



DOI:10.12404/j.issn.1671-1815.2600566

引用格式:郑展恒, 李嘉麒, 王健, 等. 融合分组建模与多维注意力机制的轻量级说话人确认方法[J]. 科学技术与工程, 2026, 26(27): 11804-11811.

Zheng Zhanheng, Li Jiaqi, Wang Jian, et al. Lightweight speaker verification method based on fusion of grouped modeling and multi-dimensional attention mechanism[J]. Science Technology and Engineering, 2026, 26(27): 11804-11811.

电子技术、通信技术

融合分组建模与多维注意力机制的轻量级说话人确认方法

郑展恒^{1,2}, 李嘉麒^{1,2*}, 王健^{1,2}, 李琦¹, 曾庆宁^{1,2}

(1. 桂林电子科技大学信息与通信学院, 桂林 541004; 2. “认知无线电与信息处理”教育部重点实验室 (桂林电子科技大学), 桂林 541004)

摘 要 近年来,深度学习在说话人确认任务中取得显著进展,但现有模型普遍存在参数规模大、时序建模能力有限、难以在边缘设备部署等问题。为此,提出一种融合分组建模与多维注意力机制的轻量级说话人确认模型。该模型基于改进的时间延迟神经网络(time-delay neural network, TDNN)结构,设计通道分组时序编码模块(channel-split temporal encoding block, CSTE-Block)与多尺度分组残差编码模块(grouped scaled Res2net, GSRes2Block),用以增强帧级特征的时序建模与多尺度表达能力;引入说话人聚合优化模块(speaker embedding refinement module block, SERMBlock)与多维注意力机制,提升全局上下文的建模深度与特征聚焦能力。全局嵌入通过注意力统计池化生成,并采用加性角度边界 Softmax 分类器与监督对比损失联合优化嵌入空间判别性。实验在 CN-Celeb 数据集上进行,结果表明,所提模型在等错误率(equal error rate, EER)和最小检测代价函数(minimum detection cost function, minDCF)方面分别达到 14.82% 和 0.694 8,参数量较 ECAPA-TDNN(emphasized channel attention, propagation and aggregation in time delay neural network)减少约 69.4%;同时在 VoxCeleb1 跨语种验证集上展现出良好的鲁棒性。复杂度分析显示,该模型兼具高性能与高效率,具备优良的部署适应性与实际应用潜力。

关键词 深度学习;说话人确认;轻量化;分组建模;多维注意力;联合损失函数

中图法分类号 TN912.34; 文献标志码 A

Lightweight Speaker Verification Method Based on Fusion of Grouped Modeling and Multi-dimensional Attention Mechanism

ZHENG Zhan-heng^{1,2}, LI Jia-qi^{1,2*}, WANG Jian^{1,2}, LI Qi¹, ZENG Qing-ning^{1,2}

(1. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; 2. Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology), Guilin 541004, China)

[Abstract] In recent years, deep learning has made significant progress in speaker verification tasks. However, existing models often suffer from large parameter sizes, limited temporal modeling capability, and poor deployability on edge devices. To address these issues, a lightweight speaker verification model incorporating grouped modeling and multi-dimensional attention mechanisms was proposed. The model was built upon an improved time-delay neural network (TDNN) architecture and introduced a channel-split temporal encoding block (CSTE-Block) and a grouped scaled Res2Net block (GSRes2Block) to enhance frame-level temporal modeling and multi-scale feature representation. In addition, a speaker embedding refinement module block (SERMBlock) and multi-dimensional attention were employed to improve global contextual modeling and feature focusing. A global speaker embedding was generated via attentive statistical pooling and jointly optimized using an additive angular margin Softmax classifier and supervised contrastive loss to enhance the discriminative power of the embedding space. Experiments on the CN-Celeb dataset demonstrate that the proposed model achieves an equal error rate (EER) of 14.82% and a minimum detection cost function (minDCF) of 0.694 8, with a parameter reduction of approximately 69.4% compared to the ECAPA-TDNN(emphasized channel attention, propagation and aggregation in time delay neural network) baseline. The model also exhibits strong cross-lingual robustness on the VoxCeleb1 evaluation set. Complexity analysis

收稿日期: 2026-01-26 修订日期: 2026-03-06

基金项目: 国家自然科学基金(62064003); 认知无线电与信息处理教育部重点实验室项目(CRKL230103)

第一作者: 郑展恒(1979—),男,汉族,陕西榆林人,硕士,高级实验师。研究方向:信号处理。E-mail: glzzh@guet.edu.cn。

* 通信作者: 李嘉麒(2001—),男,汉族,陕西宝鸡人,硕士研究生。研究方向:说话人确认。E-mail: 1419371311@qq.com。

投稿网址: www.stae.com.cn

further confirms that the model offers a favorable trade-off between performance and computational efficiency, showing excellent potential for real-world deployment.

