Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 905-915
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144215

Hans汉斯

基于改进Transformer的语音情感识别研究

高　瞻，曾金芳，韩　鎏，陈　晨

湘潭大学物理与光电工程学院，湖南　湘潭

收稿日期：2025年3月28日；录用日期：2025年4月23日；发布日期：2025年4月29日

摘　要

随着人工智能技术的迅速发展，语音情感识别作为人机交互领域的关键技术，对于提升交互体验和理解用户意图具有重要意义。传统的语音情感识别方法在特征提取和模型泛化能力上存在一定的局限性。Transformer模型因其强大的自注意力机制，能够有效捕捉长序列数据中的依赖关系，在自然语言处理等领域取得了显著成果，为语音情感识别提供了新的思路。本文基于改进的Transformer模型展开语音情感识别研究，充分利用LSTM在处理长期依赖方面的优势，以及Transformer在捕捉全局依赖和并行处理方面的能力。实验结果表明，这样的改进可以提高单一模型的准确率。

关键词

语音情感识别，LSTM，Transformer

Research on Speech Emotion Recognition Based on the Improved Transformer

Zhan Gao, Jinfang Zeng, Liu Han, Chen Chen

School of Physics and Optoelectronics, Xiangtan University, Xiangtan Hunan

Received: Mar. 28th, 2025; accepted: Apr. 23rd, 2025; published: Apr. 29th, 2025

Abstract

With the rapid development of artificial intelligence technology, speech emotion recognition, as a key technology in the field of human-computer interaction, is of great significance for enhancing the interaction experience and understanding user intentions. Traditional speech emotion recognition methods have certain limitations in feature extraction and model generalization ability. The Transformer model, due to its powerful self-attention mechanism, can effectively capture the dependencies in long-sequence data and has achieved remarkable results in fields such as natural language processing, providing new ideas for speech emotion recognition. This paper conducts research on

文章引用: 高瞻, 曾金芳, 韩鎏, 陈晨. 基于改进 Transformer 的语音情感识别研究[J]. 应用数学进展, 2025, 14(4): 905-915. DOI: 10.12677/aam.2025.144215

speech emotion recognition based on the improved Transformer model, making full use of the advantages of LSTM in handling long-term dependencies and the capabilities of Transformer in capturing global dependencies and parallel processing. The experimental results show that such an improvement can increase the accuracy of a single model.

Keywords

Speech Emotion Recognition, LSTM, Transformer

1. 引言

随着信息技术的飞速发展，人机交互的智能化需求日益迫切。语音情感识别作为实现自然、高效人机交互的关键技术，近年来受到了广泛关注。其通过对语音信号中的韵律特征、声学特征等进行分析，进而推断说话者的情感状态，在智能教育、医疗诊断、智能安防等众多领域展现出巨大的应用价值。

然而，目前语音情感识别技术仍面临诸多挑战。不同个体在表达相同情感时的语音特征存在显著差异，且情感的复杂性与多样性使得准确分类极为困难。同时，数据的不平衡性以及环境噪声干扰等问题，也严重制约了识别系统的性能提升。一般语音情感识别(Speech Emotion Recognition, SER)模型包括三个模块：语音预处理、特征提取和情感识别。

本文旨在探讨语音情感识别中的关键技术，通过创新的算法与模型优化，致力于提高识别准确率。提出基于 Transformer 网络模型，并结合长短期记忆网络(Long Short-Term Memory, LSTM)模型进行语音情感识别。

2. 相关基础知识及研究现状

2.1. 语音情感识别

在语音情感特征提取方面，当前语音情感特征提取主要分为低级特征和深度特征两个类别[1]。低级特征涵盖基音、强度、梅尔频率倒谱系数、线性预测倒谱系数等。而深度特征指的是通过深度网络提取到的特征，它能够学习语音信号中的高级时频特征。在语音情感识别模型研究方面，早期语音情感识别主要运用传统机器学习算法，如隐马尔可夫模型、高斯混合模型等[2]。目前大多采用深度学习网络模型，例如卷积神经网络[3] (Convolutional Neural Networks, CNN)、时序神经网络(Temporal Convolutional Network, TCN)以及 LSTM 等[4]。在语音情感识别研究中存在很多不足，主要包括：(1) 情感具有主观性，这就导致其定义容易出现不一致的情况，(2) 不同数据库的条件存在差异。此外，在整个流程中，特征提取环节里的特征选择对最终结果影响极大。

