基于时频注意力Conformer的多尺度短语音说话人识别模型

杨 璐,张邦成,杨俊美*,曾德炉
(华南理工大学电子与信息学院,广东广州 510630)

摘 要: 基于短语音的识别任务由于数据短缺、特征提取不精确,是说话人识别(Speaker Recognition,SR)领域目前面临的挑战之一. 针对数据量匮乏场景下的短语音声纹特征提取和身份识别,本文设计了一种基于时频注意力和卷积增强的短语音说话人识别网络. 本文在Transformer编码器中引入时频注意力和卷积,提出一种称为时频注意力Conformer(Time-Frequency Attention Convolution-augmented Transformer,TFA-Conformer)的模块,充分利用时频域通道中的信息来计算从全局到局部的有效性权重,帮助模型捕获精确的声学特征,使得特征编码器在短语音(3 s以内)环境下生成具有高判别性的说话人特征向量. 本文在标准说话人数据集TIMIT和ST-CMDS上评估了所提出的有监督训练网络模型,在短语音条件下,其识别准确性等指标相比主流方法平均提升4.837%,并且在更短时间和更少数据量的语音段识别中有平均2.799%的相对提升. 本文提出模型的参数更少且计算复杂度更低,其适用于短语音场景的同时也更轻量化.

关键词: 说话人识别;短语音;时频域;自注意力;Conformer;声纹特征
基金项目: 广东省自然科学基金(No.2023A1515011281)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)08-2658-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241114

TFA-Conformer Based Network for Short Utterance Speaker Recognition

YANG Lu, ZHANG Bang-cheng, YANG Jun-mei*, ZENG De-lu
(*School of Electronics and Information Technology*, *South China University of Technology*, *Guangzhou*, *Guangdong* 510630, *China*)

Abstract: The recognition task based on short utterances is one of the challenges in the field of speaker recognition (SR) due to data scarcity and inaccurate feature extraction. In scenarios with limited data, this paper proposes a short utterance speaker recognition network based on time-frequency (T-F) attention and convolutional enhancement for feature extraction and identity recognition. We introduce a time-frequency attention module and a convolution module in the transformer encoder to propose a module called time-frequency attention conformer (TFA-Conformer), which helps the model capture precise acoustic features by utilizing information from T-F channels to calculate validity weights from global to local perspectives, thereby enabling the feature encoder to produce highly discriminative speaker embeddings under short utterance speech conditions (3 s or less). We evaluate the proposed supervised training network on datasets under short utterance conditions, and the recognition accuracy and other metrics of the proposed method are improved by 4.837% on average, higher than those of the mainstream methods. In condition with shorter duration and less data, the proposed method shows a relative improvement of 2.799% on average. Furthermore, it requires fewer parameters and lower computational complexity, making it not only suitable for short utterance scenarios but also more lightweight.

Key words: speaker recognition; short utterance; time-frequency domain; self-attention; Conformer; voiceprint features
Foundation Item(s): Natural Science Foundation of Guangdong Province (No.2023A1515011281)

收稿日期:2024-12-11;录用日期:2025-05-29;责任编辑:朱梅玉
*通讯作者:杨俊美

1　引言

随着语音信息在生物识别认证领域的日益广泛应用,关于说话人识别的研究也逐步深入发展,其主要任务是提取输入语音的声学特征,根据特征反映的音色、音强等语音特性,分析源说话人的身份. 说话人识别任务按输入内容可区分为文本依赖型和文本无关型,其中文本依赖型要求说话人在训练和测试阶段使用相同或预定义的文本内容,而文本无关型则不受特定短语限制,此时系统仅依赖说话人的声音特征完成说话人识别,难度相对文本依赖型任务更大;按任务目标和应用场景可区分为开集任务和闭集任务,开集任务中训练和测试时的说话人可能不同,属于说话人验证(Speaker Verification)问题,即确认输入语音是否属于某个特定目标说话人的1∶1比对二元判断任务,常用于设备解锁、银行交易等身份核验场景,而闭集任务则需要系统在已知有限说话人集合中进行匹配,即确定输入语音属于已知集合中的哪一个说话人的1∶N匹配多分类任务,常用于刑侦中匹配嫌疑人发言、会议记录中标注说话人等自动身份标注场景. 本文主要探讨针对短语音的文本无关的闭集说话人识别任务,短语音说话人识别具体指在持续时长3 s以内的语音片段中,准确识别或验证说话人身份的技术,是说话人识别领域近年热点研究方向之一.

传统说话人识别依赖大规模数据统计实现,在5~10 s标准语音条件下已能够取得良好性能,如高斯混合模型-通用背景模型(Gaussian Mixture Model-Universal Background Model,GMM-UBM)[1]和$\boldsymbol{i}$向量[2],在标准长语音语料库中得到高达98%以上的识别准确率,但随着语音时长减短,识别性能显著下降,传统算法对短语音的适配性有待提高. 随着深度学习发展,说话人识别的研究重心逐渐转向神经网络编码方法,利用深层网络建模提取说话人的语音特征进行身份分析,性能较依赖语料库的质量. 目前大多数说话人识别方法,如基于DNN的$\boldsymbol{d}$向量[3]、$\boldsymbol{x}$向量[4]等,在大规模语音数据集中使用深层网络提取高维特征,展现出优秀的身份识别性能,这些数据集通常包含人均几十条的长时间语音段,在实际应用场景中这种长段语音识别的情况并不常见,短语音运用场景相对更广,而常规识别算法在语音时长减小到3 s以内时,错误率可能增加15%~30%,于是如何在短语音条件下实现高准确率的识别逐渐成为说话人识别领域的研究重点.