[**Keywords**] deep learning; speaker verification; lightweight modeling; grouped modeling; multi-dimensional attention; joint loss function

说话人确认^[1] (speaker verification, SV)旨在从语音信号中提取表征说话人身份的特征,并判断测试语音是否来自特定注册说话人,广泛应用于金融风控、远程认证、智能安防和语音交互等^[2-6]场景。受限于噪声干扰、语音非平稳特性及个体发音差异,传统方法在实际部署中面临鲁棒性不足等挑战。

早期以高斯混合模型与通用背景模型构建的 i-vector 系统^[7],通过提取如梅尔频率倒谱系数 (mel-frequency cepstral coefficients, MFCC) 或对数梅尔滤波器组能量 (filter banks, FBank) 等声学特征获得低维说话人向量,配合 PLDA (probabilistic linear discriminant analysis) 判别器完成匹配判定。尽管该方法在部分场景下表现稳定,但在变长语音、跨信道或复杂环境下的泛化能力受限。

随着深度学习的发展, x-vector 系统^[8]基于时间延迟神经网络^[9]引入帧级建模与统计池化机制,显著提升了说话人特征的判别性。后续研究引入 ResNet^[10]提升特征表达深度, ECAPA-TDNN (emphasized channel attention, propagation and aggregation in time delay neural network)^[11]则融合多尺度残差结构、通道注意力机制^[12]与多层特征聚合^[13],有效增强了嵌入性能,成为当前主流方法之一。

然而,上述模型结构复杂、参数冗余、推理开销大,不利于在边缘设备、移动终端等资源受限平台部署。为提升模型实用性,研究者尝试采用分组卷积^[14]、通道重排、通道剪枝与网络重构实现模型轻量化,同时引入注意力机制增强鲁棒性。但这些方法多源自图像领域,难以充分适配语音任务中帧级建模、跨通道交互及嵌入优化等需求,导致语音特征表达能力受限。

为此,提出一种融合分组建模与多维注意力机制的轻量级说话人确认模型 (grouped split attention-

enhanced temporal embedding TDNN, GSATE-TDNN)。该方法基于改进的 TDNN 框架,集成多维注意力机制与模块化设计,兼顾模型判别力与计算效率,在复杂语音场景下展现出良好性能与部署适应性。本文研究的主要特点如下。

(1) 引入通道分组时序编码模块 (channel-split temporal encoding block, CSTE-Block), 提升模型对帧级时间依赖的建模能力。

(2) 提出级联式多尺度分组残差编码模块 (grouped scaled Res2net block, GSRes2Block), 通过通道分组、级联式建模与注意力融合实现跨尺度语音特征的高效表征。

(3) 通过构建说话人聚合优化模块 (speaker embedding refinement module block, SERMBlock), 弥补通道交互不足与关键帧特征不突出的局限,进一步强化判别表示。

(4) 结合注意力统计池化 (attentive statistics pooling, ASP) 与联合对比损失优化嵌入空间结构,增强嵌入的类间分离性与类内紧致性,提高复杂语音条件下的泛化能力。

(5) 在 CN-Celeb 与 VoxCeleb1 数据集上开展系统评估,结果表明该模型在识别精度、模型规模与计算效率方面优于现有对比方法,具备良好的部署潜力。

1 模型

GSATE-TDNN 模型架构由帧级建模模块、多尺度残差编码模块、特征聚合模块和全局嵌入生成模块构成,如图 1 所示。

输入语音首先提取对数梅尔滤波器组能量 (FBank) 特征,形成尺寸为 $[F, T]$ 的二维声谱图 (F, T 分别为在频域和时域维度的长度)。随后输入

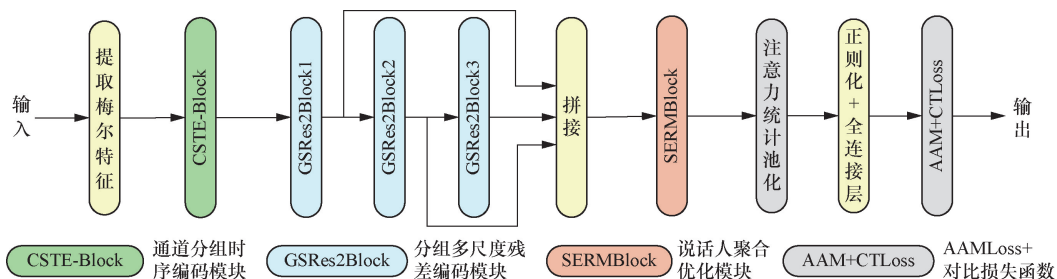


图1 GSATE-TDNN 模型结构

Fig.1 GSATE-TDNN model structure

CSTE-Block,通过分组门控机制建模帧级时间依赖关系,增强局部与全局上下文表达。

在多尺度编码阶段,模型依次堆叠三层 GSRes2Block,各层内部通过通道分组、扩张卷积与轻量化注意力机制实现不同感受野下的并行建模,增强帧级特征的多尺度表达能力。三层输出特征通过级联融合后送入 SERMBlock,以进一步优化语义表征并强化关键说话人特征的通道聚焦性。