语音情感识别主要流程如图 1 所示，包括：预处理，特征提取，特征训练，模型测试。

2.2. 语音情感数据库

语音数据是语音信号输入的起始点。与普通语音数据库不同，语音情感数据库收录的是蕴含特定情感的语音数据。一个品质优良的语音情感数据库，对提升语音情感识别的准确率大有助益。在语音情感识别研究中，语音情感库主要分为两种类型：一种是依据特定需求构建的具有针对性的语音情感库；另

一种则是官方提供的公用情感数据库。考虑到自行构建一个有效的语音库流程繁杂，本文采用公用的情感数据库。常见的语音情感数据库如表 1 [5]。

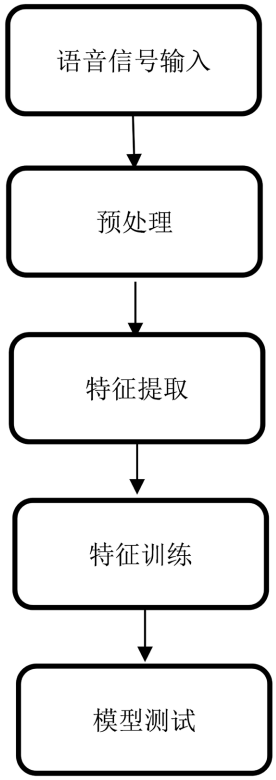

Figure 1. The main links of speech emotion recognition
图 1. 语音情感识别的主要环节

Table 1. Common speech emotion databases
表 1. 常见语音情感数据库

情感数据库	语言	情感
EMODB	德语	厌恶、愤怒、悲伤、高兴、恐惧、无聊、中性
CASIA	普通话	恐惧、惊讶、中性、生气、悲伤、惊讶、高兴
IEMOCAP	英语	惊讶、厌恶、悲伤、高兴、中性、兴奋、愤怒
ACCorpus SR	普通话	愤怒、高兴、悲伤、恐惧、中性
Belfast	德语	愤怒、高兴、悲伤、恐惧、中性

2.3. 特征提取

原始语音信号的数据量庞大且信息冗余，直接用于情感识别效率低下。因此，需要从原始语音信号中提取能够有效表征情感特征的参数，以降低数据维度，提高识别系统性能。

语音情感识别研究中常见的特征提取方法包括：

梅尔频率倒谱系数(MFCC)：基于人类听觉系统的非线性频率感知特性。该方法先将原始语音信号进行分帧加窗处理，然后对每一帧信号进行快速傅里叶变换(FFT)，将时域信号转换到频域，得到频谱。接着，通过一组梅尔滤波器组对频谱进行滤波，将线性频率转换为梅尔频率，模拟人类听觉系统对不同频

率声音的感知特性。之后，对滤波后的结果取对数，并进行离散余弦变换(DCT)，得到 MFCC 系数。MFCC 系数能够有效表征语音信号的频谱包络特征，在语音情感识别中被广泛应用[6]。

线性预测倒谱系数(LPCC)：基于线性预测编码(LPC)理论。该方法假设当前语音样本可以由过去若干个语音样本的线性组合来逼近。首先，通过对原始语音信号进行分帧处理，然后对每一帧信号进行线性预测分析，估计出一组线性预测系数(LPC 系数)，这些系数描述了语音信号的自回归模型。接着，通过对 LPC 系数进行转换，得到 LPCC。LPCC 系数能够反映语音信号的声道特性，在语音情感识别中也具有较好的性能表现。

2.4. 预处理

在语音情感识别研究中，数据预处理是提升识别准确率的关键步骤。常见的数据预处理包括：降噪处理、归一化和数据增强等。

1) 降噪处理

在实际采集的语音数据常混入各种噪声，如环境噪音、设备电子干扰等。这些噪声会干扰语音信号中的情感特征，降低识别准确率。常见的处理方法包括谱减法和小波变换法。

2) 归一化

在语音情感识别中，不同特征维度的数据往往具有不同的尺度和分布范围。例如，某些特征的取值范围可能在 0~1 之间，而另一些特征的取值可能在几百甚至几千。这种数据尺度的差异会导致机器学习模型在训练过程中对不同特征的重视程度不均衡，使得模型更倾向于学习尺度较大特征的信息，而忽略尺度较小特征的重要性，从而影响模型的整体性能和泛化能力。因此，为了消除数据尺度差异对模型训练的影响，需要对语音数据进行归一化处理。