利用小规模语料库对说话人语音特征进行建模时,亟需针对短语音的高判别性特征提取和建模方法,以提升说话人识别算法的准确率,近年的相关研究重在引入数据增强训练策略和新的模块结构[5]来弥补短时语音特征提取不足的问题,如生成式网络[6,7]、通道权重编码[8,9]和自注意力机制[10,11]等,并从时序建模[12,13]和跨模态融合[14]等方向深入研究. 然而,关于短语音的识别算法研究仍然不成熟,短语音不仅信息量有限,而且更容易受到噪声干扰,如何在短时语音环境中提取具有高判别性的声学特征,以及在保持模型轻量化的前提下提升建模鲁棒性,都是亟待解决的重要问题,这也是本文的研究主旨.

本文提出了一种基于时频注意力Conformer的特征处理模块,通过将时频域通道注意力和卷积增强融入Transformer编码器来建模短语音的多方面特征,配合传统说话人识别编码模块,弥补提取的特征向量信息不足的缺陷. 与说话人识别领域最先进的方法相比,本文提出的模型在短语音识别上表现出更高的准确性和鲁棒性. 本文主要贡献总结如下:(1)与其他依赖大量数据训练的识别算法不同,本文创新性地提出了在没有预训练条件或数据增强的情况下实现短语音识别任务的有效模型;(2)验证了从全局到局部的时频域信息和通道注意力机制在短语音说话人识别中的重要性,并提出了一种基于时频域和卷积增强自注意力机制的特征建模方法;(3)所提出的算法在数据量极其匮乏的任务中,展现出比主流算法更高的识别准确性,为短时语音的场景应用提供了可靠的效果支撑.

2　相关研究

近年来,短语音说话人识别研究主要围绕数据增强方法和深层特征提取算法展开,数据增强方法由传统信号预处理增强训练策略逐渐发展到基于深度生成式的增强算法,如基于GMM-UBM提出的生成对抗网络短语音样本补偿算法[6]和Wav2Vec[7],分别利用条件生成对抗网络和对比学习策略将短语音样本补偿为包含充分说话人身份信息的语音样本,有效解决短语音数据不足的问题,但却让训练流程繁琐化,提升训练成本. 因此,目前针对短语音说话人识别的更多方法专注于研究特征提取网络,提升声纹特征编码质量.

为了充分提取短语音特征,编码器需要具备深层特征挖掘的能力. 时间延迟神经网络是一种适用于建模序列依赖信号的神经网络模型,常用于序列信号编码,并在近年针对短语音的特征提取网络中广泛应用. 基于TDNN提出的因子化TDNN[15]和扩展TDNN[16]通过多粒度时序建模,增强针对短语音的判别性表征能力,两者都显著提升持续时长2 s左右的短语音识别性能;Yu等[17]提出的动态化TDNN(Dynamic TDNN)引入时间动态性,能适应不同时长语音信号的时变特性;Liu等[8]将TDNN结合多频率通道信息来提取特征,所提出的多尺度频率-通道注意力(Multi-scale Frequency-channel Attention,MFA)通过双路径设计表征多通道的

短语音说话人特征；Zi 等[9]提出了基于 TDNN 衍生的 Res2Block 构建的双向采样编码，实现多尺度特征聚集；Ecapa-TDNN[18]是 TDNN 体系中最具代表性的算法之一，引入了信息聚合、通道注意力机制进一步提升了说话人识别准确性，它的特征编码器由多个压缩激励残差块（SE-Res2Block）构成，该模块包括一个依赖通道上下文信息的维度缩放卷积结构和压缩激励（Squeeze-and-Excitation，SE）块[19]. SE 块在不同特征向量之间进行全局平均池化，并选择性地强调关键特征，这启发了本文频率通道上注意力权重计算方法的设计，但其使用的高维通道设计和统计池化层的全连接层导致模型参数量极大. 针对这一问题，本文设计了非级联半步连接的 SE-Res2Block，将其作为帧级特征编码器，轻量化结构的同时完成短语音声纹特征的初步提取，这部分内容将在下一节模型结构中作详细介绍.

在建模全局特征的编码器结构中，自注意力机制[20]展示了其独特优势. 在短语音说话人识别任务中，研究者们将 Transformer 的自注意力机制用于声纹特征编码器[10~12,21,22]，获取具备全局上下文信息的特征向量，有效提取全局依赖的身份信息；为了实现多角度信息挖掘，进一步提升短时语音特征提取的精细度，近年来衍生了不少关注多域特征的算法，ResdefNet[13]分阶段计算时频域信息的方法显著提升了识别精度，Wang 等[23,24]提出了并行运算的时域和频域分路，分隔域间干扰的同时考虑双域信息；此外还有利用时域、频域和谱域多维度信息提出的三域特征联合学习算法[14]，从多个维度丰富短语音的特征信息，虽然注意力机制在全局性的上下文特征捕获上表现优异，但对短时局部特征的提取能力存在局限性.

