文章编号:1671-4598(2026)04-0137-07 DOI:10.16526/j.cnki.11-4762/tp.2026.04.018 中图分类号:TN912.34 文献标识码:A

基于改进 ECAPA-TDNN 的语种识别方法研究

李 伟[1],姚 镭[2],刘 成[1]
(1.上海大学 微电子学院,上海 201800;2.临港实验室,上海 201100)

摘要:传统的语音识别模型难以很好地克服不同语种之间的差异性,并且特征提取不全面造成识别准确率不理想,为了解决这个问题,提出了一种基于改进 ECAPA-TDNN 的语种识别和分类方法;在 ECAPA-TDNN 的基础上,设计了以多头自注意力机制为核心的编码器模块,通过并行的注意力头对不同时间步的特征序列进行加权聚合,捕获不同时间步之间的全局依赖关系;设计了以多层残差块为核心的前端卷积模块,卷积层采取多层卷积和残差连接的方式,强化了多层次特征的提取能力,并针对语音信号特性,采取在时间和频率维度分开降采样的方式,生成更丰富的特征;在对输入音频进行预处理得到 FBank 特征后,进一步使用数据增强手段,再输入至优化后的语种识别模型;经过实验证明,改进后的模型在 Common Voice 和 CSS10 数据集上的识别准确率分别达到了 92.91%和 90.39%,相比 ECAPA-TDNN 提升了 5.23%和 4.78%,同时针对前端卷积模块和注意力编码器模块设计了消融实验,证明了各模块的有效性;上述结果表明,提出的方法在 ECAPA-TDNN 的基础上增强了对于语音的局部和全局特征的提取,提高了在语种识别任务上的性能。

关键词:语种识别;ECAPA-TDNN;多头自注意力;特征提取;卷积

Research on Spoken Language Identification Method Based on Improved ECAPA-TDNN

LI Wei[1], YAO Lei[2], LIU Cheng[1]
(1. School of Microelectronics, Shanghai University, Shanghai 201800, China;
2. Lingang Laboratory, Shanghai 201100, China)

Abstract: It is difficult for traditional speech recognition models to effectively overcome the difference of different languages, and incomplete feature extraction leads to a suboptimal recognition accuracy. To address this issue, a language identification and classification method based on improved emphasized channel attention, propagation and aggregation time-delay neural network (ECAPA-TDNN) is proposed. Based on the ECAPA-TDNN, an encoder module for multi-head self-attention mechanism as a center is designed, the weighted aggregation of feature sequences across different time steps is performed through parallel attention heads, and global dependencies are captured between various time steps. Meanwhile, a front-end convolutional module based on a multi-layer residual block as a center is designed. The convolutional layer adopts a multi-layer convolution and residual connection approach, enhancing the capability of multi-level feature extraction. For the characteristics of speech signals, it performs separate downsampling in both time and frequency dimensions to generate more diverse features. After the input audio is preprocessed to obtain the FBank features, the data augmentation technique is further applied and input into the optimized language recognition model. Experimental results demonstrate that the enhanced model achieves the recognition accuracies by 92.91% and 90.39% on Common Voice and CSS10 datasets respectively, with the improvements of 5.23% and 4.78% over the original ECAPA-TDNN. Also, ablation experiments are conducted on both the front-end convolutional module and attention encoder module, verifying the effectiveness of each component. These results indicate that based on the ECAPA-TDNN, the proposed approach enhances the extraction for both local and global speech features, significantly improving a performance in language identification tasks.

Keywords: language identification; ECAPA-TDNN; multi-head self-attention; feature extraction; convolution

收稿日期:2025-03-31; **修回日期**:2025-05-13。
基金项目:临港实验室攻关任务(LG-GG-202402-05-02)。
作者简介:李 伟(2000-),男,硕士研究生。
通讯作者:刘 成(1980-),男,博士,副教授。
引用格式:李 伟,姚 镭,刘 成.基于改进 ECAPA-TDNN 的语种识别方法研究[J].计算机测量与控制,2026,34(4):137-143,163.