常见的归一化方法包括：

最小－最大归一化(Min-Max Normalization)：也称为离差标准化，是一种简单且常用的归一化方法。该方法的基本思想是将数据集中的每个特征值映射到一个指定的区间，通常是[0, 1]。具体计算公式为：

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

其中，X 是原始特征值，$X_{\min}$ 和 $X_{\max}$ 分别是该特征在数据集中的最小值和最大值，X_{norm} 是归一化后的特征值。

最小－最大归一化方法简单直观，能够有效地保留数据的原始分布特征，并且在数据特征的取值范围已知且相对稳定的情况下，具有较好的归一化效果。然而，该方法对数据中的异常值较为敏感，因为异常值会极大地影响取值，从而导致归一化后的数据偏离其正常分布范围。

Z-分数归一化(Z-Score Normalization)：也称为标准差标准化，是一种基于数据的均值和标准差进行归一化的方法。该方法的核心思想是将数据集中的每个特征值转换为以均值为中心，以标准差为度量单位的标准化值，使得归一化后的数据具有零均值和单位标准差。具体计算公式为：

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{2}$$

其中，X 是原始特征值，μ 是该特征在数据集中的均值，σ 是该特征在数据集中的标准差，X_{norm} 是归一化后的特征值。

Z-分数归一化方法能够有效地消除数据的量纲影响，使得不同特征维度的数据具有可比性，并且在数据特征的分布较为稳定且近似服从正态分布的情况下，具有较好的归一化效果。此外，该方法对数据

中的异常值具有一定的鲁棒性，因为异常值在计算均值和标准差时会被平均化处理，从而减少了异常值对归一化结果的影响。然而，Z-分数归一化方法在数据特征的分布严重偏离正态分布时，可能会导致归一化后的数据分布出现扭曲，从而影响模型的性能。

3) 数据增强

在语音情感识别研究中，高质量的标注数据对于训练准确的模型至关重要。然而，实际获取大规模的标注语音情感数据往往面临诸多困难，例如数据采集成本高、标注过程繁琐且需要专业知识等，这导致可用的标注数据量有限。数据量不足会使得模型在训练过程中无法充分学习到语音情感的各种特征和模式，从而导致模型的泛化能力较差，在面对新的未见过的数据时，识别准确率会显著下降。为了缓解标注数据不足对模型性能的影响，数据增强技术应运而生。

常见的数据增强方法包括：

添加噪声：在原始语音数据中添加各种类型的噪声，如高斯白噪声、粉红噪声、环境噪声(如办公室嘈杂声、街道交通噪声等)，模拟实际环境中语音信号受到噪声干扰的情况。通过添加噪声进行数据增强，不仅可以增加数据的多样性，还可以使模型在训练过程中学习到如何从噪声背景中提取有效的语音情感特征，从而提高模型对噪声的鲁棒性和在实际复杂环境中的识别性能。

时间拉伸与压缩：对原始语音信号的时间轴进行拉伸或压缩操作。时间拉伸是指将语音信号在时间上进行扩展，使得语音听起来变慢；时间压缩则是将语音信号在时间上进行缩短，使得语音听起来变快。通过时间拉伸与压缩操作，可以生成具有不同语速的语音数据，增加数据的丰富性。由于语速的变化可能会影响语音情感的表达，因此模型在训练过程中可以学习到不同语速下语音情感的特征模式，从而提高模型对不同语速语音情感的识别能力。

2.5. 长短期记忆网络模型

长短期记忆网络(Long Short-Term Memory, LSTM)是一种特殊的循环神经网络(RNN)，由 Sepp Hochreiter 和 Jürgen Schmidhuber [7]在 1997 年提出。它有效地解决了传统 RNN 在处理长序列数据时遇到的梯度消失或梯度爆炸问题，从而在时间序列预测、自然语言处理、语音识别等众多领域得到了广泛应用。

LSTM 的结构整体如下：

细胞状态(Cell State)：LSTM 的核心在于细胞状态，它就像一条贯穿整个网络的传送带，信息可以在上面相对容易地流动。细胞状态允许模型在序列处理过程中保留长期信息，并且可以选择性地更新和传递这些信息。

门控机制(Gating Mechanisms)：LSTM 通过三种门控机制来控制细胞状态的信息流动，即输入门(Input Gate)、遗忘门(Forget Gate)和输出门(Output Gate)。这些门控机制由一个 Sigmoid 神经网络层和一个点乘运算组成，Sigmoid 层输出一个介于 0 和 1 之间的数值，用于表示对应门的开启程度，0 表示完全关闭，1 表示完全打开。

