doi:10.3969/j.issn.1673-9833.2025.06.005

融合 MFCC 和 Wav2vec 特征的对话情感识别方法

刘旭东[1]，王坤侠[1, 2]

（1. 安徽建筑大学 电子与信息工程学院，安徽 合肥 230601；
2. 安徽省古建筑智能感知与高维建模国际联合研究中心，安徽 合肥 230601）

摘　要：针对语音信号传统手工特征动态信息捕捉不足的问题，引入 Wav2vec 2.0 模型以提取语音信号中的长距离依赖关系，并通过特征融合方式得到充分的情感特征表示。通过提取语音信号中最具有代表性的 MFCC 特征，且用 Wav2vec 提取特征以弥补 MFCC 在动态信息捕捉上的不足，获得了更为丰富和具有代表性的语音情感特征。通过对交叉注意力机制的运用，将语音声学特征与上下文信息进行融合，以获得更加全面和准确的特征表示。最终，通过 Transformer 网络实现了对情感状态的精准预测。通过在 MELD 和 EEIDB 数据集上进行实验，得知本文提出的方法在加权 F_1-Score 指标上分别达到了 44.32% 和 65.50%，从而验证了其有效性和优越性。

关键词：手工特征；长距离；特征融合；上下文信息；对话情感识别；梅尔频率倒谱系数

中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1673-9833(2025)06-0029-08

引文格式：刘旭东，王坤侠 . 融合 MFCC 和 Wav2vec 特征的对话情感识别方法 [J]. 湖南工业大学学报，2025，39(6)：29-36.

An Emotion Recognition Method in Conversation Integrating MFCC and Wav2vec Features

LIU Xudong[1]，WANG Kunxia[1, 2]

（1. School of Electronic and Information Engineering，Anhui Jianzhu University，Hefei 230601，China；
2. Anhui International Joint Research Center for Ancient Architecture Intellisencing and Multi-Dimensional Modeling，Hefei 230601，China）

Abstract：In view of the issue of insufficient capture of dynamic information in traditional manual features of speech signals, Wav2vec 2.0 model is introduced for an extraction of long-range dependencies in speech signals, thus obtaining sufficient emotional feature representations based on feature fusion. By extracting the most representative MFCC features from speech signals, with Wav2vec adopted to compensate for the lack of MFCC in capturing dynamic information, richer and more representative speech emotion features can be obtained. By utilizing the cross attention mechanism, the acoustic features of speech are integrated with contextual information so as to obtain a more comprehensive and accurate feature representation. Consequently, an accurate prediction of emotional states can be achieved through Transformer networks. Through experiments on MELD and EEIDB datasets, it is found that the proposed method achieves 44.32% and 65.50% in weighted F_1-Score metrics, respectively, which verifies its effectiveness and superiority in performance.

收稿日期：2024-06-14

基金项目：安徽省住房城乡建设科学技术计划基金资助项目（2023-YF004，2023-YF113）；安徽建筑大学智能建筑与建筑节能安徽省重点实验室开放课题基金资助项目（IBES2022ZR02）

作者简介：刘旭东，男，安徽建筑大学硕士生，主要研究方向为语音与对话情感识别，E-mail：xdliu@stu.ahjzu.edu.cn

通信作者：王坤侠，女，安徽建筑大学副教授，博士，主要研究方向为人工智能，情感计算等，E-mail：kxwang@ahjzu.edu.cn

Keywords：work feature；long range；feature fusion；contextual information；emotion recognition in conversation(ERC)；Mel-scale frequency cepstral coefficients(MFCC)

随着人工智能和自然语言处理技术的迅速发展，对话情感识别（emotion recognition in conversation, ERC）任务受到更多关注。ERC 旨在通过分析对话内容，识别并理解对话参与者的情感状态。这项技术在智能客服、心理健康监测、社交媒体分析、教育系统和虚拟助手等领域具有广泛应用。

在智能客服领域，ERC 能帮助系统实时感知客户情绪，提供更加贴心和个性化的服务。在心理健康监测领域，分析对话情感有助于评估和干预心理问题。在社交媒体分析领域，ERC 可用于监测用户情绪倾向，为企业和政府决策提供支持。教育系统利用 ERC 识别学生情感状态，优化教学策略，提升学习效果。虚拟助手和智能家居通过 ERC 技术变得更加智能和人性化，能够更好地回应用户需求。

准确的情感识别不仅能提升人机交互体验，使之更加自然顺畅，还能为情感计算和分析提供强有力的支持。ERC 技术通过深入理解对话情感，推动相关领域的创新发展，满足日益增长的个性化需求，带来更加智能和贴心的用户体验。