0 引言

语音处理作为人工智能领域的重要研究方向之一，涵盖了语音识别、语音合成及语音增强等多种技术[1]，近年来在深度学习技术的推动下取得了显著进展。其中，语种识别（LID，language identification）是语音处理中的一个关键任务，通过分析语音信号特征实现语种的智能识别，在构建多语言智能系统中具有重要地位[2]。作为多语种语音识别系统的前置模块，其识别准确度直接影响后续处理流程的可靠性；在跨语言机器翻译系统中，语种识别技术可自主选择适配的语音识别与翻译引擎，实现处理流程的全自动化。这种技术特性使其成为构建智能语言服务体系不可或缺的核心组件[3]。在具体的应用场景方面，该技术已展现出显著的实用价值，例如，在智能客服系统中[4]，让机器能够自动判断客户所用的语言以便转接对应语言的人工客服，或者调用相关的语音识别引擎进行处理；在语音翻译系统中，语种识别可以辅助选择合适的语音识别引擎和机器翻译引擎，避免人工选择的麻烦。另外，语种识别在虚拟会议、音频资料检索、智能对话系统也有所应用[5-6]。

语音处理技术的发展经历了从基于规则的传统方法到依靠深度学习的现代技术的演变。早期，研究者主要依赖梅尔频率倒谱系数（MFCC，Mel frequency cepstral coefficient）[7]、线性预测倒谱系数（LPCC，linear predictive cepstral coefficient）等手工特征提取方法，再结合诸如高斯混合模型（GMM，Gaussian mixed model）和支持向量机（SVM，support vector machine）之类的传统分类器[8]。然而，随着语言数据复杂性和多样性的增加，这些传统方法在捕捉复杂的语音特征方面逐渐显得力不从心。为了应对这一挑战，深度学习模型，包括卷积神经网络（CNN，convolutional neural network）、循环神经网络（RNN，recurrent neural network）及长短期记忆网络（LSTM，long short-term memory）等方法被逐渐开发和使用，显著提升了语音识别和语种分类的性能[9]。

在诸多深度学习方法中，深度神经网络（DNN，deep neural network）作为一种区分性建模方法，可以直接对输入语音特征进行语种分类[10]，并且是对语音进行帧级别的分类，在语音识别准确性方面取得了实质性的进步。有研究人员基于声学特征和音位特征对语种识别的有效性[11]，融合了声学特征和音位信息，从语音信号中提取声学特征，送入到一个共享的 CNN 模块，同时针对 LID 任务和音素分割任务进行优化，但这个过程复杂度很高。文献［12］提出了一种音素时序神经模型（PTN，phonetic temporal neural net），使用语音识别中的音素区分网络生成音素的后验概率，在对这一后验概率进行 RNN 建模。PTN 可以将语音片段转换为音素后验概率的向量序列，不仅时序解析度更高，且包含的音素信息更加丰富。另一方面，PTN 中的语言模型是连续序列上的神经语言模型，可以实现更复杂的时序建模。

对于语种识别任务来说，不同语种之间的差异性提高了音频数据的复杂度，增加了识别的困难性。近年来，Transformer 模型的引入为语音处理带来了新的突破[13]。得益于其自注意力机制，Transformer 能够有效捕捉语音信号中的长程依赖和全局特征，在语音处理任务中展现了出色的表现[14]。注意力机制最早由文献［15］提出，已经在自然语言处理、语音和计算机视觉的各种任务中彻底改变了序列建模和转换模型[16]。总的来说，它允许模型专注于输入或输出序列的特定部分，而不受元素之间距离的限制。

然而，尽管 Transformer 在处理长序列数据方面具有优势，其在捕捉局部特征时的表现仍不及其他特化的深度学习模型，它在处理短时序列和局部特征时的效率较低，在声学语种分类中，音频信号通常具有强烈的时序性和局部特征[17]，尤其是在低频和高频信息的捕捉上，传统的 Transformer 架构可能无法充分有效地捕获这些局部信息。

另一方面，如 ECAPA-TDNN（Emphasized Channel Attention，Propagation and Aggregation Time-delay Neural Network）[18]等在说话人识别任务中表现良好的模型在长程语音韵律特征方面的捕捉能力不足。ECAPA-TDNN 的局限性在多语种复杂场景中尤为显著。例如，在处理声调语言（如汉语）与非声调语言（如德语）混合的语音流时，其局部卷积结构难以捕捉跨语种的全局韵律差异（如声调变化、重音分布），导致混淆相似音素的语言类别。此外，在低信噪比环境下（如会议录音或客服对话），ECAPA-TDNN 对长时语音特征的建模能力不足，可能错误聚合噪声片段与目标语音的时频特征，造成语种误判。对于具有部分相似音素但语法差异显著的语言，ECAPA-TDNN 因缺乏全局上下文建模，易将局部音段特征错误关联至目标语种。为了解决这一问题，本文在 ECAPA-TDNN 的基础上，加入了为语种识别所设计的注意力编码器模块，并优化了前端卷积模块（FCM），将 Transformer 的全局特征捕捉能力与 ECAPA-TDNN 的局部特征提取优势相结合，以优化语种分类任务中的模型性能。通过实验验证，该模型在多个公开数据集上的表现优于传统方法，证明了其在声学语种分类领域的有效性与潜力。