在建模短语音的局部特征时，循环网络和卷积神经网络（Convolutional Neural Networks，CNN）表现优异. 姜珊等[25]结合双向循环网络和门控循环单元，同时利用过去和未来的上下文信息解决短语音中提取局部信息不完整的问题；Li 等[26]采用残差连接的深层 CNN 提取端到端的说话人特征向量，在 2~5 s 短语音识别中取得显著提升；上述编码结构虽然能有效建模局部时序信息，但忽略了全局上下文建模能力. Conformer[27,28]模型将卷积结构引入 Transformer 编码器，在有效提取全局特征的同时大大提升了神经网络捕获局部特征的能力，并增强了模型的可解释性，其采用的深度可分离卷积[29]在不影响模型性能的情况下大幅度降低了计算复杂度和参数量，但 Conformer 仅考虑时域建模，模型缺失部分频域信息，对短时深度特征的解析能力仍有改进空间.

本文参考 Conformer 的思路并在多头自注意力机制中引入时频注意力池化块，分路并行计算时频域信息以避免域间信息相互干扰[30,31]. 分路信息最终通过矩阵乘法得到时频注意力权重矩阵 $\mathbf{TF}_{\text{att}} \in \mathbb{R}^{T\times F_{\text{dim}}}$，其中 T 和 F_{dim} 分别代表输入数据包含帧数和频率通道数. 该矩阵在点对点的细度上对应说话人特征向量 $\boldsymbol{h}_t \in \mathbb{R}^{T\times F_{\text{dim}}}$. 在此基础上本文提出时频注意力 Conformer 模块用作特征权重细化，结合注意力机制全局多域建模能力和卷积网络局部建模能力，为短语音识别集成了多层次的声纹信息.

3 模型结构

一般而言，基于深度学习的说话人识别模型框架由预处理模块、编码器和分类器组成[32]. 预处理阶段中，原始数据经过预加重和双端检测处理后，通过梅尔滤波器组转化为有效突出人声特征的梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCCs），这将作为模型特征编码器的原始信号输入.

本文提出网络的完整架构如图 1 所示，预处理后的 MFCCs 通过一维卷积神经网络实现维度扩展，然后进入多层 SE-Res2Block 构成的帧级特征编码器提取特征，编码器后紧接的时频注意力 Conformer 模块生成强调时频域和上下文局部信息的帧级特征，起到特征池化的作用. 随后经过含有块级特征均衡模块的分类器，帧级特征转化为能够匹配真实标签的块级特征，即说话人身份预测概率. 输出概率将与真实的说话人标签对比，计算交叉熵损失.

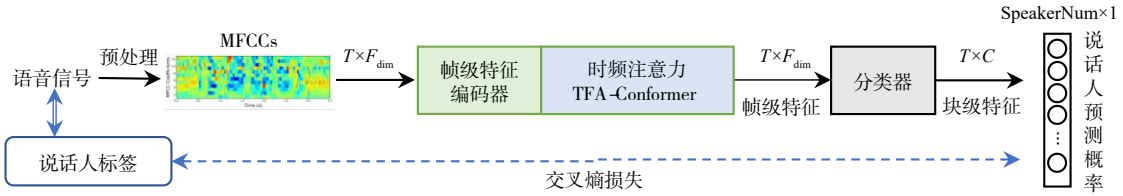


图 1 说话人识别网络有监督训练框架

3.1 非级联半步连接的 SE-Res2Block

SE-Res2Block[18]可用作帧级特征编码，其由两个卷积神经网络层夹着的 Res2Net[33]和一个 SE 块[19]组成，其中 Res2Net 将向量维度分为 n 个卷积组，通过缩小组内感受野提取多尺度特征，该特征随后通过 SE 块计算强调关键频率通道的权重.

本文提出的帧级特征编码器的详细结构如图 2 所示，首先使用一维卷积神经网络来扩充特征向量的维

度,增强特征表达能力. 在Macaron-Net[34]的启发下,本文引入半步残差连接堆叠N个SE-Res2Block,其中每一层的输入为上一层输出加上半倍的其余辅助层输出,通过加快传递权重更新提升训练效率. 与Ecapa-TDNN中的输出拼接形式不同,本文通过半步残差连接将多层输出的结果传递并汇聚到最后一层,直接使用最后一层的输出替代原始多层特征的简单拼接,输出隐藏特征向量维度由原来的$N\times C$缩减为最后单层SE-Res2Block的输出维度C,这能有效减少后端建模网络的参数量,并且由于输出向量只直接依赖于最后一个编码层,避免多层特征重复处理,减少跨层参数互扰. 当编码层输出向量维度缩减到原来的$1/N$时,也能有效避免由于向量维度过大且短语音样本量匮乏产生的多种训练问题,如过拟合、梯度爆炸和硬件受限等.

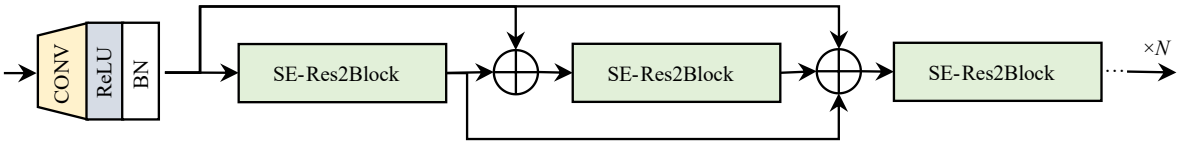


图2　帧级特征编码器

最终,帧级特征编码器输出包含通道注意力的多尺度局部特征,由于SE-Res2Block仅关注时变特征,为了得到固定维度的全局表征向量,本文紧接着提出时频注意力Conformer模块,捕获时频域信息的同时实现全局多层级池化,为短语音特征建模提供更深层的信息.