1 相关工作

对话情感识别的研究 [1] 涉及多种技术和方法。传统的情感识别方法主要依赖于手工特征和机器学习方法 [2-3]，例如早期的特征提取方法包括短时傅里叶变换（short-time Fourier transform，STFT）、线性预测编码（linear predictive coding，LPC）等，但是这些方法都存在一定的缺陷。目前，使用较为广泛的特征提取方法包括梅尔频率倒谱系数（Mel-scale frequency cepstral coefficients，MFCC）和倒谱分析（cepstral analysis）[4] 等。

近年来，随着深度学习技术的兴起，情感识别也迎来了新的发展。深度学习方法能够自动从语音数据中学习特征，例如基于卷积神经网络（convolutional neural networks，CNN）、递归神经网络（recurrent neural networks，RNN）以及注意力机制（attention mechanism）模型 [5-7]，在情感识别任务中都表现出了良好的性能。这类方法避免了手工特征设计的复杂性，能够自动提取更为丰富和有效的特征。此外，还有自监督学习方法，如 Wav2vec 和 Wav2vec 2.0[8-9] 技术，通过在大量未标注的语音数据上进行预训练，学习语音的潜在特征表示。

S. P. Mishra 等 [4] 尝试使用基于 MFCC 的熵特征来提高情感分类的准确率。Sun C. J. 等 [10] 则对 Wav2vec 2.0 模型进行了微调，从中提取情感嵌入，并预训练情感特征提取器，取得了显著的效果。M. Khan 等 [11] 利用 CNN 处理原始的语音信号和文本数据，并将它们馈送到相应的编码器进行判别和语义特征提取。

在模型网络方面，Liang X. W. 等 [12] 利用 LSTM（long short-term memory）网络的优势，提出了一种称为说话人信息增强的长短时记忆（SL-LSTM）的循环结构，用于跟踪不同说话人的情绪状态，从而增强对话中的情绪学习。Lian Z. 等 [13] 使用 Transformer 网络提出了一个半监督多模态交互网络，利用层次递归网络和混合融合策略整合多模态特征，并在语音和文本特征下进行了验证。Luo J. C. 等 [14] 则结合语音和相应的转录文本进行了情感检测，采用平行自注意和交叉注意力机制动态捕捉音频和文本模态间及模态内的相互作用，构建了中级融合和残差模块，用于建模长期上下文依赖关系。

此外，多模态情感识别的发展 [15] 也极其迅速，这些方法综合了文本、语音、图像等多种模态，用来提高情感识别的准确性和鲁棒性。尽管对话情感识别研究现已取得了显著的进展，但是仍然面临着诸多挑战 [16]。例如对话需要考虑上下文的依赖性和连续性，如何有效对对话的上下文信息建模是一个关键问题。还有，情感表达具有高度的主观性和多样性，不同的个体对同一情感的表达方式可能存在明显的差异。另外，现实情境中具体的对话数据通常存在噪声和不完整性，这也就导致了对情感识别模型的鲁棒性提出了更高要求。

本文在多模态情感线数据集（MELD）[17] 和老人情感交互数据集（EEIDB）[18] 上展开实验，旨在通过处理音频文件进行语音数据下的对话情感识别任务。实验主要采用 Transformer 网络，并使用 librosa 工具提取语音数据的声学特征。然后利用预训练的 Wav2vec 2.0 模型，从原始音频波形中学习有用的包含动态信息和上下文关联的特征表示。本文合理地利用交叉注意力机制，通过改进交叉注意力后的注意力矩阵计算两类特征的权重，并且进行特征融合。最后，

使用 Transformer 模型网络进行分类，验证前面工作的有效性和鲁棒性。实验结果表明，本文提出的模型在 MELD 和 EEIDB 数据集上均取得了优异的表现。具体的模型图如图 1 所示。