1 ECAPA-TDNN 和注意力机制

1.1 ECAPA-TDNN 模型

ECAPA-TDNN 模型是近年来在语音识别领域取得

显著进展的深度神经网络模型。它在经典的 TDNN（Time-Delay Neural Network）模型基础上进行了优化，主要包括强调通道注意力机制和改进的时间延迟神经网络结构。ECAPA-TDNN 通过引入通道注意力机制（Channel Attention），堆叠多个 TDNN 层和 Res2Net 模块，有效地聚焦于输入特征中最具代表性的部分，从而提升语音识别任务中的特征表示能力，其结构如图 1 所示。

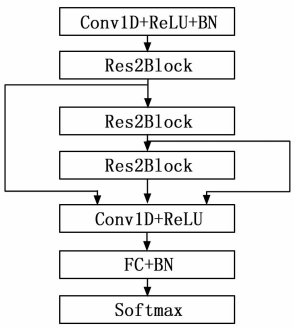

图 1　ECAPA-TDNN 结构

然而，尽管 ECAPA-TDNN 在多种语音处理任务中表现出色，但在用于声学语种分类任务时，仍然存在一些不足和问题。首先，ECAPA-TDNN 主要依赖于卷积层和时间延迟机制，这对于捕捉长序列中的全局信息和长期依赖关系存在一定的局限性。语种分类通常需要分析语音信号中的长时依赖关系，这要求模型能够有效地处理跨越较长时间步的特征信息，而 ECAPA-TDNN 在这方面的能力相对较弱。其次，ECAPA-TDNN 的主要设计仍然局限于局部特征的提取，无法充分挖掘全局信息。对于声学语种分类任务，音频信号中的全局语义信息对于语种区分起着重要的作用[19]。

此外，传统的 ECAPA-TDNN 模型并未有效结合自注意力机制，这使得它在处理复杂语种的音频特征时，可能无法充分关注到输入信号中的重要部分，尤其是在面对多语种的复杂情况时，模型的表现可能会受到限制[20]。因此，ECAPA-TDNN 在语种分类任务中，仍然存在着长序列建模能力弱以及全局信息提取不足等问题。为了克服这些不足，本文将自注意力机制与 ECAPA-TDNN 结合，通过引入自注意力机制和全局特征建模，进一步提升模型在声学语种分类任务中的性能。

1.2　多头自注意力机制

多头自注意力机制（MHSA，multi-head self-attention）引入自 Transformer 架构，多头注意力机制是在自注意力机制的基础上发展起来的，是自注意力机制的变体，主要通过使用多个独立的注意力头，分别计算注意力权重，并将它们的结果进行拼接或加权求和，从而获得更丰富的表示，其结构如图 2 所示。

注意力机制实际上是查询向量 $\boldsymbol{Q}$（Query）和一组

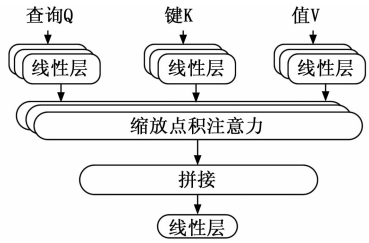

图 2　多头自注意力机制示意图

键 $\boldsymbol{K}$（Key）—值 $\boldsymbol{V}$（Value）向量对到输出的映射。具体来说，输出向量是值向量的加权求和，其中值向量的权重通过计算查询向量和键向量之间的兼容性获得。假设每个查询向量 $\boldsymbol{Q}$ 和键向量 $\boldsymbol{K}$ 的维度都是 d_k，值向量 $\boldsymbol{V}$ 的维度是 d_v。查询 $\boldsymbol{Q}$ 和每个键 $\boldsymbol{K}$ 之间的兼容性函数计算为它们的点积，并除以 $\sqrt{d_k}$ 进行缩放。为了获得值的权重，缩放后的点积值通过 softmax 函数传递：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

多头自注意力机制对于 h 个注意力头和模型中的维度 d_m 维的查询、键和值通过线性投影分别投影 h 次到 d_k、d_k 和 d_v 维度。每个头按照公式（1）执行注意力操作。h 个 d_v 维的输出被拼接起来，并通过另一个投影矩阵投影回 d_m 维度：

$$\text{MultiHeadAttn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}^O \tag{2}$$

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{3}$$

其中：$\boldsymbol{W}^Q$，$\boldsymbol{W}^K$，$\boldsymbol{W}^V$，$\boldsymbol{W}^O$ 是投影矩阵。多头注意力机制允许模型在不同的表示空间中联合关注序列的不同部分。