3.2　时频注意力Conformer模块(TFA-Conformer)

Transformer编码器通过多头自注意力机制建模向量之间的全局序列依赖关系,能处理具有长时连续特征的语音信号. 为了充分利用时域和频域提供的身份信息来获取全局上下文中局部位置之间的依赖关系,并使特征嵌入向量更适用于短语音任务,本文提出在帧级特征编码器后接入一个时频注意力Conformer模块,其结构如图3所示,它在Transformer编码器中引入时频注意力池化和夹层式深度可分离卷积,计算多视角特征的时频注意力加权向量. 类似于SE-Res2Block编码模块,TFA-Conformer模块的内部层同样使用半步残差连接方式.

首先,由于原始自注意力主要在频率域通道计算权重依赖关系,往往忽视了时间域中的能量分布,而这在建模说话人身份特征向量时也起着重要作用[32]. 于是本文在自注意力模块的子层间引入时频注意力池化块,其详细结构如图4所示.

该结构由一个时域分支和一个频域分支组成,特征向量在平均池化后输入到时频域分路计算注意力权重,其中的平均池化层分别从时间域$\boldsymbol{T}_{\text{att}}\in\mathbb{R}^{1\times T}$和频率域$\boldsymbol{F}_{\text{att}}\in\mathbb{R}^{F_{\text{dim}}\times 1}$提取注意力权重信息:

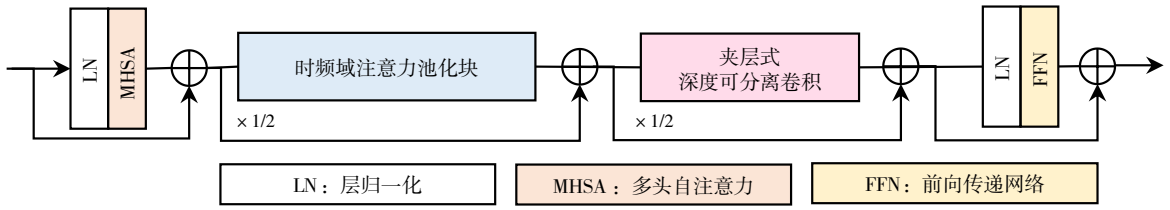


图3　时频注意力TFA-Conformer模块

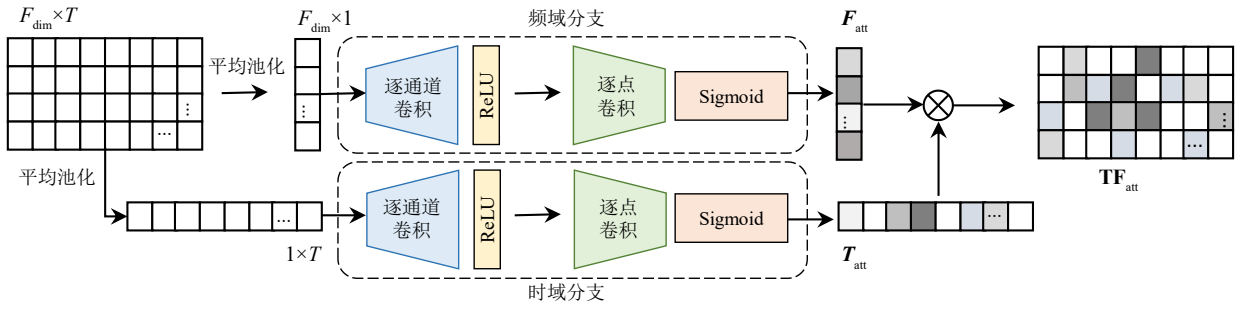


图4　时频注意力池化块

$$T_{\text{att}}(j)=\frac{1}{F_{\text{dim}}}\sum_{i=1}^{F_{\text{dim}}}\mu_{Fi}(j) \tag{1}$$

$$F_{\text{att}}(j)=\frac{1}{T}\sum_{i=1}^{T}\mu_{Ti}(j) \tag{2}$$

其中，$T_{\text{att}}(j)$和$F_{\text{att}}(j)$分别表示平均池化后的时域和频域一维向量的第j个维度数据，$\mu_{\cdot i}(j)$代表在·方向上第i个向量的第j维度的特征参数．与文献[24]中不同的是，本文在平均池化后的时频域分路中引入深度可分离卷积层进行域内特征提取，以增强模块表达细节的能力．深度可分离卷积将普通卷积的计算过程分为逐通道阶段和逐点阶段，逐通道卷积和逐点卷积分别实现域内维度扩展和压缩．本文在逐通道卷积引入空洞率，使域内分路在更广的感受野下捕获时频域能量信息．根据文献[29]中阐述的深度可分离卷积原理，在时频域分支实现的一维深度可分离卷积与常规卷积结构的参数量对比如下，默认卷积核大小k远小于输入特征维度F，此时$F-k+1\approx F$，其中N_{sep}和N_{std}分别表示深度可分离卷积和传统卷积的参数个数，$N_{\text{depthwise}}$和$N_{\text{pointwise}}$分别表示逐通道卷积和逐点卷积的参数量，C_{in}和C_{out}分别表示输入与输出通道数：

$$\begin{aligned}\frac{N_{\text{sep}}}{N_{\text{std}}}&=\frac{N_{\text{depthwise}}+N_{\text{pointwise}}}{k\cdot(F-k+1)\cdot C_{\text{in}}C_{\text{out}}}\\&=\frac{k\cdot(F-k+1)\cdot C_{\text{in}}+F\cdot C_{\text{in}}C_{\text{out}}}{k\cdot(F-k+1)\cdot C_{\text{in}}C_{\text{out}}}\\&\approx\frac{k\cdot F\cdot C_{\text{in}}+F\cdot C_{\text{in}}C_{\text{out}}}{k\cdot F\cdot C_{\text{in}}C_{\text{out}}}=\frac{k+C_{\text{out}}}{k\cdot C_{\text{out}}}\end{aligned} \tag{3}$$