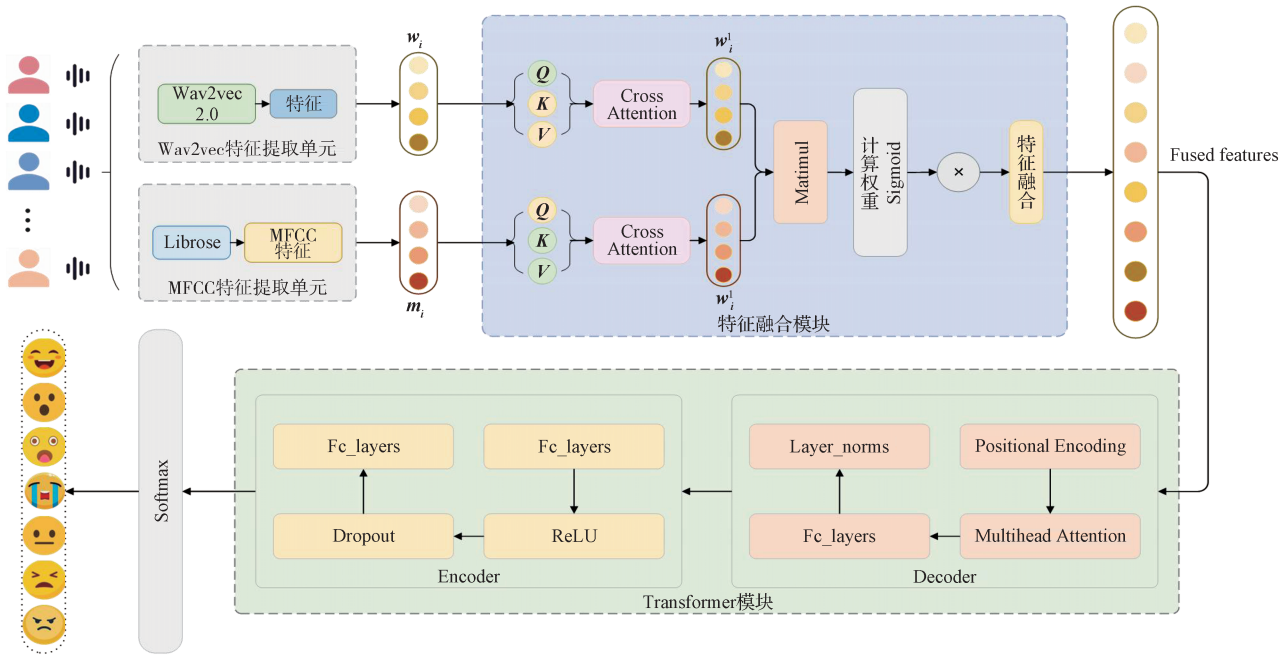

图 1　模型整体结构图

Fig. 1　Overall structure diagram of the model

2　融合 MFCC 和 Wav2vec 特征模型

在对话情感识别任务中，对话信息由一组元组组成，表示为$\{(u_1, p_1), (u_2, p_2), \cdots, (u_n, p_n)\}$。其中，$p_i$ 为当前的说话人，u_i 为当前说话人所说的话语，并且每个 u_i 都对应一个情感标签 y_i。每个对话可能包含两个或多个说话人，并且每个说话人的话语数量可能不同。对话情感识别任务不仅要获得语音信号中丰富的情感特征，还要结合上下文关系获得更加充分的上下文特征表示。

2.1　特征提取

2.1.1　MFCC 特征提取

MFCC 是一种被广泛应用于语音识别的特征提取方法，它结合人类听觉系统的特性，提取出对语音信号最具代表性的特征。MFCC 通过将语音信号映射到梅尔频率尺度，并使用离散余弦变换（DCT）来提取倒谱系数，具有较好的抗噪能力和鲁棒性，能够有效应对不同环境下的语音数据。此外，MFCC 特征能够较好地模拟人耳对不同频率的感知能力，使得情感识别系统在面对复杂的语音数据时，仍能提取出关键的情感信息。

具体而言，预加重通过滤波器增强高频成分，补偿语音信号中的能量衰减；分帧与加窗将语音信号分割成小的时间窗口（通常为 20~30 ms），并对每帧加窗以减少频谱泄漏。快速傅里叶变换（FFT）将时域信号转换为频域信号，获取频谱信息。梅尔滤波器组使用一组梅尔滤波器对频谱进行加权，生成梅尔频率能量谱。对数运算对能量谱进行对数变换，然后通过 DCT 提取倒谱系数，得到最终的 MFCC 特征。

本研究采用 Librosa 工具 [19] 提取维度为 300 的 MFCC 特征，以充分捕捉语音数据中的情感特征。Librosa 作为一个强大的音频处理库，不仅简化了 MFCC 特征的提取过程，还保证了提取特征的一致性和高效性。通过提取 300 维的 MFCC 特征，能够更加详细地表征语音信号的频谱信息，增加提取出来的特征的有效性。其中，MFCC 特征的提取过程如图 2 所示。

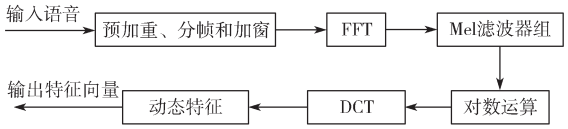

图 2　MFCC 特征提取流程

Fig. 2　MFCC feature extraction process

2.1.2　Wav2vec 2.0 特征提取

Wav2vec 2.0 被视为语音处理领域的一项重要突破，其自监督学习方法赋予了它独特的优势。该技术通过从大规模未标记的音频数据中学习特征表示，能够利用深度学习模型自动掌握最佳的特征表达方式，这为语音识别以及其他相关任务提供了强大的支持。其所具备的能力不仅限于简单地提取特征，而且能够深入理解音频信号背后的语义和语境。