遗忘门(Forget Gate)：遗忘门决定了细胞状态中哪些信息需要被丢弃。它接收上一时刻的隐藏状态和当前时刻的输入，通过 Sigmoid 函数计算出一个遗忘系数，其中和分别是遗忘门的权重矩阵和偏置向量。然后，将遗忘系数与上一时刻的细胞状态进行点乘运算，得到经过遗忘门处理后的细胞状态。这样，遗忘门就可以根据当前输入和上一时刻隐藏状态，动态地决定细胞状态中哪些信息需要被保留，哪些信息需要被遗忘。

输入门(Input Gate)：输入门负责控制当前时刻的输入信息进入细胞状态。它同样接收上一时刻的隐藏状态和当前时刻的输入，通过 Sigmoid 函数计算出一个输入系数，其中和分别是输入门的权重矩阵和偏置向量。同时，通过一个 Tanh 函数生成一个候选细胞状态，其中和分别是生成候选细胞状态的权重矩

阵和偏置向量。然后，将输入系数与候选细胞状态进行点乘运算，得到经过输入门处理后的输入信息。最后，将经过遗忘门处理后的细胞状态与经过输入门处理后的输入信息相加，得到当前时刻更新后的细胞状态。这样，输入门就可以根据当前输入和上一时刻隐藏状态，动态地控制当前输入信息进入细胞状态，并与细胞状态中保留的长期信息进行融合，从而实现对细胞状态的更新。

输出门(Output Gate)：输出门用于控制细胞状态中哪些信息将被输出到当前时刻的隐藏状态。它接收上一时刻的隐藏状态和当前时刻的输入，通过 Sigmoid 函数计算出一个输出系数，其中和分别是输出门的权重矩阵和偏置向量。然后，将输出系数与当前时刻更新后的细胞状态进行点乘运算，得到经过输出门处理后的输出信息。最后，将经过输出门处理后的输出信息通过一个 Tanh 函数进行变换，得到当前时刻的隐藏状态。这样，输出门就可以根据当前输入和上一时刻隐藏状态，动态地控制细胞状态中哪些信息将被输出到当前时刻的隐藏状态，从而实现对隐藏状态的更新，并将隐藏状态传递到下一个时刻或作为模型的输出。

2.6. 双向长短期记忆网络模型

双向长短期记忆网络(Bidirectional Long Short-Term Memory, Bi-LSTM)是在 LSTM 基础上发展而来的一种循环神经网络架构，它通过同时处理正向和反向的序列信息，显著提升了模型对序列数据中复杂依赖关系的捕捉能力。Bi-LSTM 网络主要由输入层、两个方向相反的 LSTM 层、合并层以及输出层构成。在处理序列数据时，输入序列同时被送入正向 LSTM 层和反向 LSTM 层。正向 LSTM 层按照序列的时间顺序从前往后处理数据，而反向 LSTM 层则按照逆时间顺序从后往前处理数据。这两个 LSTM 层各自独立地学习序列在不同方向上的特征表示。随后，正向 LSTM 层和反向 LSTM 层在每个时间步的隐藏状态通过合并层进行合并。合并方式通常有拼接(concatenate)和相加(add)两种，其中拼接是将正向和反向的隐藏状态按维度拼接在一起，这种方式能够保留两个方向隐藏状态的全部信息，从而得到一个包含更丰富信息的特征向量；相加则是将正向和反向的隐藏状态对应元素相加，这种方式相对简单，能够突出两个方向隐藏状态中共同的信息部分。最后，合并后的特征向量被送入输出层，输出层根据具体的任务需求，如分类、回归等，对特征向量进行进一步的处理，最终得到模型的输出结果。

双向 LSTM 网络通过其独特的结构设计，能够同时从正向和反向两个方向对序列数据进行处理和特征提取，有效地捕捉序列数据中的长距离依赖关系和上下文信息。

2.7. Transformer 神经网络

传统神经网络处理输入信息时，往往是独立对待每一个输入，未能充分挖掘前后输入之间具有的关系。例如，在处理一段文字时，如果只关注每个单独的单词，而没有捕捉单词之间的语义联系，那么对整句话的理解可能会产生偏差。像 CNN 和 RNN 这样的神经网络模型，确实在信息关联过程中存在一定的局限性。

于是，为了更有效地捕捉序列中元素之间的关系，Ashish Vaswani 等人[8]提出了 Transformer 模型。这个模型与以往的网络模型有很大的区别，它是由多头注意力机制和前馈神经网络组成。这种设计使得模型能够在任意两个序列元素之间建立直接的联系，从而更好地捕获元素间的关联。