所以，利用逐通道和逐点的深度可分离卷积机制，相比常规卷积结构能有效减少训练参数，在配置时频注意力池化块中$C_{\text{out}}=8$且$k=7$的条件下，计算成本减少为常规卷积的三分之一，这有利于构建更深的神经网络．两个域分支将最终通过矩阵乘法合并，得到时频域联合注意力映射谱：

$$\mathbf{TF}_{\text{att}}(i,j)=F_{\text{att}}(i)\otimes T_{\text{att}}(j) \tag{4}$$

其中，$\mathbf{TF}_{\text{att}}(i,j)\in\mathbb{R}^{F_{\text{dim}}\times T}$，$\otimes$代表向量乘法．通过时频域谱映射，模型能够更有效地综合时频域特征，提高语音建模性能．

其次，在多尺度特征聚合MFA-Conformer[27]的启发下，在Transformer编码器的两个子层间也引入深度可分离卷积，但在原始结构的基础上额外增添了一个逐点卷积层和GLU激活函数层，前置逐点卷积调整通道分布，结合GLU的动态门控机制实现通道筛选，其详细的卷积结构如图5所示．引入的卷积层专注局部特征建模，高效建模局部时序依赖，弥补多头自注意力只强调全局序列依赖建模的不足．

3.3 压缩激励的块级特征均衡模块

为了与说话人标签的独热编码(one-hot encoding)匹配，本文引入块级特征均衡[25]模块如图6，将帧级特征转换为与说话人标签对齐的块级特征．该模块主要包括平均池化层和用于维度匹配的全连接层．此外，本文在块级均衡之前插入一个SE块，以计算频率通道的注意力权重．这在特征融合时强调了有效的频率点信息，减少了冗余信息的传递．特征将被输入到平均池化层，生成在连续时间段内的平均特征向量：

$$\mu_T(i)=\frac{1}{T}\sum_{t=1}^{T}h_t(i) \tag{5}$$

其中，$h_t(i)$代表在经过压缩激励模块后输出向量的第t个分帧块的第i维度上的特征参数．紧接着的块级均衡模块计算流程如下：

$$F_{\text{out}}=\delta\left(f_{c2}\left(I\left(f_{c1}\left(\mu_T\right)\right)\right)\right) \tag{6}$$

其中，f_{c1}和f_{c2}代表控制维度扩张与收缩的全连接层，δ代表softmax函数，它将输出预测的说话人概率$F_{\text{out}}\in[0,\ 1]$．I表示L2归一化层，能够将数据分布映射到单位圆范围内．最后的输出格式是[Batch, SpeakersNum]，能够匹配说话人分类标签进行交叉熵损失函数计算．

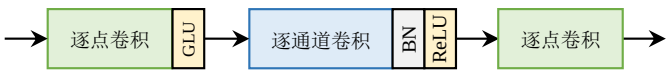


图5 夹层式深度可分离卷积

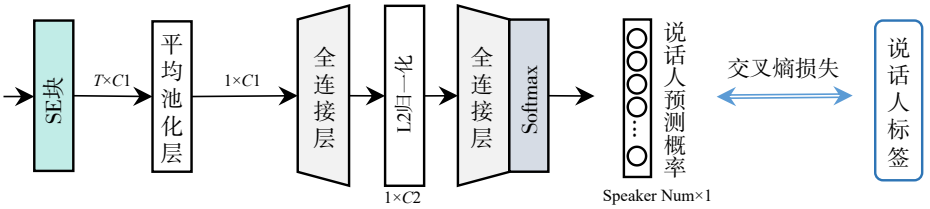


图6 块级特征均衡模块

4　实验

4.1　数据集

为了适配短语音任务,本文在实验中选用的数据集包括TIMIT[34]和ST-CMDS. TIMIT数据集由来自八个方言区的630名说话人组成,而ST-CMDS是一个中文语音数据集,包含来自855名说话人的十多万条语句. 为了将实验场景限制在短语音条件下,本文限制每个说话人的可训练语句数量为6条,并限制每条语音的持续时长为2.5 s. 基于短时连续性特征,输入特征由72维MFCCs特征F_{dim}组成,窗口长度为25 ms,帧移为10 ms. 其中,本文将网络的输入时间步长T设置为256,采样率设置为16 000 Hz.

4.2　模型参数设置

在本研究中,本文选择了以下在说话人识别中表现出色的主流算法作为基准:Bi-GRU[25]、ResCNN[25]、Ecapa-TDNN[18]和MFA-Conformer[27],表1给出了模型参数设置的详细说明.