Wav2vec 2.0 的突出之处在于其自监督学习能力和对长距离依赖的捕获能力。通过对未标记音频数据进行对比学习，模型可以自动学习最优的特征表示方式，无需人工标注，大幅减少了数据标注的成本和工作量。此外，模型还能够捕捉音频信号中的长距离依赖关系，充分表征语音数据中丰富和深层的上下文信息，从而更好地理解和分析语音内容。另外，其还可以直接从原始音频中提取特征，无需依赖人工设计的复杂特征，因此避免了手工特征设计带来的麻烦和不确定性，使得所提取的特征更加全面、有效且具有普适性。

本文中，课题组利用预训练的 Wav2vec 2.0 模型（见图 3）提取了音频数据的 768 维特征。这些特征不仅包含了充分的音频原始信息，还潜藏着丰富的情感特征。通过这些特征，能够更准确地捕捉到语音数据中蕴含的动态上下文信息和情感信息，为后续的情感识别任务提供强有力的支持和保障。

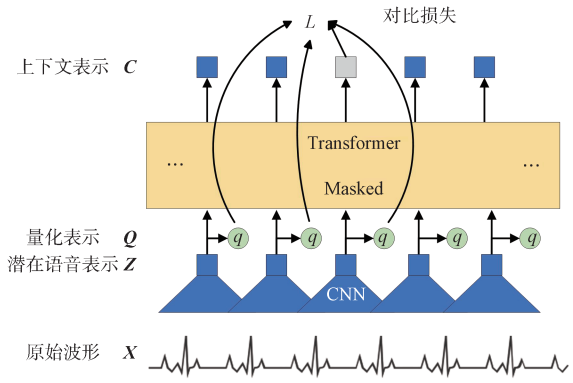

图 3 Wav2vec 2.0 模型结构示意图

Fig. 3 Wav2vec 2.0 model structure schematic diagram

2.2 改进交叉注意力机制的特征融合

交叉注意力机制（cross attention mechanism）[20] 在深度学习中具有重要作用，它超越了单一的自注意力机制，通过精确地计算不同模态或序列之间的相关性，将多种信息源巧妙地融合在一起，从而显著提升模型的性能和表现能力。本研究中，将这一机制应用于两种不同类型的特征，利用其强大的信息融合能力，成功地实现了两类信息的完美融合。

具体而言，使用交叉注意力机制进行特征融合时，查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$、值矩阵 $\boldsymbol{V}$ 的来源通常是不同模态的输入特征。本研究中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵来源于两种不同类型的输入特征，如若 $\boldsymbol{Q}$ 为 MFCC 类型的特征，则 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 为 Wav2vec 特征类型的特征。当两类特征分别输入交叉注意力机制后，会分别生成对应的注意力矩阵。本文通过对输出的注意力矩阵进行处理，借助多层感知机（multilayer perceptron，MLP）网络和 Sigmoid 激活函数，进一步计算出两类特征在总体中所占的权重。随后，通过这些权重应用于这两种类型特征，实现两类特征的进一步融合。

计算输出两类特征之间的权重，能够更深入地挖掘特征之间的关联关系，进而实现对交叉注意力输出的两类特征进行更加充分的融合。其中，交叉注意力计算公式如下：

$$CrossAttention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d_K}\right)\boldsymbol{V}, \tag{1}$$

式中 d_K 为键向量的维度。

本文引入了 Wav2vec 2.0，这不仅丰富了手工特征，同时也弥补了 MFCC 在捕捉动态变化信息方面的不足。Wav2vec 的引入能够更全面地捕捉到语音数据的上下文特征，从而为情感识别任务提供更为准确和丰富的特征表示。通过融合这些特征表示，得到更丰富、更具代表性的语音声学特征和上下文信息的表示形式，这将为后续的情感识别任务提供更为有力的支持和保障。最终，将这些融合后的特征输入 Transformer 模型网络，以预测相关情感类别。权重计算公式如下：

$$Weighted = sigmoid(\boldsymbol{Wx}+B), \tag{2}$$

$$\boldsymbol{F\_features} = M\_Weighted\times\boldsymbol{M\_features}+W\_Weighted\times\boldsymbol{W\_features}。 \tag{3}$$

式（2）（3）中：$\boldsymbol{Wx}$ 为输入向量 $\boldsymbol{x}$ 与权重矩阵 $\boldsymbol{W}$ 的矩阵乘法；B 为添加的偏置项；$\boldsymbol{F\_features}$ 为融合后的特征，$M\_Weighted$ 为计算得出的 MFCC 特征权重；$W\_Weighted$ 为计算出的 wav2vec 特征权重；$\boldsymbol{M\_features}$ 为 MFCC 特征矩阵；$\boldsymbol{W\_features}$ 为 Wav2vec 特征矩阵。

2.3 Transformer 网络

融合后的特征表示不仅简单地结合了不同类型的信息，而且包含了更加丰富和有代表性的语音数据表征。研究过程中，课题组意识到对话情感识别任务需要考虑到对话的上下文关系和历史情感状态，因此将 Transformer 网络作为预测模型。Transformer 模型有其独特的自注意力机制，能够捕捉到长距离的依赖关系，因此在处理长序列和复杂语义关系的任务中表现出色。本研究选择 Transformer 模型的一个重要原因，是它能够更好地理解对话信息中的连续性和上下文关系。