聚合阶段采用注意力统计池化生成全局说话人嵌入,随后通过 BN(batch normalization) 层、全连接层完成特征变换与压缩。最终输出嵌入在训练阶段通过 AAM-Softmax 与对比损失函数联合优化,以提升嵌入空间的类间判别性与类内一致性。

1.1 通道分组时序编码模块

为提升语音帧级特征的时间建模能力,并兼顾计算效率与结构紧凑性,提出通道分组时序编码模块(CSTE-Block),其结构如图2所示。该模块主要由分组时序建模路径、轻量残差掩码路径与恒等残差连接三部分组成。

输入的 Fbank 特征首先被暂存,并沿通道维度划分为两组,分别送入两条建模路径。主干路径中,特征经一维卷积升维后送入通道分组时序卷积(temporal split convolution, TS-Conv)进行分组建模,该结构通过对通道子组并行建模以提取时间依赖信息,有效降低参数规模,提升在资源受限设备中的推理效率。随后,经卷积降维并接入 PReLU 激活函数输出建模结果。另一支路通过轻量残差掩码机制对初始输入进行时序注意力建模,学习跨通道的时序注意力权重,以突出高相关帧并抑制冗余干扰。该掩码被用于对主干建模结果进行加权增强,从而提升关键语义的响应强度。最终,主干增强路径的输出与原始输入通过恒等映射完成残差融合,获得判别性更强的帧级时序表示。

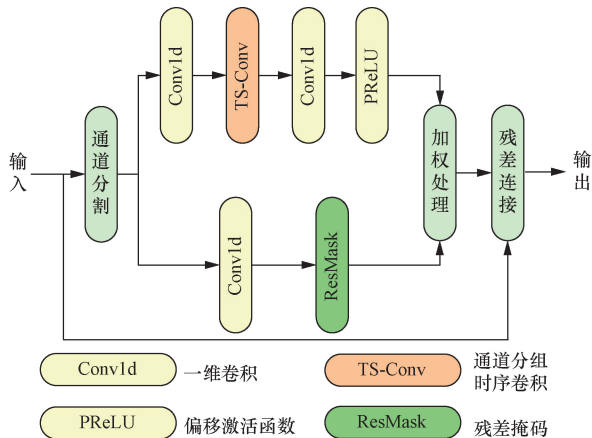


图2 通道分组时序编码模块

Fig.2 CSTE-Block

CSTE-Block 以分组计算与门控加权为核心,兼顾表达能力与结构轻量化,为后续多尺度残差结构提供了高质量语义输入。

1.2 分组多尺度残差编码模块

为提升模型对多尺度上下文语音特征的建模能力,提出 GSRes2Block。该模块在 Res2Net 框架基础上引入通道分组与轻量化注意力机制,主要由分组深度可分离卷积链、跨组信息融合、残差连接及通道注意力门控组成。在兼顾建模性能与计算效率的同时,实现跨尺度语音特征的高效表征。其中单一模块结构如图3所示。

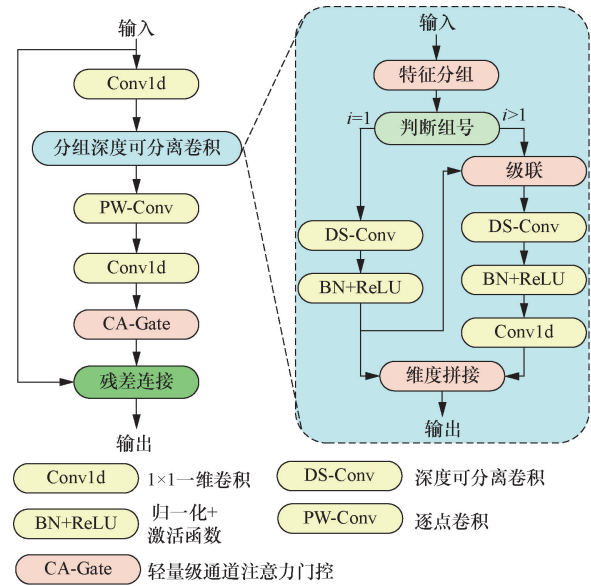


图3 分组多尺度残差编码模块

Fig.3 GSRes2Block

具体而言,输入特征首先通过 1×1 卷积进行通道压缩,随后进行分组深度可分离卷积,将特征划分为多个子组,分组方式表达式为

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_s], \quad \mathbf{X}_i \in \mathbf{R}^{B \times \frac{C}{s} \times T} \quad (1)$$

式(1)中: $\mathbf{X}_i$ 为第 i 个特征子组; B 为批量样本数; C 为总通道特征维度; s 为分组数; T 为时间步长。

第一个子组采用深度可分离卷积提取局部特征,结合批归一化与非线性激活函数提升特征稳定性与建模能力,输出特征 $\mathbf{Y}_1$ 表达式为

$$\mathbf{Y}_1 = \text{ReLU} \{ \text{BN} [\text{DWConv}(\mathbf{X}_1)] \} \quad (2)$$