Transformer 模型通过多个自注意力层，对输入进行多通道的线性变换，并分别计算出注意力得分。所有这些结果再经过一次全连接操作，最后通过线性变换得出最终输出。与 RNN 和 CNN 在水平方向传播不同的是，Transformer 在每一层上都可进行并行计算，因此，Transformer 在处理效率、性能和精度上，都较传统的 RNN 和 CNN 表现出更优的表现。

在近几年的语音情感识别研究中，Transformer 也逐渐被运用其中。在 2023 年，Wagner J [9]等对基

于 Transformer 的架构在语音情感识别中的应用进行了深入分析。作者在多个预训练的 wav2vec 2.0 和 Hubert 变体上进行微调，针对 MSP-Podcast 数据集的唤醒度、支配度和效价维度进行实验，并使用 IEMOCAP 和 MOSI 数据集测试跨语料库泛化能力。研究发现基于 Transformer 的架构在效价预测方面表现出色，且比基于 CNN 的基线更稳健。2024 年，Ting Dang [10]等人提出了一种基于 CNN-Transformer 和多维度注意力机制的语音情感识别网络，旨在对语音中不同粒度的局部和全局信息进行建模，并捕获语音信号中的时间、空间和通道依赖性。实验在 IEMOCAP 和 Emo-DB 数据集上进行，结果表明该方法显著提高了性能。

2.8. 本文模型

综合以上研究现状，本文选择使用学习语音信号长期相关性和关注情感分类的混合模型，利用 Transformer 网络中的注意力机制关注情感部分，使用 MFCC 作为特征，并将其输入到所提出的混合 LSTM-Transformer 分类器中识别结果。将 LSTM 与 Transformer 结合，可以充分利用 LSTM 在处理长期依赖方面的优势，以及 Transformer 在捕捉全局依赖和并行处理方面的能力。本研究的网络模型结构如图 2 所示。

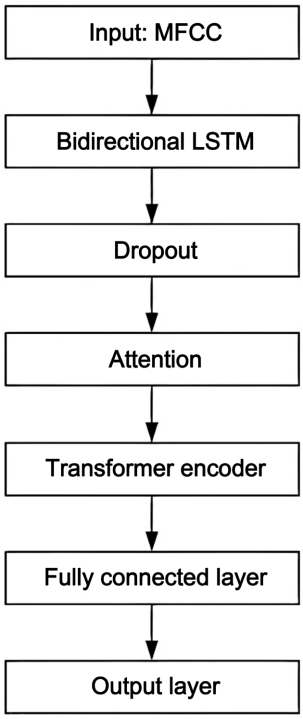

Figure 2. Network model structure
图 2. 网络模型结构

输入序列首先通过嵌入层转化为向量表示，LSTM 层处理这些向量，捕捉序列中的局部时间依赖性，LSTM 层的输出再作为 Transformer 层的输入，Transformer 层利用自注意力机制进一步捕捉全局依赖关系，最后，将 LSTM 和 Transformer 的输出进行融合(如拼接、加权求和等)，并输入到全连接层进行分类、生成或序列标注等任务。

在这个结构图中：

输入序列(Input)：原始的序列数据，如文本、时间序列等。

嵌入层：将输入序列转化为向量表示，这些向量作为后续层的输入。

LSTM 层：处理这些向量，捕捉序列中的长期依赖关系。

Transformer 层：接收 LSTM 层的输出，利用自注意力机制捕捉全局依赖关系。

融合层(Fully connected)：将 LSTM 和 Transformer 的输出进行融合，得到最终的特征表示。设 Transformer 的输出为 $\boldsymbol{T} \in RdT$，LSTM 的输出为 $\boldsymbol{L} \in RdL$，并且假设 $dT = dL = d$ (若维度不同，可先通过全连接层将其映射到相同维度)。引入两个可学习的权重参数 α 和 β，那么加权求和后的融合输出 $\boldsymbol{F}$ 可以表示为：

$$\boldsymbol{F} = \alpha \boldsymbol{T} + \beta \boldsymbol{L} \tag{3}$$

通常为了简化计算，会对权重进行归一化处理，即 $\alpha + \beta = 1$，此时可以只使用一个可学习的权重参数 ω，令 $\alpha = \omega$，$\beta = 1 - \omega$，则公式变为：

$$\boldsymbol{F} = \omega \boldsymbol{T} + (1 - \omega)\boldsymbol{L} \tag{4}$$

最后，将融合后的输出 $\boldsymbol{F}$ 通过一个全连接层进行线性变换，得到最终的输出 $\boldsymbol{O}$：

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{F} + \boldsymbol{b} \tag{5}$$