与传统的 RNN 模型相比，Transformer 模型的多头注意力机制实现了并行计算，这一特性极大地提高了模型的训练效率和推理速度。基于其并行计算能力，Transformer 模型能够轻松处理更大规模的模型和更长序列的数据，从而提供了更丰富、更深入的情

感识别能力。因此，本研究选择 Transformer 网络作为预测模型，它能够更好地满足对话情感识别任务的需求，可为情感分析领域的研究和应用带来更为深远的影响。

2.4　损失函数

为了解决对话数据集中样本不均衡的问题，本研究中采用加权交叉熵损失函数（weighted-cross entropy loss）[21] 进行实验。加权交叉熵在传统交叉熵损失函数的基础上引入了每个类别的权重，以更好地调整不同类别之间的重要性。

在加权交叉熵损失函数中，为每个类别分配了特定的权重，这些权重反映了该类别在整个数据集中的重要性。通过这种方式，模型在训练过程中更加关注那些在数据集中比例较小但重要性更高的类别，从而提高了模型对不同类别的敏感度。其中，加权交叉熵损失函数的计算公式如下：

$$L=-\sum_{i=1}^{C} w_i \cdot \boldsymbol{y}_i \cdot \log\left(\hat{y}_i\right)。\tag{4}$$

式中：$\boldsymbol{y}_i$ 为真实标签的 one-hot 编码向量（即对于类别 i，$\boldsymbol{y}_i$ 取值为 1，其它类别取值为 0）；$\hat{y}_i$为模型的预测概率，即类别 i 的预测概率；w_i 为类别 i 的权重；C 为类别总数。

在实际计算损失函数时，每个样本的损失值都会根据其所属类别的权重进行调整。因此可以自定义分配不同的权重，以更精确地反映不同类别样本的重要性，进而使模型能够更好地适应数据集的特征分布。这种加权处理的策略不仅优化了模型的训练效果，还增强了模型的泛化能力，使其在真实场景中的表现更加稳健。

2.5　评价指标

为了更全面地验证模型的有效性，本研究采用加权 F_1 值（weighted_F_1，$F_{1\text{-w}}$）和准确率 [22] 作为评估指标。加权 F_1 值能够综合考虑不同类别的不平衡性，通过平衡精确率和召回率来提供更全面的性能评估；而准确率则提供了对整体分类效果的直观衡量。这两种评估指标的结合，确保能够全方位地评估模型在各种情感分类任务中的表现，从而验证模型在实际应用中的有效性和鲁棒性。

加权 F_1 值是所有类别的 F_1 值按其权重进行加权平均的结果，其中，N_i 为类别 i 的样本数，N 为总样本数，P_i 为精确率，R_i 为召回率。

$$F_{1\text{-w}}=\sum_{i=1}^{C}\frac{N_i}{N}\cdot\frac{2\cdot P_i\cdot R_i}{P_i+R_i}。\tag{5}$$

准确率（accuracy）是评估分类模型性能的基本指标之一，为模型预测正确的样本数量 N_{correct} 占总样本数量 N 的比例，其计算式如下：

$$Accuracy=N_{\text{correct}}/N。\tag{6}$$

3　实验

3.1　数据集

3.1.1　MELD 数据集

MELD 数据集 [17] 是从电影《老友记》上摘取的片段，构成一个多模态数据集，既包括文本信息，也包括与文本对应的音频和视频信息。MELD 数据集包括 1 443 段对话，总共 13 708 句，平均每段对话有 9.5 句话，每句话平均持续时间为 3.6 s，包含生气、厌恶、悲伤、快乐、中性、惊喜和恐惧 7 种情感。与此同时，每句话也拥有相应的情感标签，分为积极、消极和中性 3 类。

3.1.2　EEIDB 数据集

EEIDB 数据集 [18] 是关于老年人情感交互的数据库。该数据集收集了 14 位演员在电视剧《空巢姥爷》剧情中的对话，包括视频和音频信息，包含愤怒、开心、悲伤、中立等情感。该数据集包括从 118 个对话中提取的 810 个语音视频摘录。情绪注释由 18 名参与者（10 名男性和 8 名女性）进行评估。

此外，这些数据库还记录了演员在日常对话中的情绪转变，数据集的具体信息见表 1。

表 1　数据集的具体信息

Table 1　Specific information of the dataset

Dataset	#Conversations			#Utterances			#Classes
	Train	Test	Val	Train	Test	Val	
MELD	1 039	280	114	9 989	2 610	1 109	7
EEIDB	77	26	13	824	227	124	4