其余子组通过与前一组的输出级联后输入下一组,通过特征级联实现跨组上下文建模,捕捉多尺度时序依赖,输出特征 $\mathbf{Y}_i$ 表达式为

$$\mathbf{Y}_i = \text{ReLU} \{ \text{BN} [\text{DWConv}(\text{Concat}[\mathbf{Y}_{i-1}, \mathbf{X}_i])] \} \quad (3)$$

所有子组的输出进行维度拼接,并通过逐点卷积恢复整体特征维度,形成多尺度融合表征 $\mathbf{Z}$, 表达式为

$$\mathbf{Z} = \text{PWConv}(\text{Concat}[\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_i]) \quad (4)$$

同时,引入轻量级注意力门控,对分组处理后的特征进行空洞卷积增强长时依赖建模,并通过全局平均池化、门控激活与逐通道变换利用全局上下文信息对通道进行显式加权,引导模型聚焦于语义显著的通道维度,提升对说话人区分性特征的关注。

模型主干网络串联部署三个 GSRes2Block 模块,分别聚焦于不同深度层次的多尺度信息提取。各阶段输出通过跳跃连接方式在通道维度拼接,融合浅层与深层多尺度语义,有效提升帧级特征的表达完整性,为后续聚合优化模块提供更丰富的特征基础。

相较 ECAPA-TDNN 中采用的 SE-Res2Net 结构,GSRes2Block 在保持多尺度建模能力的同时,减少了全连接注意力所带来的参数与计算负担,通过分组结构与轻量门控策略显著减少计算成本,适配于资源受限设备的实时部署场景。

1.3 说话人聚合优化模块

多尺度残差编码后的帧级特征仍存在跨通道信息交互不足及关键帧表示不突出的问题,为进一步提升多尺度残差编码后的帧级特征在语义聚合和关键帧建模方面的表现,设计说话人聚合优化模块(SERMBlock),其结构如图4所示。

首先,模块输入首先通过分组深度可分离卷积在各子通道进行上下文建模,并送入两个并行的分组门控循环单元(grouped gated recurrent unit, G-GRU)以提取不同子空间的长短期依赖。随后通道重排(channel shuffle)^[15]操作打破通道孤立性,实现跨组语义交互,增强跨通道表征能力。

为提升通道维度的判别性,引入说话人感知通道优化子模块(speaker-aware channel refinement block, SACRBlock)。该模块借鉴前馈网络思想,结合深度可分离卷积与通道差异建模机制,旨在以较低的计算成本强化通道间的互补关系与表达能力。

SACRBlock 由主路径和两条残差路径组成,主

路径用于建模全局信息特征,残差路径用于保存基础通道特征以参与后续差异建模。具体而言,输入特征 $\mathbf{X}_{\text{shuffle}}$ 首先通过 3×3 深度卷积提取基础通道特征表达,再接入 GeLU 激活函数与随机失活 Dropout 提升非线性建模能力与鲁棒性,基础通道表征 $\mathbf{X}_{\text{base}}$ 可表示为

$$\mathbf{X}_{\text{base}} = \text{Drop}\{\text{GeLU}[\text{DWConv}(\mathbf{X}_{\text{shuffle}})]\} \quad (5)$$

得到的基础通道特征被暂存后使用分解卷积(decomposition convolution, D-Conv)进行结构解耦,经过激活函数处理保留全局语义特征;第一条恒等残差路径保存的基础通道特征通过消减与全局语义特征的相似部分,获得更关注局部细节的特征 $\mathbf{X}_{\text{local}}$,表达式为

$$\mathbf{X}_{\text{local}} = \mathbf{X}_{\text{base}} - \text{GeLU}[\text{D-Conv}(\mathbf{X}_{\text{base}})] \quad (6)$$

随后,结合可学习缩放因子(ElementScale),实现通道级差异响应的自适应加权,进一步增强关键语义维度,加权融合输出 $\mathbf{Y}$,表达式为

$$\mathbf{Y} = \mathbf{X}_{\text{base}} + \alpha \mathbf{X}_{\text{local}} \quad (7)$$

式(7)中: α 为可学习缩放因子,其初值设为较小常数(如 1×10^{-5}),用于抑制初始训练阶段残差注入的扰动,随着训练逐步强化通道的语义区分性。两路径输出通过可学习缩放因子进行差异调控并加权融合,实现通道级差异响应的自适应加权,进一步增强关键语义维度。

在帧级建模方面,为强化对说话人判别帧的建模能力,引入时间选择性增强模块(temporal selective enhancement, TSE),通过时间维度平均池化与注意力加权机制,为每一帧分配动态权重,强化语义集中帧、抑制冗余帧干扰,从而提升嵌入表征对关键内容的响应性。