2　本文方法和模型构建

本文提出了一种基于改进 ECAPA-TDNN 的语种分类方法，其总体框架如图 3 所示，旨在结合 Transformer 架构的全局建模能力与 ECAPA-TDNN 在语音特征提取方面的优势，以提升语种分类的性能和效率。该方法的核心在于通过优化模型结构和参数，改进音频处理流程，从而增强模型的整体表现。

首先对语音数据进行预处理，提取 FBank 特征。音频数据样本会先转化为 float32 类型，采样率设置为 16 kHz，并进行音量归一化，以确保输入数据的一致性。提取 Fbank 特征的过程包括将音频波形转化为数字信号，通过离散傅里叶变换（FFT）获得频谱信息，并使用梅尔滤波器对不同频率段的能量进行采集，以适应人耳的听觉感知。

对音频的预处理步骤如下：1）对输入的音频信号预加重；2）分帧，即将音频分成小段，每帧持续约 20～40 ms，同时每帧之间会有一定重叠；3）对每一帧信

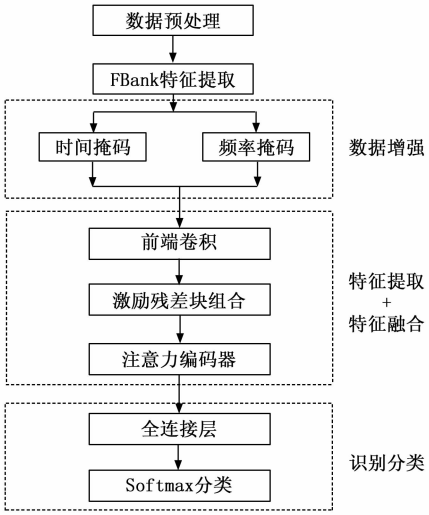

图 3 本文方法的流程图

号加窗函数。

然后进行快速傅里叶变换，得到频谱信息，使用 80 个梅尔滤波器采集不同频率段的能量，将频谱能量映射到 Mel 频率尺度上，每个滤波器集中在一个特定的频率范围，模拟人耳对不同频率的感知，得到 FBank 特征，如图 4 所示。

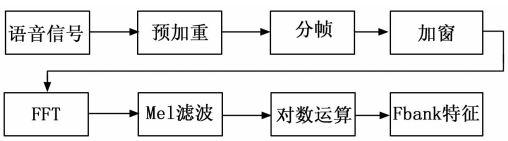

图 4 FBank 特征提取方法

在完成数据预处理后，接下来进行数据增强，以提高模型的鲁棒性。具体方法是在频谱图上随机掩盖时间和频率区域。通过这种频率掩码和时间掩码操作，可以有效扩展训练样本，增强模型对不同输入的适应能力。

最后，经过处理后的数据将输入到改进后的本文模型中进行声学语种分类。该模型的结构包括特征提取、特征融合和分类 3 个主要部分。特征提取部分由卷积层和基本残差块组成，改进的前端卷积模块使得模型能够提取更复杂的特征，尤其是在高阶特征的提取上表现优异。

这个模型的整体结构可以分为几个主要部分：特征提取部分，特征融合部分以及分类部分。模型结构如图所示，经过前期预处理的数据输入进来，首先经过特征提取部分。前端卷积模块包含一个卷积层和两个残差块。3 个连续的激励残差块，每个残差块通过 Res2Net 结构和 SE 机制进行特征处理和通道关系自适应调整。每个 Block 的输出加上之前的输入形成残差连接。最后，将 3 个 Block 的输出在通道维度上拼接形成一个大的特征图。将拼接后的特征图输入到注意力编码器中，进行全局特征建模。通过一个 1D 卷积层、ReLU 激活和批归一化，对特征进行进一步处理。分类部分则使用注意力聚合与全连接层，最终通过 softmax 函数实现语种分类，具体结构如图 5 所示。