本实验中只使用 MELD、EEIDB 数据集中的音频信息。MELD 使用数据集本身划分好的训练集、验证集和测试集进行实验，数据分别为 9 989, 1 109, 2 610 个话语。EEIDB 按 7:1:2 的比例划分为训练集、验证集和测试集。由于每个类别分布极其不均衡，故使用加权 F_1 值来评估实验结果。

3.2　实验环境

本实验以对话音频信息为训练数据，每轮将完整的对话内容输入模型进行训练。实验环境为 Windows 系统，所采用的 GPU 型号为 GeForce RTX 4060。在 PyTorch 环境下，设定学习率为 0.000 1，权值衰减率为 5E−4，优化器选用 AdamW，损失函数采用加权交叉熵。最后，采用 Softmax 函数进行情感分类，并使用加权 F_1 值（$F_{1\text{-w}}$）作为评估指标，以验证模型的有效性。

3.3 验证实验结果

本实验在上述参数设置下，在 MELD 和 EEIDB 数据集上进行验证。其中，在 MELD 数据集上加权 F_1 值为 44.32%，准确率为 46.21%；在 EEIDB 数据集上加权 F_1 值为 65.50%，准确率为 71.28%。

通过与 EmbraceNet(EmbraceNet 体系结构网络）、DialogueRNN（对话框架网络）、MM-DFN（多模态动态融合网络）和 SCA（交叉注意单元）等模型在 MELD 数据集上进行对比（实验结果见表 2），得知本文模型对语音模态下情感的分类性能有所提升。相比于 MM-DFN 提升了 1.6%，相比于 SCA 提升了 1.22%。以上结果表明，本文提出的模型具有一定的优越性。

在 MELD 数据集中，各类情感的预测结果如表 3 所示。分析表中数据可知，由于该数据集中情感类别分布不均衡，Neutral 情感的样本数量远多于其他情感的样本数量，而 Fear 情感和 Disgust 情感的样本数量则相对较少。这种不平衡性增加了对后两类情感进行有效训练的难度。通过对比表 3 中各模型的具体预测结果也验证了这一问题，可以明显地看出这一问题对模型性能的影响。另外，EEIDB 数据集中，得到的各个类型情感的预测结果如下：Neutral 为 71.43%，Anger 为 37.50%，Sadness 为 83.33%，Happy 为 44.44%。

表 2 不同模型下的 F_{1-w} 值对比

Table 2 Comparison F_{1-w} values under different models

模型	EmbraceNet[23]	DialogueRNN[6]	MM-DFN[24]	SCA[14]	Our Model
F_{1-w}/%	40.00	41.79	42.72	43.10	44.32

表 3 不同模型下的具体识别率对比

Table 3 Comparison of specific recognition rates under different models %

Model	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	F_1 Score
bcLSTM	61.86	19.34	2.90	14.71	15.74	6.06	25.85	39.08
EmbraceNet	48.50	16.70	0.00	0.00	45.50	0.00	60.00	40.00
DialogueRNN	65.57	20.47	5.56	14.01	13.17	5.13	35.18	41.79
SCA	62.90	20.12	0.00	6.70	22.65	0.00	53.21	43.10
Our Model	69.08	29.18	2.00	19.71	30.60	1.47	30.43	44.32

3.4 消融实验结果

为验证本文提出的特征融合方法的有效性，接下来进行消融实验。表 4 展示了单独使用 MFCC 特征、Wav2vec 特征以及两者融合后的实验结果。

表 4 不同特征下的消融实验结果对比

Table 4 Comparison of ablation experimental results under different features %

数据类型	F_{1-w} 值（MELD/EEIDB）	准确率（MELD/EEIDB）
Audio MFCC	39.03/62.23	40.36/66.47
Wav2vec 2.0	40.48/59.83	42.08/65.34
Audio MFCC+Wav2vec 2.0	44.32/65.50	46.21/71.28

表 4 所示 MELD 数据集的实验结果表明，与单独使用 MFCC 特征的结果相比，融合特征在加权 F_1 值上提升了 5.29%，在准确率上提升了 5.85%；与单独使用 Wav2vec 特征的结果相比，融合特征在加权 F_1 值上提升了 3.84%，在准确率上提升了 4.13%。这些结果充分证明了融合后的特征包含了更丰富的语音声学特征和上下文信息，能够更好地表征语音数据中的对话信息和情感信息，从而显著提升了实验的准确性和鲁棒性。

可观察到，在处理样本数量较少的情感类别时，模型的预测准确性显著降低，这进一步验证了情感分类不均衡对训练效果的负面影响。为了解决这一问题，课题组采用了加权交叉熵损失函数来验证其在实验中的有效性。实验结果显示，使用加权交叉熵损失函数后，模型的预测性能有了显著的提升。图 4 和图 5 为在 MELD 数据集中使用不同损失函数的混淆矩阵图。