最终,融合后的特征经过 1×1 卷积、批归一化与激活函数映射至目标通道维度,并与初始输入残差连接,输出聚合后的帧级语音表示。该模块在保持结构轻量的同时,增强了特征聚合与通道建模能力,为后续嵌入生成与系统判别性能提供坚实基础。

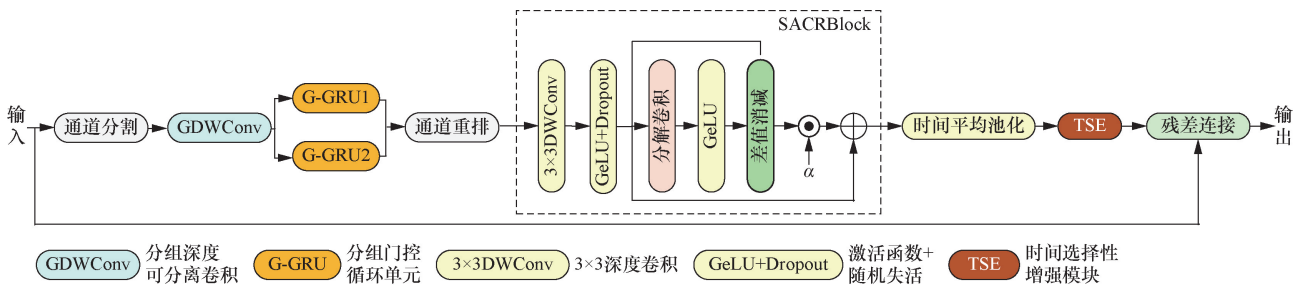


图4 说话人聚合优化模块

Fig. 4 SERMBlock

1.4 全局嵌入生成与联合损失优化

经多层特征建模后,帧级特征送入注意力统计池化(ASP)层。该模块通过学习的注意力权重对帧级表示进行加权建模,计算其加权均值与标准差并拼接,生成固定维度的段级说话人表示。随后,该表示通过映射层压缩至低维嵌入空间,用于后续的相似度计算与判别训练。

在损失函数设计方面,采用加性角度边界 Softmax 分类器 (additive angular margin softmax, AAM-Softmax)^[16] 作为主损失函数,引入角度边界提升类间区分性,其计算公式为

$$L_{\text{AAM}} = -\frac{1}{N} \sum_{i=1}^N \lg \exp S[\cos(\theta_{y_i}) - m] / \{ \exp S[\cos(\theta_{y_i}) - m] + \sum_{j \neq y_i} \exp S \cos(\theta_j) \} \quad (8)$$

式(8)中: S 为缩放因子; m 为角度间隔; θ_{y_i} 为第 i 个样本与其真实类别中心之间的角度; N 为批次大小。

同时,融合监督式对比损失^[17] (supervised contrastive loss, SupConLoss) 与温度归一化对比损失^[18] (normalized temperature-scaled cross-entropy loss, NT-XentLoss), 进一步压缩类内分布、增强语义一致性。其计算公式分别为

$$L_{\text{SC}} = \sum_{i=1}^N \frac{-1}{|P(i)|} \sum_{p \in P(i)} \lg \frac{\exp(z_i z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i z_a / \tau)} \quad (9)$$

$$L_{\text{NT}} = - \sum_{i=1}^N \lg \frac{\exp[\text{sim}(z_i, z_j) / \eta]}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp[\text{sim}(z_i, z_k) / \eta]} \quad (10)$$

式中: $P(i)$ 为与样本 i 同类的正样本集合; $A(i)$ 为当前批次所有样本集合; z_i 为样本 i 在嵌入空间中的特征表示; τ 为温度系数,通常设为 0.07; η 为温度参数,通常设为 0.5。

三者联合构成总损失函数,定义为

$$L = L_{\text{AAM}} + \lambda_1 L_{\text{SC}} + \lambda_2 L_{\text{NT}} \quad (11)$$

式(11)中: λ_1 与 λ_2 分别为 SupConLoss 和 NT-XentLoss 的权重。在实验过程中,对两种对比损失函数的权重进行多次调试和自适应优化权重,发现当初始权重 $\lambda_1 + \lambda_2 = 0.01$ 时,可以在不引起参数膨胀的情况下达到模型分类的最佳效果,在提升模型表达能力的同时,兼顾了稳定性和计算效率。

通过上述联合优化策略,模型在保持轻量结构的同时显著提升嵌入空间的判别性与泛化能力,为复杂场景下的说话人确认任务提供更加稳定可靠的表示基础。

2 实验结果与性能分析

2.1 数据集

为验证所提模型在复杂语音场景下的识别性能与跨语种鲁棒性,实验选用 CN-Celeb^[19] 与 Vox-Celeb1^[20] 数据集进行训练与评估。

CN-Celeb 数据集由中国科学院声学研究所发布,涵盖 CN-Celeb1 与 CN-Celeb2 两部分,共计约 4 000 名说话人、65 万条语音片段,累计时长超过 800 h,语音内容涉及脱口秀、新闻、访谈等多种真实场景,具备高语言复杂性与噪声多样性。采用 CN-Celeb 的全部语音作为训练集,使用其官方验证集进行模型性能评估。