其中，$\boldsymbol{W} \in Rdout \times d$ 是全连接层的权重矩阵，$\boldsymbol{b} \in Rdout$ 是偏置向量，$dout$ 是最终输出的维度。

输出层(Output)：根据融合后的特征表示进行分类、生成或序列标注等任务。

3. 实验过程

本文实验的操作系统为 Windows 10，使用的编程语言为 Python 3.7，开发工具为 PyCharm 2022，开发框架为 TensorFlow1.13.1，CPU 版本为 Inter Core™ i5-7300HQ，GPU 版本为 NVIDIA GeForce GTX1050Ti 8GB。

3.1. 数据集

本文采用的是 EMO-DB，即柏林情感语音数据库(Berlin Database of Emotional Speech)。由柏林工业大学录制的德语情感语音库，由 10 位演员(5 男 5 女)对 10 个语句(5 长 5 短)进行 7 种情感(中性/neutral、生气/anger、害怕/fear、高兴 happy、悲伤/sadness、厌恶/disgust、无聊/boredom)的模拟得到，共包含 800 句语料，采样率 48 kHz (后压缩到 16 kHz)，16 bit 量化。语料文本的选取遵从语义中性、无情感倾向的原则，且为日常口语化风格，无过多的书面语修饰。语音的录制在专业录音室中完成，要求演员在演绎某个特定情感前通过回忆自身真实经历或体验进行情绪的酝酿，来增强情绪的真实感。经过 20 个参与者(10 男 10 女)的听辨实验，得到 84.3%的听辨识别率。

3.2. 数据处理

音频加载与 MFCC 特征提取：

使用 librosa.load 函数加载音频文件，该函数将音频文件读取为 NumPy 数组，并返回音频时间序列 y 和采样率 sr。

使用 librosa.feature.mfcc 函数计算音频的 MFCC 特征。提取的 MFCC 特征矩阵被转置，以匹配(时间步，特征数)的形状，这通常是深度学习模型输入所期望的格式。

标签提取与编码：

遍历指定路径下的所有.wav 音频文件，根据文件名提取情感标签。这里假设文件名中包含了情感标签的信息，并且这个信息被用来从 emotion_labels 字典中检索具体的标签名称。

使用 LabelEncoder 对标签进行编码，将字符串标签转换为整数标签。这是因为大多数机器学习算法，特别是深度学习模型，通常期望输入是数值型的。

数据集封装：

将音频文件路径、标签和任何可选的数据转换(如标准化)封装在 MfccEmotionDataset 类中，这个类继承自 PyTorch 的 Dataset 类。len 方法返回数据集中的样本总数。

getitem 方法根据索引加载单个样本，包括音频文件的 MFCC 特征和对应的标签。这个方法还负责应用任何可选的数据转换，并处理音频长度不一致的问题。

处理音频长度不一致：

由于不同音频文件的长度可能不同，而深度学习模型通常要求所有输入具有相同的形状，因此需要对 MFCC 特征进行截断或填充，以匹配目标长度(在这个例子中是 128)。

如果 MFCC 特征的时间步数超过目标长度，则进行截断；如果少于目标长度，则在末尾添加零填充。

可选的数据转换：

通过 transform 参数，可以在提取 MFCC 特征后应用额外的数据预处理步骤，如标准化或归一化。这些步骤对于提高模型性能和稳定性通常很重要。

单独的样本处理：

get_one 方法提供了一种处理单个音频文件的方式，而无需创建整个数据集对象。这对于测试或单独处理文件很有用。

3.3. 实验数据和分析

本实验整体数据从损失率，准确率和混淆矩阵三方面分析实验情况。在整个训练过程中，损失值都呈现出平稳的下降的趋势，说明模型正在逐渐优化其参数，整个训练过程都属于正常，没有出现明显的过拟合现象。实验过程中的损失值变化如图 3 所示。