表1　模型参数

模型分层	参数设置
1D CNN扩张维度	通道数C=512 (通道数代表输出层神经元数量)
SE-Res2Blocks	模块数量N=3
	Res2Net分组数n=6
多头自注意力机制	通道数C=512 注意力头数h=4
时频注意力池化块	通道数C=32
	逐通道卷积(k=7,d=3)
深度可分离卷积	逐通道卷积(k=17)
块级均衡模块	平均层通道数$C1$=512
	全连接层通道数$C2$=1 024
	输出通道数C=SpeakerNum

为了更好地适配短语音条件,将每种方法的嵌入特征维度F_{dim}统一设置为72维,更改输出维度为语料库对应的说话人数量. 除上述基础张量维度修改之外,Bi-GRU和ResCNN无其他参数更改;Ecapa-TDNN设置SE-Res2Block通道数为{512 512 512};MFA-Conformer设置Conformer模块数量为3.

4.3　训练步骤和评价标准

本文使用TITAN X GPU进行实验. 优化器采用Adam,初始学习率设定为0.000 5,每经过650步骤下降3%. 训练采用batchsize为64,并将数据集按3∶1∶1的比例划分为训练集、验证集和测试集. 以下是交叉熵损失计算的详细公式:

$$H(p,q)=-\sum_{i=1}^{\mathrm{Sp_{num}}}q(x_i^e)\log\left(p(x_i^e)\right) \tag{7}$$

其中,$\mathrm{Sp_{num}}$表示说话人总数,$q(\cdot)$表示真实说话人标签的独热值,$p(x_i^e)$表示说话人被预测为x_i^e的可能概率. 本文使用准确率作为评估说话人分类性能的准则:

$$\mathrm{ACC}=\frac{\mathrm{TP}}{\mathrm{Total}} \tag{8}$$

其中,TP代表分类正确的实例数量,Total代表样本总数,最终结果将分别计算验证集和测试集的识别准确率. 为了避免因语音样本数量和质量分布不均衡引起对准确率计算的干扰,本文也将平均精确率和平均召回率作为评估准则,即分别计算每个说话人的精确率和召回率结果的平均值:

$$\mathrm{Precision_{avg}}=\frac{1}{\mathrm{Sp_{num}}}\sum_{i=1}^{\mathrm{Sp_{num}}}\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i} \tag{9}$$

$$\mathrm{Recall_{avg}}=\frac{1}{\mathrm{Sp_{num}}}\sum_{i=1}^{\mathrm{Sp_{num}}}\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i} \tag{10}$$

其中,TP_i和FN_i分别代表第i个说话人的样本正确分类和错误分类的个数,FP_i代表将非第i个说话人的样本错误分类为当前说话人的个数.

4.4　对比实验及消融实验

首先,本文在TIMIT和ST-CMDS数据集上展开实验. 为了在短语音条件下完成训练,本节限制每个说话人的可训练语句数量为6条,每条语音的持续时长为2.5 s. 本文提出的算法将与四种先进的说话人识别算法进行比较. 表2展现了在两个数据集上不同算法的准确率性能、参数量和浮点运算次数(FLoating Point Operations,FLOPs). 下面四项是消融实验,其展现了本文引入Transformer编码器中的深度可分离卷积模块、块级均衡模块中的SE块和时频注意力池化块的有效性.

由于帧级特征编码器去除了输出级联,且在时频注意力Conformer模块中引入了深度可分离结构,卷积核运算无需混合所有通道数据,这进一步降低模型结构的复杂度,使其参数总量低于其他模型. 同时,帧级特征编码器引入的半步残差连接提升了运算效率,FLOPs指标反映其模型计算复杂度也低于大部分对比算法. 根据表3可以看出本文的方法在分类精确率和召回率上的性能表现优于其他方法,在减少了模型参数量的同时在测试数据集上表现更好. 根据精确率和召回率的概念,可以验证本文提出的方法相对其他说话人识别算法针对该数据集中的少数类具有更好的识别表现,受多数类造成的样本类总体分布偏移影响较小,能够更好地处理类别不平衡的应用场景.

对比其他说话人识别算法,上述实验验证了本文算法显著的识别性能,主要受益于以下几点模型结构的创新与调整:

(1)编码器主要利用TDNN结构处理短时语音信号,通过时间延迟机制显式建模时长依赖关系,分层时

表2 各种算法的实验结果

算法	参数量/M	FLOPs/G	验证准确率/%		测试准确率/%	
			TIMIT	ST	TIMIT	ST
Bi-GRU[25]	6.59	4.92	89.84	88.61	89.06	88.28
ResCNN[26]	10.9	7.92	96.09	93.36	95.31	92.19
Ecapa[18]	6.76	4.90	97.27	95.70	96.09	94.53
MFA-L3[27]	7.47	3.06	96.48	94.14	95.27	94.53
Ours(-Conv)	5.51	2.81	97.27	96.88	95.70	94.92
Ours(-SE)	6.05	3.46	95.31	96.48	94.53	95.70
Ours(-TFA)	6.30	3.45	96.88	96.09	94.14	95.31
Ours	6.31	3.46	98.83	97.27	97.66	96.88

注:“-Conv”“-SE”和“-TFA”分别表示去除对应模块的消融对比实验.