VoxCeleb1 数据集包含 1 251 名说话人,语音来自 YouTube 视频,覆盖多种语言与口音,具有典型的跨语种、远场、多背景噪声特性。为测试模型的跨语种泛化能力,实验仅以 CN-Celeb 为训练语料,不使用 VoxCeleb1 中任何样本进行训练,直接在其测试集上进行推理验证。

所有语音数据均重采样至 16 kHz,并统一执行音量归一化与静音裁剪等预处理。训练过程中采用 PK-Sampler 策略构建每批样本,以保证说话人类别分布均衡并提升训练稳定性。

2.2 实验设置

实验在 Windows 11 环境下进行,硬件平台包括 Intel Core i5-12600KF 处理器 (3.7 GHz) 与 GeForce RTX 4060Ti 显卡,模型基于 PyTorch 2.3.0 框架实现,Python 版本为 3.10.14, CUDA 版本为 11.8。

语音数据采样率统一为 16 kHz,训练采用 Adam 优化器,初始学习率设为 0.001,批量大小为 64。学习率调度使用 Warmup Cosine 策略,预热周期为 5 轮,最大学习率为 0.001,最小学习率降至 1×10^{-5} 。模型总训练轮数为 60,训练过程中保存验证集表现最优的模型参数。

联合损失中,监督式对比损失 (SupConLoss) 与温度归一化对比损失 (NT-XentLoss) 的权重均设为 0.005,以平衡主分类损失与嵌入空间优化目标。

2.3 消融实验

为验证各子模块在整体性能中的有效性,以 ECAPA-TDNN 为基线,采用逐步构建策略,依次引入 CSTE-Block、GSRes2Block、SERMBlock 及联合对比损失函数,在 CN-Celeb 数据集上开展消融实验,实验结果如表 1 所示。

基础模型在未加入任何模块的情况下,等错误率 (equal error rate, EER) 为 15.79%,最小检测代价函数 (minimum detection cost function, minDCF)

表 1 CN-Celeb 消融实验

Table 1 CN-Celeb ablation experiment

CSTE-Block	GSRes2Block	SERMBlock	对比损失函数	EER/%	minDCF
×	×	×	×	15.79	0.7404
√	×	×	×	15.92	0.7385
√	√	×	×	15.56	0.7165
√	√	√	×	15.09	0.7021
√	√	√	√	14.82	0.6948

为 0.740 4。引入 CSTE-Block 后,minDCF 降至 0.738 5,说明通道分组结构虽未显著改善错误率,但对帧级时序建模的稳定性具有一定提升作用。继续加入 GSRes2Block 后,EER 降至 15.56%、minDCF 降至 0.716 5,表明多尺度残差结构在提升跨尺度语义表达方面具有显著效果。进一步引入 SERM-Block,模型性能持续提升,EER 降至 15.09%,minDCF 降至 0.702 1,说明该模块通过通道交互与帧选择机制有效增强了特征聚合能力。最终,联合引入对比损失函数后,模型在两个指标上均达到最优:EER 为 14.82%,minDCF 为 0.694 8,验证了联合监督策略对嵌入空间结构优化的正向作用。上述实验结果表明,各模块在提升模型判别性与鲁棒性方面均发挥了积极作用,整体设计具备良好的模块增益特性与系统优化潜力。

2.4 对比实验

为全面评估所提模型的综合性能与轻量化效果,选取 ResNet34^[21]、ERes2Net^[22]、ECAPA-TDNN 及 CAM++^[23]等主流说话人确认模型作为对比方法,在 CN-Celeb 数据集上进行实验。所有模型均在相同训练策略与预处理流程下构建,评估指标包括参数量(Params)、等错误率(EER)与最小检测代价函数(minDCF)。实验结果如表 2 所示,其中加粗部分表示本文模型指标。

表 2 展示了不同模型在测试集上的性能指标对比。结果表明,所提模型在 EER 和 minDCF 指标上分别达到 14.82%与 0.694 8,在各方法中分别取得最优和次优结果,同时参数量仅为 4.5 M,优于目前主流的说话人确认模型。具体而言,相较于经典深层网络 ResNet34 和 ERes2Net,所提模型通过 GTRUBlock

表 2 CN-Celeb 对比实验

Table 2 CN-Celeb comparative experiment

模型	Params/M	EER/%	minDCF
ResNet34	6.7	15.99	0.7212
ERes2Net	6.6	15.18	0.6694
ECAPA-TDNN	14.7	15.79	0.7404
CAM++	7.2	15.55	0.7158
本文模型	4.5	14.82	0.6948