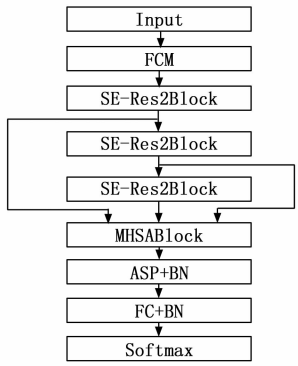

图 5 本文模型的结构示意图

2.1 前端卷积模块

为了更好地捕获某些局部频率区域中出现的特征，本文设计了一个基于多层残差块的前端卷积模块 FCM，通过多级卷积逐步实现对语音信号的特征提取，如图 6 所示。前端卷积模块（FCM）由多层残差块和卷积操作构成，其设计旨在高效提取语音信号的时频局部特征。模块输入为 80 维 FBank 特征（形状为 [Batch，1，80，T]，其中 80 为频率维度，T 为时间步数）。初始卷积层采用 3×3 的卷积核，步幅为（1，1），将输入通道从 1 扩展至 32，输出特征图尺寸为 [Batch，32，80，T]。随后，特征依次通过两层残差块，每层包含两个 ResBlock 单元。每个 ResBlock 的卷积核尺寸为 3×3，步幅在频率维度设置为 2、时间维度保持 1，通过两次下采样将频率维度从 80 逐步压缩至 20（80→40→20），同时保留时间维度的原始分辨率。残差块内部通过跳跃连接融合输入与输出特征，当输入输出通道数不匹配时，采用 1×1 卷积调整维度以适配残差相加操作。最后，最终卷积层采用 3×3 卷积核，步幅在时间维度设置为 2、频率维度保持 1，将时间步数压缩至原始输入的 1/8（T→T/8），输出特征形状为 [Batch，32，10，T/8]。通过合并频率与通道维度，模块最终输出高维特征 [Batch，320，T/8]，为后续模块提供兼具局部细

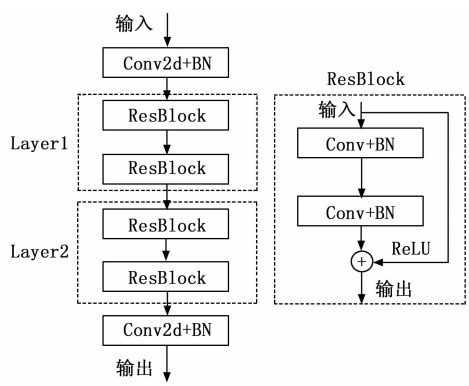

图 6 前端卷积模块的结构示意图

节与全局结构的语种识别特征。

模型的输入数据是提取得到的 FBank 特征，具有明显的二维结构，即时间维度和频率维度，前端卷积模块使用卷积操作提取时频信息。在时间维度上，捕捉语音随时间变化的动态特征（如语速、节奏等），在频率维度上，提取语音中不同频段的能量分布特征（如语种特征依赖的共振峰等）。通过这些卷积操作，实现对语音信号短时间动态和频谱特性的局部建模。

由于后续的注意力编码器模块对长序列建模的计算成本较高，因此在 FCM 中通过多层卷积和降采样操作，逐步减少时间步的数量，降低后续模型的计算复杂度。聚合时间步的信息，增强模型对整体时序特征的全局感知能力。同时，增加了通道数来构建高维特征空间，通过卷积操作增加通道，将原始输入映射到更高维度的特征空间，目的是在高维特征空间中表示更丰富的语种相关信息，提供更多的特征维度供后续模块进行建模和特征提取。FCM 使用残差模块对低级别特征进行多层次提取，残差模块可以学习更深层次的特征表示，同时避免梯度消失或爆炸的问题，不同卷积层聚合了从低层次到高层次的语音特征，形成具有层次化的特征表示。

此外，FCM 模块在一定程度上可以提高模型泛化能力。FCM 使用二维卷积对输入的时频特征进行建模，提取了语音信号在时间和频率维度的局部模式（如共振峰、时序动态特征）。局部特征通常是通用的，与具体语言、说话人和环境无关，因此模型可以更好地泛化到未见过的语言或语音样本。FCM 使用多层残差块，通过逐层的卷积操作和残差连接提取不同层次的特征，低层特征主要是语音信号的基础特性，如频谱结构。高层特征则包含更复杂的语种相关模式。层次化的特征表示可以更全面地描述输入数据，从而提升模型对多样化数据的适应能力。

2.2　注意力编码器模块

本文设计了一个注意力编码器模块，针对语种识别的需求，模块结构和参数设置都经过了优化，以适配音频特征的全局建模要求并充分利用全局与局部的时序信息。该模块基于堆叠的多层注意力编码器结构，具体采用两层编码器，每层由多头自注意力机制和前馈神经网络组成。多头机制能够并行提取信息，增强对复杂音频特征的建模能力。同时，模块特征维度与输入数据特性一致，确保高效特征表示。其中超参数设置如表 1 所示。