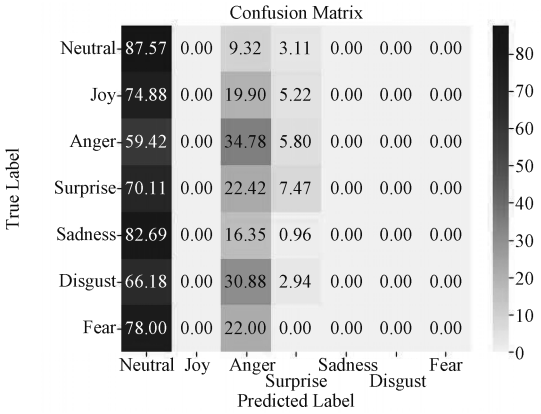

图 4 交叉熵损失函数下的混淆矩阵图

Fig. 4 Confusion matrix diagram under cross-entropy loss function

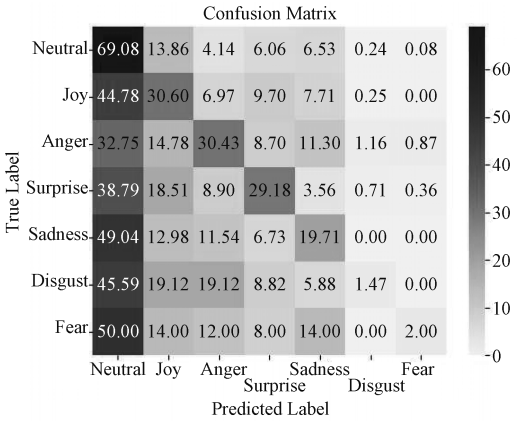

图 5　加权交叉熵下的混淆矩阵图

Fig. 5　Confusion matrix diagram under a weighted cross-entropy

通过对比使用传统交叉熵损失函数和加权交叉熵损失函数的混淆矩阵，可以明显看出，加权交叉熵损失函数通过为不同类别分配权重，使模型更加关注样本比例较少的类别，从而提高了模型对不同类别的敏感度，使结果更加均衡。这样的改进不仅提升了模型在少数类情感上的预测准确性，也增强了整体模型的鲁棒性和可靠性。

4　结语

本文对对话情感识别任务进行了探索和实验。首先，观察到在处理样本数量较少的情感类别时，模型的预测准确性明显降低，这进一步验证了情感分类不均衡对训练效果的负面影响。为了解决这一问题，采用了加权交叉熵损失函数，实验结果证明这种方法能够显著提升模型的预测性能，使其更加均衡和可靠。其次，引入了 Wav2vec 技术，以弥补 MFCC 特征在捕捉动态信息方面的不足，从而获得更丰富和具有代表性的语音情感特征。通过交叉注意力机制的应用，成功融合了语音声学特征和上下文信息，进一步提高了模型的表现能力。最后，使用 Transformer 网络对情感状态进行了预测，取得了良好的实验结果。

在未来的工作中，课题组将进一步探索和优化模型架构，以提高对话情感识别任务的准确性和鲁棒性。此外，还可以考虑更广泛地应用其他自监督学习技术和模型预训练策略，以进一步改进特征表示和模型性能。另外，扩大数据集的规模和多样性也是未来研究的重要方向，这将有助于提高模型的泛化能力和实用性。综上所述，本研究为对话情感识别领域的进一步研究和实践提供了有益的经验和启示。

参考文献：

[1]　陈晓婷，李　实 . 对话情绪识别综述 [J]. 计算机工程与应用，2023，59(3)：33−48.
CHEN Xiaoting，LI Shi. Survey on Emotion Recognition in Conversation[J]. Computer Engineering and Applications，2023，59(3)：33−48.

[2]　NOROOZI F，SAPIŃSKI T，KAMIŃSKA D，et al. Vocal-Based Emotion Recognition Using Random Forests and Decision Tree[J]. International Journal of Speech Technology，2017，20(2)：239−246.

[3]　CHEN L F，SU W J，FENG Y，et al. Two-Layer Fuzzy Multiple Random Forest for Speech Emotion Recognition in Human-Robot Interaction[J]. Information Sciences，2020，509：150−163.

[4]　MISHRA S P，WARULE P，DEB S. Speech Emotion Recognition Using MFCC-Based Entropy Feature[J]. Signal, Image and Video Processing，2024，18(1)：153−161.

[5]　ALLUHAIDAN A S，SAIDANI O，JAHANGIR R，et al. Speech Emotion Recognition Through Hybrid Features and Convolutional Neural Network[J]. Applied Sciences，2023，13(8)：4750−4764.

[6]　MAJUMDER N，PORIA S，HAZARIKA D，et al. Dialogue RNN：An Attentive RNN for Emotion Detection in Conversations[EB/OL]. (2019−05−25)[2024−06−10]. http://arxiv.org/pdf/1811.00405.

