尔兹曼机。载于 ICML，第 807–814 页，2010年。29

[121] Paul-Ambroise Duquenne、Holger Schwenk 和 Benoˆıt Sagot。SONAR：句子级多模态和语言无关表示。*arXiv* 电子印本，第 arXiv–2308 页，2023年。30，106

[122] Srinivas Parthasarathy 和 Carlos Busso。通过多任务学习联合预测唤醒度、效价和支配度。在 *Interspeech*，第 2017年 卷，第 1103–1107 页，2017年。31[123] Soumya Dutta 和 Sriram Ganapathy。用于多模态情感识别的 HCAM——层次式交叉注意力模型。*arXiv* 预印本 *arXiv:2304.06910*，2023年。33，60，61[124] Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。监督式对比学习。神经信息处理系统进展，33:18661–18673，2020年。49[125] Soumya Dutta、Smruthi Balaji 和 Sriram Ganapathy。用于对话中多模态情感识别的专家混合模型。*Computer Speech & Language*，100:101965，2026年。52[126] Parishad BehnamGhader、Vaibhav Adlakha、Marius Mosbach、Dzmitry Bahdanau、Nicolas Chapados 和 Siva Reddy。LLM2Vec：大语言模型其实是强大的文本编码器。在 第一届语言建模会议，2024年。53

[127] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。LoRA：大语言模型的低秩适配。*ICLR*，1(2):3，2022年。 53，71

[128] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型引导语言-图像预训练。在 国际机器学习会议，第 19730–19742 页。PMLR，2023年。54，68[129]Christian Szegedy、Wei Liu、Yangqing Jia、Pierre Sermanet、Scott Reed、Dragomir Anguelov、Dumitru Erhan、Vincent Vanhoucke 和 Andrew Rabinovich。通过卷积不断深入。在 *IEEE*计算机视觉与模式识别会议论文集，第 1–9 页，2015年。55[130] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Doll´ar。用于密集目标检测的焦点损失。在*IEEE*国际计算机视觉会议论文集，第 2980–2988 页，2017年。56，69

[131] Wei Ai、Yuntao Shou、Tao Meng 和 Keqin Li。DER-GCN：用于多模态对话情感识别的对话与事件关系感知图卷积神经网络。*IEEE*神经网络与学习系统汇刊，2024年。59，60[132] Yuntao Shou、Wei Ai、Jiayi Du、Tao Meng 和 Haiyan Liu。用于对话中多模态情感识别的高效长距离潜在关系感知图神经网络。*CoRR*，2024年。59，60

[133] Yuntao Shou、Tao Meng、Fuchen Zhang、Nan Yin 和 Keqin Li。重新审视基于广义状态空间模型和概率引导融合的多模态情感学习。*CoRR*，2024年。59，60

[134] Chengwen Zhang、Yaohui Liu 和 Bo Cheng。用于多模态情感识别的 MoE 多模态图注意力网络框架。载于 *ICASSP 2025-2025 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，第 1–5 页。IEEE，2025年。59，60，61

[135] Tao Shi 和 Shao-Lun Huang。MultiEMO：一种基于注意力、具备相关性感知能力的多模态融合框架，用于对话情感识别。载于 计算语言学协会第 *61* 届年会论文集（第 *1* 卷：长论文），第 14752–14766 页，2023年。60，61

[136] Bobo Li、Hao Fei、Lizi Liao、Yu Zhao、Chong Teng、Tat-Seng Chua、Donghong Ji 和 Fei Li。重新审视对话多模态情感识别中模态与上下文的解耦和融合。载于第*31*届*ACM*国际多媒体会议论文集，第5923–5934页，2023年。60

[137] Soumya Dutta、Smruthi Balaji、Viveka Salinamakki、Sriram Ganapathy 等人。Abhinaya——一个用于自然条件下语音情感识别挑战赛的系统。载于 *Proc.Interspeech 2025*，第 4663–4667 页，2025年。67[138] Koji Okabe、Takafumi Koshinaka 和 Koichi Shinoda。用于深度说话人嵌入的注意力统计池化。载于 *Interspeech 2018*，第 2252–2256 页，2018年。67[139] Sanyuan Chen 等人。BEATs：使用声学分词器进行音频预训练。载于 国际机器学习大会，第 5178–5193 页。PMLR，2023年。68[140] Hugo Touvron 等人。LLaMA：开放且高效的基础语言模型。 *arXiv* 预印本 *arXiv:2302.13971*，2023年。68