其中，d_{model}为输入特征维度，等于前置模块输出通道数的 3 倍，即 channels×3，n_{head}代表注意力头的数量，设置为 8，能够高效处理特征的多维模式，$dim_{\text{feedforward}}$表示前馈网络的隐藏层维度，设置为 2 048，确保足够的非线性变换能力，d_{ropout}设置为 0.1，用于防止过拟合；num_{layers}表示编码器层的堆叠数量，设置为 2。

表 1　注意力编码器参数设置

参数	设置值
d_{model}	channels×3
n_{head}	8
$dim_{\text{feedforward}}$	2 048
d_{ropout}	0.1
num_{layers}	2

注意力编码器模块基于 Transformer 架构，接收前端卷积模块输出的拼接特征（形状为［Batch，channels×3，T/8］）。输入阶段，特征首先转置为［Batch，T/8，channels×3］，以适配 Transformer 的序列处理格式。模块核心为两层堆叠的 Transformer 编码器层，每层包含多头自注意力机制与前馈神经网络。多头自注意力机制将特征分割为 8 个独立子空间，并行计算时间步间的全局依赖关系。具体地，每个子空间通过线性投影生成查询（Query）、键（Key）、值（Value）矩阵，并基于缩放点积注意力计算权重，最终拼接 8 个子空间输出。前馈神经网络由两个全连接层构成，隐藏层维度为 2 048，激活函数为 ReLU，该设计目的在于增强特征的局部非线性表达能力。每层编码器的输出通过残差连接与层归一化（LayerNorm）融合，以稳定训练过程并加速收敛。最终，特征转置回［Batch，channels×3，T/8］，保留与后续池化层的兼容性。

该模块的核心是多头自注意力机制，能够通过计算序列中每个时间步与其他时间步的关系捕捉长距离依赖。其基本原理是通过将输入特征序列映射到查询、键和值 3 个子空间，计算每个时间步与其他时间步的相关性分数（Attention Score），并基于这些分数加权聚合特征。模型将特征维度分成 8 个子空间，每个子空间的维度为 192，独立计算注意力权重，从而捕获不同语种特征在多样子空间下的依赖关系，多头机制能够并行提取信息，增强对复杂音频特征的建模能力。

输入音频经过前端卷积处理后生成（batch size，512×3，时间步）的特征图，特征通过通道展开和转置后，输入自注意力层的形状为（batch size，时间步，1 536），其中时间步根据卷积模块设定动态调整。为了对局部特征进行非线性映射和增强，在每个时间步独立应用前馈神经网络（FNN），由两个全连接层和 ReLU 激活函数组成，FNN 不仅可以提升特征的表达能力，还能增强模型对复杂模式的拟合能力，处理音频中更细致的语种差异。同时，通过在自注意力层和 FFN 层添加捷径连接（Shortcut Connection），直接将输入特征与输出相加，缓解梯度消失问题并稳定训练过程，并在每个子层后进行标准化，确保特征分布一致性，提升收敛

速度和数值稳定性。

该注意力模块接收由前端卷积模块输出的特征图，初始形状为（batch size，通道数，时间步）。在输入该模块之前，特征被转置为（batch size，时间步，特征维度）。在每一层注意力编码器中，多头自注意力机制计算全局依赖关系，聚合时间序列特征；前馈神经网络对局部特征进行非线性映射和增强；通过残差连接和层归一化，稳定特征表达并提升训练效率。经过多层堆叠，生成的输出特征包含丰富的全局上下文信息和时间依赖特性。最后，特征被转置回原始形状（batch size，通道数，时间步），便于后续池化操作和分类模块的进一步处理。

3 实验及结果讨论

3.1 数据集和性能评价指标

本文实验在多语种数据集 Common Voice 与 CSS10 上进行[21]。两个数据集均选取了包含汉语、日语、法语、德语、俄语共 5 种语言在内的音频数据，共 17 500 条音频片段，以九比一的比例划分训练集与测试集。其中，Common Voice 数据集有 9 000 条训练数据和 1 000 条测试数据。CSS10 数据集有 6 750 条训练数据和 750 条测试数据。相关数据集以及训练集、测试集的构成如表 2 所示。

表 2 数据集构成 条

数据集	汉语	日语	法语	德语	俄语	训练集	测试集	总计
Common-Voice	2 000	2 000	2 000	2 000	2 000	9 000	1 000	10 000
CSS10	1 500	1 500	1 500	1 500	1 500	6 750	750	7 500