[7]　CHEN Z Z，LI J W，LIU H，et al. Learning Multi-Scale Features for Speech Emotion Recognition with Connection Attention Mechanism[J]. Expert Systems with Applications，2023，214：118943−118952.

[8]　BAEVSKI A，ZHOU Y，MOHAMED A，et al. wav2vec 2.0：A Framework for Self-Supervised Learning of Speech Representations[J]. Advances in Neural Information Processing Systems，2020，33：12449−12460.

[9]　曹荣贺，吴晓龙，冯　畅，等 . 基于 Wav2vec2.0 与语境情感信息补偿的对话语音情感识别 [J]. 信号处理，2023，39(4)：698−707.
CAO Ronghe，WU Xiaolong，FENG Chang，et al. Wav2vec2.0 and Context Emotional Information Compensation Based Dialogue Speech Emotion Recognition[J]. Journal of Signal Processing，2023，39(4)：698−707.

[10]　SUN C J，ZHOU Y，HUANG X，et al. Combining Wav2vec 2.0 Fine-Tuning and ConLearnNet for Speech Emotion Recognition[J]. Electronics，2024，13(6)：1103.

[11]　KHAN M，GUEAIEB W，EL SADDIK A，et al. MSER：Multimodal Speech Emotion Recognition Using Cross-Attention with Deep Fusion[J]. Expert Systems with Applications，2024，245：122946−122956.

[12]　LIANG X W，ZOU Y，XU R F. SI-LSTM：Speaker Hybrid Long-Short Term Memory and Cross Modal

Attention for Emotion Recognition in Conversation[EB/OL]. [2024−06−02]. https://arxiv.org/abs/2305.03506v3.

[13] LIAN Z，LIU B，TAO J H. SMIN：Semi-Supervised Multi-Modal Interaction Network for Conversational Emotion Recognition[J]. IEEE Transactions on Affective Computing，2023，14(3)：2415−2429.

[14] LUO J C，PHAN H，REISS J. Cross-Modal Fusion Techniques for Utterance-Level Emotion Recognition from Text and Speech[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island：IEEE，2023：1−5.

[15] 李佳泽，梅红岩，贾丽云，等 . 动态时间序列建模的多模态情感识别方法 [J/OL]. 计算机工程与应用，2023：1−14. [2024−06−05]. http://kns.cnki.net/kcms/detail/11.2127. TP. 20231130. 1701.028.html.
LI Jiaze，MEI Hongyan，JIA Liyun，et al. Multimodal Emotion Recognition Method Based on Dynamic Time Sequence Modeling[J/OL]. Computer Engineering and Applications, 2023：1−14. [2024−06−05]. http://kns.cnki.net/kcms/detail/11.2127.TP. 20231130.1701.028.html.

[16] 郭 续，买日旦·吾守尔，古兰拜尔·吐尔洪 . 基于多模态融合的情感分析算法研究综述 [J]. 计算机工程与应用，2024，60(2)：1−18.
GUO Xu，MAIRIDAN Wushouer，GULANBAIER Tuerhong. Survey of Sentiment Analysis Algorithms Based on Multimodal Fusion[J]. Journal of Computer Engineering & Applications，2024，60(2)：1−18.

[17] PORIA S，HAZARIKA D，MAJUMDER N，et al. MELD：A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations[EB/OL]. [2024−06−02]. 2018: 1810.02508. https://arxiv.org/ abs/1810.02508v6.

[18] WANG K X，ZHU Z B，WANG S D，et al. A Database for Emotional Interactions of the Elderly[C]//2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). Okayama：IEEE，2016：1−6.

[19] BABU P A，SIVA NAGARAJU V，VALLABHUNI R R. Speech Emotion Recognition System with Librosa[C]//2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). Bhopal：IEEE，2021：421−424.

[20] HOU R，CHANG H，MA B，et al. Cross Attention Network for Few-Shot Classification[C]//Advances in Neural Information Processing Systems 32, Volume 6 of 20：32nd Conference on Neural Information Processing Systems (NeurIPS 2019).Vancouver：s.n.，2019：32.

[21] HO Y，WOOKEY S. The Real-World-Weight Cross-Entropy Loss Function：Modeling the Costs of Mislabeling[J]. IEEE Access，2020，8：4806−4813.

[22] CHICCO D，JURMAN G. The Advantages of the Matthews Correlation Coefficient (MCC) over F_1 Score and Accuracy in Binary Classification Evaluation[J]. BMC Genomics，2020，21(1). DOI:10.1186/s12864-019-6413-7.

[23] XIE B J，SIDULOVA M，PARK C H. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion[J]. Sensors，2021，21(14)：4913−4929.

[24] HU D，HOU X L，WEI L W，et al. MM-DFN：Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore：IEEE，2022：7037−7041.

（责任编辑：廖友媛）