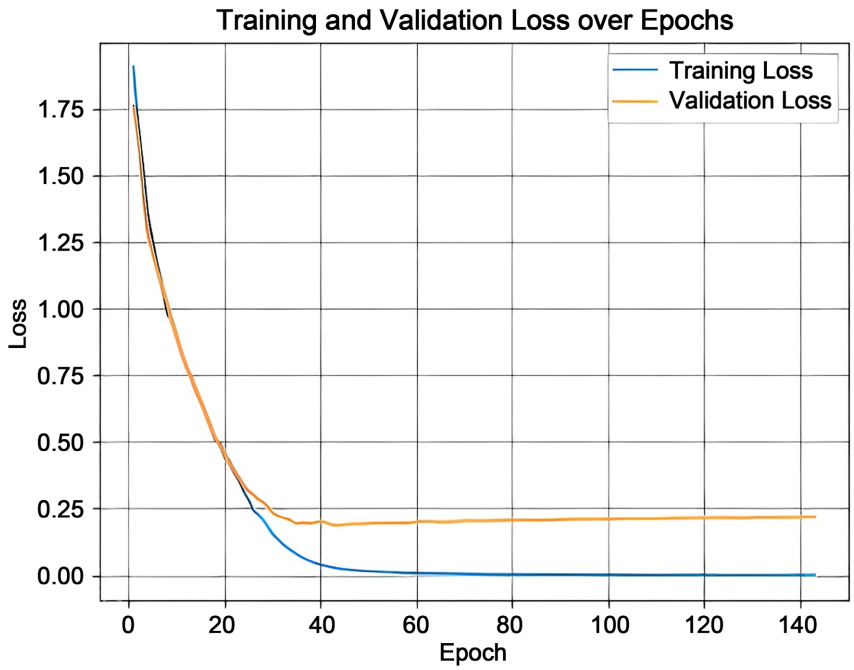

Figure 3. Training and validation loss over epochs
图 3. 各训练轮次的训练损失和验证损失

对七种语音情感训练得到的预测准确率(Accuray)，结果如图 4 所示。

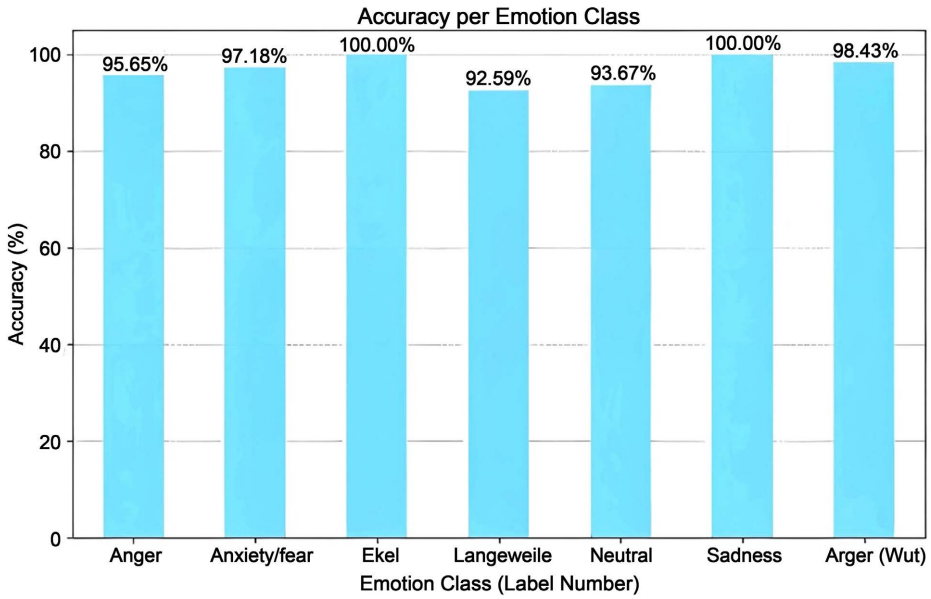

Figure 4. Accuracy per emotion class
图 4. 每个情感类别的准确

每个情感标签独立的识别率都在 92%以上，部分情感标签的识别准确率达到了 100%，说明识别率处于一个较高的水平。综合得到的平均识别率约为 93%。

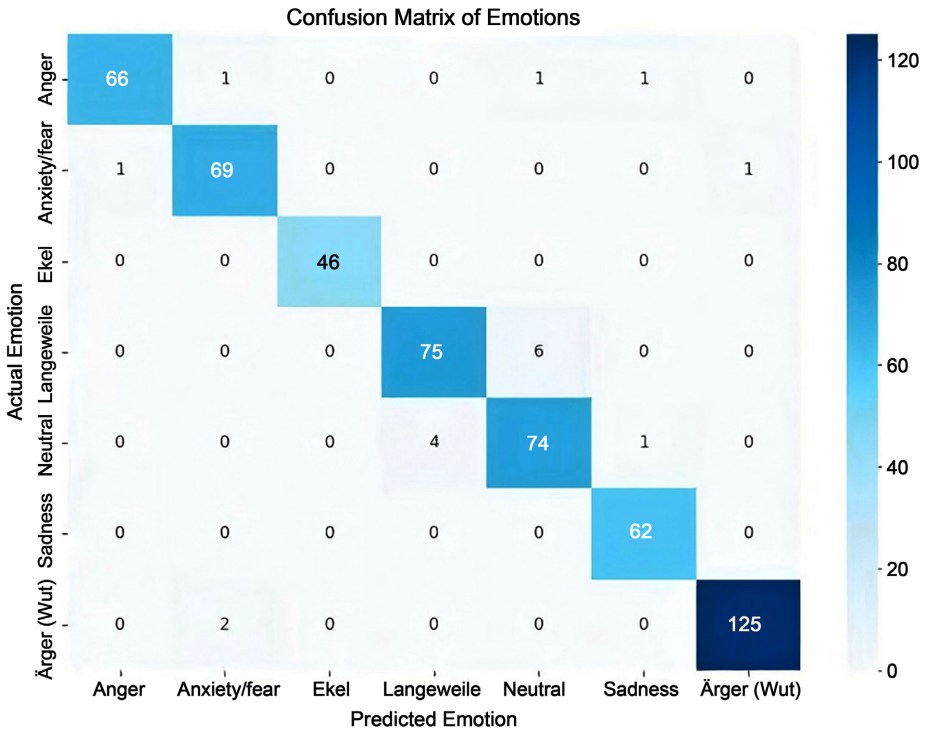

Figure 5. Confusion matrix of emotions
图 5. 情感混淆矩阵

最后再根据实验数据生成了混淆矩阵，整体结果如图 5 所示。

根据混洗矩阵可以看出，对角线元素远大于其它元素，说明该列所代表的类别被正确预测的概率较高，模型对这个类别的预测比较准确。

综合以上结果可以看出，整体平均识别率约为 93%，呈现一个较好的识别率，但是对于部分偏中性或者偏复杂的语音情感识别准确率还有待提升，混淆矩阵中 Neutral 和 Langeweile 两种情感的识别错误率比别的情绪错误率明显高一些。

4. 结论

本文基于 EMO-DB 数据集，采用了 LSTM-Transformer 的混合网络模型，对语音情感识别这一任务进行了研究和识别，最后得到的综合识别率为 93%。此外，对于一些情感表达较为隐晦或复杂的语音样本，模型的识别效果仍不理想。例如，在识别包含复杂情感或者偏中立的语音时，模型的准确识别率相对较差。在下一步研究中，需要进一步挖掘语音信号中的深层情感特征，提高模型对复杂情感的理解和识别能力。

参考文献

[1] 李美娟. 基于深度学习和特征融合的语音情感识别方法研究[D]: [硕士学位论文]. 济南: 齐鲁工业大学, 2024.

[2] Jin, Q., Li, C., Chen, S., *et al*. (2015) Speech Emotion Recognition with Acoustic and Lexical Features. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). South Brisbane, 19-24 April 2015, 4749-4753. https://doi.org/10.1109/ICASSP.2015.7178872

[3] Zhang, S., Zhang, S., Huang, T. and Gao, W. (2018) Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia*, **20**, 1576-1590. https://doi.org/10.1109/tmm.2017.2766843

[4] 陶建华, 陈俊杰, 李永伟. 语音情感识别综述[J]. 信号处理, 2023, 39(4): 571-587.

[5] 程适, 骆晓宁, 李冬城, 等. 一种基于双向 LSTM 的语音情感识别模型[J]. 长江信息通信, 2022, 35(7): 19-22.

[6] Mohan, M., Dhanalakshmi, P. and Kumar, R.S. (2023) Speech Emotion Classification Using Ensemble Models with MFCC. *Procedia Computer Science*, **218**, 1857-1868. https://doi.org/10.1016/j.procs.2023.01.163

[7] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[9] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., *et al*. (2023) Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 10745-10759. https://doi.org/10.1109/tpami.2023.3263585

[10] Tang, X., Lin, Y., Dang, T., Zhang, Y. and Cheng, J. (2024) Speech Emotion Recognition via CNN-Transformer and Multidimensional Attention Mechanism. arXiv: 2403.04743. https://doi.org/10.48550/arXiv.2403.04743