与 GSRes2Block 的联合建模有效利用帧级上下文信息,在显著压缩模型规模的同时进一步提升识别精度,表现为更低的 EER 和参数量;与当前主流高性能模型 ECAPA-TDNN 相比,尽管后者依赖多尺度建模与通道注意力机制实现嵌入判别性的增强,但其计算复杂度较高;而所提模型在保持更小参数规模的前提下,依然获得了更优的识别性能。与此同时,CAM++作为 ECAPA-TDNN 的轻量化改进版本,在引入上下文感知掩蔽模块后取得一定的精度-效率平衡,但所提模型在 EER 进一步下降的同时,仍保持更小的参数规模,进一步提升了模型的整体性能与部署适应性。

综上所述,所提模型在保证识别精度的同时显著降低了模型复杂度,具备良好的性能-效率折中能力,为资源受限设备下的说话人确认任务提供了具备实用潜力的解决路径。

2.5 复杂度分析

为进一步评估所提模型在资源受限场景下的部署潜力,需对其计算复杂度与推理效率进行分析,并与目前主流的说话人确认模型如 ResNet34、ECAPA-TDNN 及 CAM++进行对比。实验统一在 Intel i5-12600KF CPU 平台上完成,评估指标包括模型参数量(Params)、每次前向传播所需的浮点运算量(floating-point operations per second,FLOPs)以及实时因子(real-time factor,RTF)。结果如表 3 所示。

从表 3 可见,所提模型在参数量与 FLOPs 方面均显著低于其他对比方法,仅为 4.5 M 和 1.40 G,分别较 ECAPA-TDNN 减少约 69.4%和 77.2%,显示出极佳的结构紧凑性与计算效率;在 RTF 方面,GSATE-TDNN 的推理时间约为 0.011,远低于实时性阈值 1,相较于 ResNet34 和 ECAPA-TDNN 分别提升约 66.7%和 65.6%,表现出更优的处理速度;相比同样以轻量化为目标且保持模型性能均衡的 CAM++模型,GSATE-TDNN 在保持更小参数量与计算量的同时,进一步压缩 RTF 至更低水平,展现出在语音验证等对实时性和资源消耗敏感场景中的广泛适用性。

综合来看,所提方法在性能与复杂度之间实现了良好权衡,具备较强的边缘部署可行性与工程实用价值。

表 3 计算复杂度对比

Table 3 Computational complexity comparison

模型	Params/M	FLOPs/G	RTF
ResNet34	6.7	3.96	0.033
ECAPA-TDNN	14.7	6.84	0.032
CAM++	7.2	1.72	0.013
本文模型	4.5	1.40	0.011

2.6 泛化性能分析

为进一步验证所提模型在跨语种与复杂声学条件下的泛化能力,本实验在未使用 VoxCeleb1 数据的前提下,仅以 CN-Celeb 全量数据作为训练集,直接在 VoxCeleb1 验证集上进行测试。该设置用于评估模型在语言、口音及背景噪声等变化条件下的适应性。对比方法包括经典的 ResNet34 与 ERes2Net 结构、主流高性能模型 ECAPA-TDNN 以及轻量化模型 CAM++。各模型均在相同训练策略与预处理流程下进行训练,测试指标包括 EER 与 minDCF,实验结果如表 4 所示。

从结果可以看出,所提模型在 EER 与 minDCF 指标上分别达到 14.25% 与 0.873 3,优于所有对比模型。相比 ECAPA-TDNN 与 CAM++,所提模型在保持低复杂度的基础上进一步提升识别精度,体现出更强的跨语种鲁棒性。该优势归因于 GSATE-TDNN 所引入的分组建模、语义增强与注意力机制,有效提升了模型在未知语言分布下的特征一致性与判别能力。

综上,所提模型在不依赖目标语言样本训练的前提下,仍展现出优异的泛化性能,验证了其在多语言与真实复杂语音场景下的实际应用潜力。

表 4 泛化性能对比

Table 4 Comparison of generalization performance

模型	Params/M	EER/%	minDCF
ResNet34	6.7	16.95	0.923 1
ERes2Net	6.6	16.85	0.903 7
ECAPA-TDNN	14.7	16.41	0.942 5
CAM++	7.2	16.35	0.930 6
本文模型	4.5	14.25	0.873 3

3 结论

针对现有说话人确认模型在帧级时序建模能力弱、参数规模大及部署效率低等问题,提出一种融合分组建模与多维注意力机制的轻量级说话人确认模型——GSATE-TDNN。通过实验对所提出模型进行验证,得到以下结论。

(1) 引入通道分组时序编码模块(CSTE-Block),通过通道划分与并行建模机制,有效提升帧级时间依赖的建模能力,同时控制循环单元的参数规模。

(2) 提出分组多尺度残差结构(GSRes2Block),基于级联式残差建模与通道注意力调控,实现跨尺度上下文信息的高效融合与压缩表达。

(3) 构建说话人聚合优化模块(SERMBlock),集成分组深度卷积、通道重排、SACR 注意力机制与 TSE 帧选择策略,增强跨通道语义交互与关键帧关注能力。