表3 不同算法的平均精确率和召回率 单位:%

算法	平均精确率		平均召回率	
	TIMIT	ST	TIMIT	ST
Bi-GRU[25]	91.32	90.64	90.25	88.35
ResCNN[26]	91.84	90.70	91.06	90.59
Ecapa[18]	93.54	91.27	94.18	93.58
MFA[27]	96.01	95.18	94.97	93.53
Ours	98.04	96.24	97.70	97.28

序卷积有效保留语音信号局部结构信息,展现出对短时语音建模的显著优势,弥补了ResCNN在全局依赖建模上的不足,对快速变化的短时特征建模能力也优于Bi-GRU的循环结构;

(2)池化模块创新性地采用Transformer编码结构替代传统全局池化,利用多头自注意力机制的相对位置编码计算序列元素间的长距离相关性,对比Ecapa-TDNN有效提升了模型对全局特征的建模准确率;

(3)在Transformer编码器中引入了夹层式深度可分离卷积,根据时间顺序建模局部细节特征,以自适应权重融合自注意力生成的全局特征和卷积提取的局部特征;并提出了时频注意力池化块融合时域和频域信息,生成特征权重映射谱,相比MFA-Conformer的单一时域卷积,多域特征同时包含帧间动态和频谱分布特征,弥补单时域建模对频谱结构的细粒度特征(如共振峰、谐波结构)捕捉不足,对于短语音而言获取更丰富的信息有利于增强特征的表征能力,从而提升识别性能;

(4)从模型轻量化分析,本文算法引入的深度可分离卷积显著减少了模型参数量,在特征编码器中引入的半步残差连接也显著降低计算复杂度,通过结构优化实现了计算效率的突破.

4.5 不同数据量条件下的实验对比

在ST-CMDS数据集中,本文考虑设置不同的短语音原始数据条件:本文为每位说话人选择不同数量的可训练语音段并改变单个语音块的截取帧长度,然后在不同的算法下分别训练网络模型,这能帮助本文评估模型在不同语音数据量条件下的鲁棒性[35,36]. 首先,本文设置每位说话人的待训练说话人样本数量分别为3、6、9和12. 然后,本文比较在单个语音段不同持续时长的条件下的准确率,分别选择每条语音段持续时长为128、192和256帧,即分别持续大约1.25 s、1.75 s和2.5 s.

如图7展示,柱状图左侧纵坐标对应准确率指标,右侧纵坐标对应虚线展示的模型相对提升率指标. 当将语音数量减少到一定程度时本文提出的方法在指标上明显优于其他算法的表现,这说明提出算法在限制样本规模的条件下更加适用. 相比之下,本文的方法分别表现出平均2.968%和2.632%相对Ecapa-TDNN和MFA-Conformer的优势百分比,这表示该方法能够在更少数据的情境下展现出相对其他方法更好的性能.

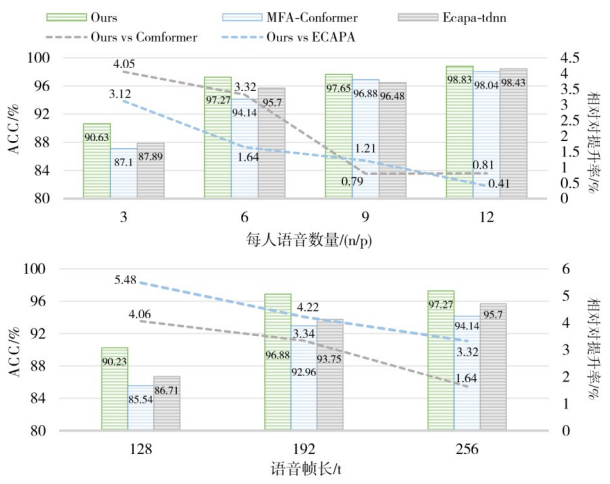


图7 不同短语音条件下的实验结果

5　结论

本文提出了一种针对短语音说话人识别的有效网络,其结合时频注意力Conformer模块,能够关注上下文局部和全局特征,同时在时频域并行分析信息以从短语音段中提取更精细的特征. 经过实验比较,本文的方法在识别准确性和模型参数效率方面展现出优势,即使在数据样本较少的情况下,它也表现出优越的性能,为样本匮乏条件下的语音识别提供了应用基础.

参考文献

[1] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41.

[2] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.

[3] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Interspeech 2017. Los Angeles: ISCA, 2017: 999-1003.

[4] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 5329-5333.

[5] HE Y Y, KANG Z H, WANG J Z, et al. Voiceextender: Short-utterance text-independent speaker verification with guided diffusion model[C]//2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway: IEEE, 2023: 1-8.

[6] HU Z F, FU Y Q, LUO Y, et al. Speaker recognition based on short utterance compensation method of generative adversarial networks[J]. International Journal of Speech Technology, 2020, 23(2): 443-450.

[7] BAEVSKI A, ZHOUH, MOHAMED A, et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations[EB/OL]. (2020-10-22) [2025-05-05]. https://arxiv.org/abs/2006.11477v3.

[8] LIU T C, DAS R K, LEE K A, et al. MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 7517-7521.

[9] ZI Y F, XIONG S W. Short-duration speaker verification by joint filter superposition-based multi-dimensional central difference feature extraction and Res2Block-based bidirectional sampling[J]. IEEE Transactions on Consumer Electronics, 2024, 70(3): 5128-5141.

[10] WANG R, AO J Y, ZHOU L, et al. Multi-view self-attention based transformer for speaker recognition[C]// ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 6732-6736.

[11] ZHU Y K, MAK B. Bayesian self-attentive speaker embeddings for text-independent speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1000-1012.