本文采用多语种识别正确率作为评价指标，其表达式为：

$$\overline{Acc} = \frac{\sum_{i=1}^{G} R}{N}, \quad i = 1,2,\cdots,G \tag{4}$$

其中：R 为每种语种的识别正确数，N 为测试集总数，G 为语种数量。

3.2 实验设置和数据增强

读取音频数据，分割音频路径和标签，分成训练数据、测试数据和分类标签 3 个路径。对语音数据进行预处理，初始化音频段，将音频数据样本转化为 float32 类型的浮点数据；音频数据的采样率设置为 16 kHz，对音频数据进行音量归一化，归一化的音量分贝值设为 −20 dB，以避免语音信号过载或过弱；使用语速扰动增强的音频增强手段，语速扰动因子的设置值从 0.9、1.0、1.1 中随机选择一个速度比例应用于音频段，使音频段的播放速度有所不同，增强模型的鲁棒性，语速扰动因子在±10%范围内能有效提升模型对语速变化的鲁棒性，且不会显著扭曲音素结构。

通过在音频数据的频谱图上掩盖随机的时间和频率区域来实现数据增强。其中频率掩码的宽度范围为（0，8），时间掩码的宽度范围为（0，10）。以频率掩码操作为例，首先获取输入张量的形状，即频谱图的批量大小，时间维度和频率维度，然后随机生成每个样本的频率掩码长度和位置，对于每个样本，在给定的频率掩码宽度范围，即（0，8）内随机生成一个频率掩码长度，具体步骤如下：生成一个张量，张量大小为（batch _ size，1），生成的掩码长度张量再将其扩展为（batch _ size，1，1）的形状，以便后续与频率维度进行广播运算。

同时，在每个样本的特征维度上，随机生成一个起始位置作为掩码的起始点。具体步骤为：生成一个张量，张量大小为（batch _ size，1），生成的掩码位置张量也将其扩展为（batch _ size，1，1）的形状。掩码生成后将掩码位置上的特征值置为零，再返回处理后的频谱图，时间掩码操作同理，在时间维度上进行掩码操作，再将处理过后的音频信息作为输入数据通过改进后的模型。

3.3 对比实验

为了验证本文提出模型对语种识别任务的有效性，使用 4 个模型分别在 Common Voice 数据集和 CSS10 数据集上进行实验，4 个模型分别是 ECAPA-TDNN，ResNetSE，TDNN 以及本文提出的模型，比较其识别准确率。实验在 PyTorch 深度学习框架之下进行，设备显卡为 NVIDIA GeForce RTX 3090，迭代次数为 100。其结果分别如表 3 和表 4 所示。

表 3 在 Common Voice 数据集上的语种识别准确率 %

模型	汉语	日语	法语	德语	俄语	平均准确率
ECAPA-TDNN	90.21	88.17	83.02	85.38	91.62	87.68
ResNetSE	88.36	81.58	78.81	83.90	86.05	83.74
TDNN	87.06	85.52	82.44	83.72	89.61	85.67
本文模型	96.55	92.47	92.06	89.96	93.51	92.91

由表 3 可知，在 Common Voice 数据集上，本文提出的模型平均识别准确率最高，达到了 92.91%，相比 ECAPA-TDNN 模型的 87.68%提高了 5.23%，相比 ResNetSE 和 TDNN 分别提高了 9.17%和 7.24%，实验中所选取的 5 种语言的识别准确率均有提升。

表 4 在 CSS10 数据集上的语种识别准确率 %

模型	汉语	日语	法语	德语	俄语	平均准确率
ECAPA-TDNN	86.97	82.59	88.64	82.55	87.30	85.61
ResNetSE	88.84	79.17	86.06	83.35	78.83	83.25
TDNN	87.91	81.56	81.55	78.32	85.01	82.87
本文模型	92.35	88.61	89.02	88.79	93.18	90.39

由表 4 可知，在 CSS10 数据集上，本文提出的模型平均识别准确率最高，达到了 90.39%，相比 ECAPA-

TDNN 模型的 85.61%提高了 4.78%，相比 ResNetSE 和 TDNN 分别提高了 7.14%和 7.52%，实验中所选取的五种语言的识别准确率也均有提升。

该实验结果证明，对于语种识别任务，本文提出的模型能有效提高语种识别准确率。

3.4　消融实验

为了验证各模块的性能和有效性，本文设计了对照模型 1 和对照模型 2 进行消融实验。其中，对照模型 1 是在 ECAPA-TDNN 的基础上采用了前端卷积模块 FCM 而没有采用注意力编码器模块，对照模型 2 则是采用了注意力编码器模块而没有采用前端卷积模块 FCM，并与同时采用了前端卷积模块和注意力编码器模块的本文模型进行对比，结果如表 5 所示。