[141] Abhimanyu Dubey 等人。LLAMA 3 模型集。arXiv 预印本 arXiv:2407.21783，2024年。xiv，68，69，73[142] Wenxuan Zhang 等人。大语言模型时代的情感分析：现实检验。载于计算语言学协会 NAACL 2024年会议论文集，3881–3906 页，2024年。68[143] Soumya Dutta 和 Sriram Ganapathy。用于对话多模态情感识别的大语言模型监督预训练。arXiv 预印本 arXiv:2501.11468，2025年。68[144] Abinay Reddy Naini、Lucas Goncalves、Ali N. Salman、Pravin Mote、Ismail R. Ulgen、Thomas Thebaud、Laureano Moro Velazquez、Leibny Paola Garcia、Najim Dehak、Berrak Sisman 和 Carlos Busso。Interspeech 2025年自然条件下语音情感识别挑战赛。载于 Interspeech 2025，第 4668–4672 页，2025年。71[145] Daya Guo 等人。DeepSeek-r1：通过强化学习激发大语言模型的推理能力。arXiv 预印本 arXiv:2501.12948，2025年。xiv，73[146] Josh Achiam 等人。GPT-4 技术报告。arXiv 预印本 arXiv:2303.08774，2023年。xiv，73[147] Soumya Dutta、Avni Jain 和 Sriram Ganapathy。结合音高与时长风格迁移的音频到音频情感转换。arXiv 预印本 arXiv:2505.17655，2025年。78[148] Adam Polyak 等人。基于离散解耦自监督表征的语音再合成。载于 Interspeech，第 3615–3619 页，2021年。79，86[149] Benjamin Van Niekerk 等人。用于改善语音转换的离散与软语音单元比较。载于 IEEE（ICASSP），第 6562–6566 页，2022年。79，85，94[150] Brecht Desplanques、Jenthe Thienpondt 和 Kris Demuynck。ECAPA-TDNN：基于时延神经网络的说话人验证中的强化通道注意力、传播与聚合。载于 Interspeech，第 3830–3834 页，2020年。80，85，95[151] Raghavendra Pappagari 等人。x向量与情感相遇：情感与说话人识别之间依赖关系的研究。载于 IEEE（ICASSP），第 7169–7173 页，2020年。80[152] Zein Shaheen 等人。利用说话人嵌入中的情感信息实现富有表现力的文本转语音。载于 INTERSPEECH，第 2038–2042 页，2023年。80

[153] Tao Li 等人。用于端到端语音合成的跨说话人情感解耦与迁移。*IEEE/ACM* 音频、语音与语言处理汇刊，30：1448–1460，2022年。80[154] Yaroslav Ganin 等人。神经网络的域对抗训练。《机器学习研究杂志》，17(1)：2096–2030，2016年。80[155] Daniel Erro 等人。基于韵律单元选择的情感转换。*IEEE*音频、语音和语言处理汇刊，18(5)：974–983，2009。81[156] Kavita Kasi 和 Stephen A Zahorian。另一种音高跟踪算法。载于*IEEE*（*ICASSP*），第1卷，I–361页，2002年。82

[157] Gallil Maimon 和 Yossi Adi。基于离散自监督单元的波形域说话风格转换。载于计算语言学协会论文集：*EMNLP 2023*，8048–8061页，2023年。82，95

[158] Ann Lee 等人. 基于离散单元的直接语音到语音翻译。载于第60届计算语言学协会年会论文集（第1卷：长论文），第3327–3339页，2022年。82[159] Ann Lee 等人. 基于真实数据的无文本语音到语音翻译。载于 NAACL:人类语言技术论文集，第860–872页，2022年。82[160] Felix Kreuk 等人. 使用离散且分解的表征进行无文本语音情感转换。载于 EMNLP论文集，第11200–11214页，2022年。82[161] Sang-gil Lee 等人. BigVGAN：采用大规模训练的通用神经声码器。载于第十一届国际学习表征会议，2023年。83, 84, 86[162] Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur. Librispeech：一个基于公有领域有声读物的自动语音识别语料库。载于2015年IEEE声学、语音与信号处理国际会议（ICASSP），第5206–5210页。IEEE，2015年。85[163] Yinghao Aaron Li、Ali Zare 和 Nima Mesgarani. StarGANv2-VC：一种多样化、无监督、非平行的自然语音转换框架。载于 Interspeech，第1349–1353页，2021年。88[164] Lo¨ıc Barrault 等人. Seamless：多语言富有表现力的流式语音翻译。arXiv预印本 arXiv:2312.05187，2023年。87

[165] Laurens Van der Maaten 和 Geoffrey Hinton. 使用 t-SNE 可视化数据。《机器学习研究杂志》，9(11)，2008。95

[166] Yingzhi Wang、Mirco Ravanelli 与 Alya Yacoubi。语音情感分段：情感何时出现？载于*2023* 年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），第 1–7 页。IEEE，2023年。103