[12] KARTHIKEYAN V, SUJA PRIYADHARSINI S. A stacked convolutional neural network framework with multi-scale attention mechanism for text-independent voiceprint recognition[J]. Pattern Analysis and Applications, 2024, 27(2): 48.

[13] ZHANG Y M, YU H, MA Z Y. Speaker verification system based on deformable CNN and time-frequency attention[C]//2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE, 2020: 1689-1692.

[14] ZI Y F, XIONG S W. Multi-fisher and triple-domain feature enhancement-based short utterance speaker verification for IoT smart service[J]. IEEE Internet of Things Journal, 2024, 11(4): 6044-6055.

[15] SNYDER D, GARCIA-ROMERO D, SELL G, et al. Speaker recognition for multi-speaker conversations using X-vectors[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 5796-5800.

[16] VILLALBA J, CHEN N X, SNYDER D, et al. State-of-

the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18[C]// Interspeech 2019. Los Angeles: ISCA, 2019: 1488-1492.

[17] YU Y Q, LI W J. Densely connected time delay neural network for speaker verification[C]//Interspeech 2020. Los Angeles: ISCA, 2020: 921-925.

[18] DESPLANQUES B, THIENPONDT J, DEMUYNCK K, et al. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[EB/OL]. (2020-08-10)[2025-05-05]. https://arxiv.org/abs/2005.07143v3.

[19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVFConference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 1: 30.

[21] ZHU H N, LEE K A, LI H Z. Serialized multi-layer multi-head attention for neural speaker embedding[EB/OL]. (2021-07-14)[2025-05-05]. https://arxiv.org/abs/2107.06493v1.

[22] MARY N J M S, UMESH S, KATTA S V. S-vectors and TESA: Speaker embeddings and a speaker authenticator based on transformer encoder[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 404-413.

[23] WANG H L, ZOU Y X, CHONG D D, et al. Environmental sound classification with parallel temporal-spectral attention[EB/OL]. (2020-05-21)[2025-06-05]. https://arxiv.org/abs/1912.06808v3.

[24] ZHANG Q Q, SONG Q, NI Z H, et al. Time-frequency attention for monaural speech enhancement[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 7852-7856.

[25] 姜珊, 张二华, 张晗. 基于Bi-GRU+BFE模型的短语音说话人识别[J]. 计算机与数字工程, 2022, 50(10): 2233-2239.
JIANG S, ZHANG E H, ZHANG H. Speaker recognition under short utterance based on Bi-GRU+BFE model[J]. Computer & Digital Engineering, 2022, 50(10): 2233-2239. (in Chinese)

[26] LI C, MA X, JIANG B, et al. Deep speaker: An end-to-end neural speaker embedding system[EB/OL]. (2017-05-05)[2025-05-05]. https://arxiv.org/abs/1705.02304.

[27] ZHANG Y, LV Z Q, WU H B, et al. MFA-conformer: Multi-scale feature aggregation conformer for automatic speaker verification[EB/OL]. (2022-11-11) [2025-05-05]. https://arxiv.org/abs/2203.15249v2.

[28] CHANG O, LIAO H, SERDYUK D, et al. Conformer is all you need for visual speech recognition[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2024: 10136-10140.

[29] ZHANG C, CHEN W, XU C. Depthwise separable convolutions for short utterance speaker identification[C]// 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Piscataway: IEEE, 2019: 962-966.

[30] ZHEN YANG, TIANLANG WANG, HAIYAN GUO, et al. Speaker verification method based on cross-domain attentive feature fusion[J]. Journal on Communications, 2023, 44(8): 89-98.

[31] ZHANG Q Q, QIAN X Y, NI Z H, et al. A time-frequency attention module for neural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 462-475.

[32] HAJAVI A, ETEMAD A. A study on bias and fairness in deep speaker recognition[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5.

[33] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE TransPattern AnalMachIntell, 2021, 43(2): 652-662.

[34] LU Y P, LI Z H, HE D, et al. Understanding and improving transformer from a multi-particle dynamic system point of view[EB/OL]. (2019-06-06)[2025-05-05]. https://arxiv.org/abs/1906.02762v1.

[35] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et

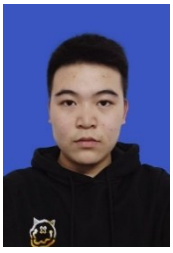
al. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. (2012-07-03)[2025-05-05]. https://arxiv.org/abs/1207.0580v1.

[36] NOVOSELOV S, VOLOKHOV V, LAVRENTYEVA G. Universal speaker recognition encoders for different speech segments duration[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5.

作者简介

杨　璐　女,2000年生. 现为华南理工大学电子与信息学院硕士研究生. 研究方向为语音信号处理领域的声纹识别.
E-mail: 202221013354@mail.scut.edu.cn

杨俊美　女,2009年3月至今在华南理工大学电子与信息学院任教. 主要研究方向为智能信号处理、自适应滤波、图像超分辨率重建、语音去混响等.
E-mail: yjunmei@scut.edu.cn



张邦成　男,2000年生. 现为华南理工大学电子与信息学院硕士研究生. 研究方向为语音信号处理领域的语音分离.
E-mail: 202221013363@mail.scut.edu.cn

曾德炉　男,在华南理工大学电子与信息学院任教. 研究方向为数学与信息等交叉理论及应用.
E-mail: dlzeng@scut.edu.cn