表 5　在 Common Voice 和 CSS10 数据集上进行的消融实验 %

模型	FCM	注意力编码器模块	Common Voice	CSS10
ECAPA-TDNN	×	×	87.68	85.61
对照模型 1	√	×	88.11	85.94
对照模型 2	×	√	91.36	88.05
本文模型	√	√	92.91	90.39

由表 5 可知，在 Common Voice 和 Css10 数据集上，采用了前端卷积模块 FCM 的对照模型 1 相比 ECAPA-TDNN 的语种识别准确率略有小幅提升，这说明前端卷积模块 FCM 通过多层残差卷积与时频解耦降采样，有效提升了模型对局部时频特征的捕捉能力，FCM 中的残差块通过逐层堆叠的 3×3 卷积操作，逐步提取语音信号的低阶到高阶特征。对照模型 1 在 Common Voice 与 CSS10 数据集上相比 ECAPA-TDNN 分别提升了 0.43%和 0.33%，表明残差结构增强了模型对短时语音模式（如声调转折）的敏感度。而 FCM 在时间维度采用步幅为 2 的降采样，而在频率维度保持完整分辨率，这种设计减少了冗余时间帧的干扰，同时保留了对语种敏感的频段信息，从而改善了跨语种混淆问题。

而采用了注意力编码器模块的对照模型 2 在两个数据集上相比，ECAPA-TDNN 的识别准确率分别提升了 3.68%和 2.44%，这表明了注意力机制能够捕捉跨时间步的语种关键特征，比如部分语种的连读规则或辅音聚类，注意力编码器模块通过计算时间步之间的相关性权重，可聚焦于语音中的长时韵律信息、语调升降等，而传统卷积结构受限于局部感受野，难以建模此类长程依赖，同时也证明了多个子空间特征融合的有效性，注意力编码器模块的 8 个注意力头将输入特征投影到不同子空间，分别关注语音信号的不同属性，这种并行处理机制增强了模型对复杂语种特征的识别能力。而同时采用了前端卷积模块和注意力编码器模块的本文模型在两个数据集上相比对照模型 1 有 4.8%和 4.45%的提升，相比对照模型 2 有 1.55%和 2.34%的提升，对比 ECAPA-TDNN 有 5.23%和 4.78%的提升，证明了本文所提出方法的有效性。

4　结束语

本文针对语种识别任务面对的语种差异性大、语音特征复杂等导致的识别准确率问题，提出了一套用于语种识别任务的新方法，在对输入音频进行预处理之后提取得到音频的 FBank 特征，并使用时间掩码、频率掩码等数据增强手段进行进一步的数据处理，输入至以 ECAPA-TDNN 模型为基础加以改进和优化得到的语种识别模型，为了提升对于局部特征的提取能力，特别设计了一个前端卷积模块，该模块以多层残差块为核心，通过多级卷积实现对语音信号的逐步特征提取，以及在时间和频率维度分开进行下采样操作，不仅能够捕获局部和全局的时频特征，还通过通道数的调整生成了高度压缩和融合的特征表示，为后续的模块提供了更高效的输入。同时，以多头自注意力机制为核心，设计了一个编码器模块，针对语种分类任务的需求，模块结构和参数设置均经过专门优化，以适配音频特征的全局建模要求，并充分利用全局与局部的时序信息。该模块具体采用两层编码器，每层由多头自注意力机制和前馈神经网络组成。

经过实验得出，本文所提出的方法在 Common Voice 数据集上的识别准确率达到了 92.91%，对比原 ECAPA-TDNN 模型的 87.68%提升了 5.23%，在 CSS10 数据集上达到了 90.39%，对比 ECAPA-TDNN 的 85.61%提升了 4.78%，结果证明了本文所提出方法的有效性和泛化性。同时，针对了前端卷积模块和注意力编码器模块分别进行了消融实验，只应用了前端卷积模块的方法对比 ECAPA-TDNN 模型在两个数据集均有小幅提升，而只应用了注意力编码器模块的方法则有 3.68%和 2.4%的提升，消融实验结果证明了前端卷积模块有效提升了模型对局部时频特征的捕捉能力和注意力编码器模块在长程特征建模方面的优势。未来，考虑增加更多语种和样本量，特别是针对一些小语种和容易混淆的语种识别进行优化。

参考文献：

[1] ALI B N，ISMAIL S，IMTINAN A，et al. Speech recognition using deep neural networks：a systematic review [J]. IEEE Access，2019，7：19143－19165.

[2] BAI Z X，ZHANG X L. Speaker recognition based on deep learning：an overview [J]. Neural Networks，2021，140：65－99.

[3] 陈　聪. 面向复杂环境的语种识别方法研究 [D]. 成都：电子科技大学，2023.

（下转第 163 页）