(4) 全局嵌入通过注意力统计池化生成,并结合 AAM-Softmax 与对比损失联合优化,有效提升了嵌入空间的判别性与泛化能力。

(5) 实验结果表明,所提模型在识别精度与计算效率之间实现了良好的平衡,与现有主流模型相比,在小幅提升识别性能的同时(EER 与 minDCF 进一步下降),模型参数量压缩至仅 4.5 M,显著降低了计算资源消耗;复杂度分析显示,所提模型在 FLOPs 与 RTF 方面均为最低,具备良好的推理效率和部署友好性;同时,该模型在未参与训练的 VoxCeleb1 验证集上表现出稳定的跨语种鲁棒性,进一步验证了其泛化能力。未来将考虑将该方法推广至更广泛的资源受限场景,如多语种语音认证、远程交互及嵌入式终端等任务中,具有良好的实际应用潜力。

参考文献

- [1] Sharma R, Govind D, Mishra J, et al. Milestones in speaker recognition[J]. Artificial Intelligence Review, 2024, 57(3): 58.
- [2] Katav A, Moshe Y, Cohen I. A framework for robust speaker verification in highly noisy environments leveraging both noisy and enhanced audio[C]//2025 33rd European Signal Processing Conference (EUSIPCO). New York: IEEE, 2025: 41-45.
- [3] Weng S, Yang L, Mao J. Effective modeling of critical contextual information for TDNN-based speaker verification[J]. arXiv Preprint, arXiv: 2509.09932, 2025.
- [4] 张葛祥, 曾鑫, 姚光乐, 等. 说话人识别综述[J]. 控制工程, 2025, 32(2): 251-264.
Zhang Gexiang, Zeng Xin, Yao Guangle, et al. A survey of speaker recognition[J]. Control Engineering of China, 2025, 32(2): 251-264.
- [5] 王华朋, 冯嘉琪. 基于深度学习的语音增强方法综述[J]. 科学技术与工程, 2025, 25(20): 8331-8346.
Wang Huapeng, Feng Jiaqi. Review of speech enhancement methods based on deep learning[J]. Science Technology and Engineering, 2025, 25(20): 8331-8346.
- [6] 万玫汐, 王华朋, 闫道申, 等. 基于改进 ECAPA-TDNN 的法庭自动说话人识别[J]. 科学技术与工程, 2024, 24(27): 11763-11773.
Wang Meixi, Wang Huapeng, Yan Daoshen, et al. Forensic automatic speaker recognition based on enhanced ECAPA-TDNN[J]. Science Technology and Engineering, 2024, 24(27): 11763-11773.
- [7] Song Y, Liu F, Shen T. PLDA in i-vector based underwater acoustic signals classification[J]. Ships and Offshore Structures, 2024, 19(3): 366-374.
- [8] Zhang Y, Liu L. Multi-task learning for X-vector based speaker recognition[J]. International Journal of Speech Technology, 2023, 26(4): 817-823.
- [9] Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Interspeech. Dresden: International Speech Communication Association, 2015: 3214-3218.
- [10] Zhou T, Zhao Y, Wu J. Resnext and res2net structures for

- speaker verification[C]//2021 IEEE Spoken Language Technology Workshop (SLT). New York: IEEE, 2021: 301-307.
- [11] Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification[J]. arXiv Preprint, arXiv: 2005. 07143, 2020.
- [12] Hu J, Shen L, Sun G. Squeeze-and-excitation net-works[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132-7141.
- [13] Liu T, 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). New York: IEEE, 2022: 7517-7521.
- [14] Tesfai H, Saleh H, Al-Qutayri M, et al. Lightweight shufflenet based on CNN for arrhythmia classification[J]. IEEE Access, 2022, 10: 111842-111854.
- [15] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: practical guidelines for efficient CNN architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer International Publishing, 2018: 116-131.
- [16] Li Z, Liu Y, Li L, et al. Additive phoneme-aware margin softmax loss for language recognition[J]. arXiv Preprint, arXiv: 2106. 12851, 2021.
- [17] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.
- [18] Ågren W. The NT-Xent loss upper bound[J]. arXiv Preprint, arXiv: 2205. 03169, 2022.
- [19] Fan Y, Kang J W, Li L T, et al. CN-Celeb: a challenging Chinese speaker recognition dataset[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2020: 7604-7608.
- [20] Nagrani A, Chung J S, Xie W, et al. VoxCeleb: large-scale speaker verification in the wild[J]. Computer Speech & Language, 2020, 60: 101027.
- [21] Jiahong L, Jie B, Yingshuang C, et al. An adaptive resnet based speaker recognition in radio communication[C]//2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT). New York: IEEE, 2021: 161-164.
- [22] Chen Y, Zheng S, Wang H, et al. An enhanced Res2Net with local and global feature fusion for speaker verification[J]. arXiv Preprint, arXiv: 2305. 12838, 2023.
- [23] Wang H, Zheng S, Chen Y, et al. Cam++: a fast and efficient network for speaker verification using context-aware masking[J]. arXiv Preprint, arXiv: 2303. 00332, 2023.