DOI:10.19651/j.cnki.emt.2518455

融合空洞卷积与多尺度注意力的说话人确认*

李嘉麒[1,2] 郑展恒[1,2] 曾庆宁[1,2] 王 健[1,2]
(1.桂林电子科技大学信息与通信学院 桂林 541004; 2.桂林电子科技大学认知无线电与信息处理教育部重点实验室 桂林 541004)

摘 要:针对复杂语音环境下 CAM++模型在特征提取与识别性能方面存在的不足,本文提出了一种融合空洞卷积与时频多尺度注意力机制的说话人确认模型 TF-DCAM。该模型首先利用空洞残差卷积与时频重聚焦机制增强特征提取能力,提升对冗余信息的抑制效果;其次引入时频多尺度注意力模块,通过通道注意力与跨纬度交互机制提升模型对关键信息的感知能力;再通过自适应掩码时序卷积模块强化长时依赖建模;最后采用对比损失函数联合优化嵌入空间结构。实验在 CN-Celeb 数据集上表明,TF-DCAM 在 EER 和 minDCF 上分别相较基线模型降低了 14.98%和 10.98%;在 VoxCeleb1 上亦展现出良好的跨语种泛化能力。结果证明所提方法在保证轻量化的同时显著提升了说话人确认性能与鲁棒性。

关键词:深度学习;说话人确认;时频多尺度注意力;空洞卷积;对比损失函数
中图分类号:TN912.34 **文献标识码**:A **国家标准学科分类代码**:510.4040

Speaker verification method based on dilated convolution and multi-scale attention mechanism

Li Jiaqi[1,2] Zheng Zhanheng[1,2] Zeng Qingning[1,2] Wang Jian[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: To address the limitations of the CAM++ model in feature extraction and recognition performance under complex acoustic conditions, this paper proposes TF-DCAM, a speaker verification model integrating dilated convolution and temporal-frequency multi-scale attention mechanisms. The model enhances feature representation through dilated residual convolution and a time-frequency adaptive refocusing unit to suppress redundant information. A temporal-frequency multi-scale attention module is introduced to improve sensitivity to key information via channel attention and cross-dimensional interaction. An adaptive masking temporal convolution module is further incorporated to model long-term dependencies effectively. Finally, a combination of contrastive loss functions is applied to jointly optimize the speaker embedding space. Experiments conducted on the CN-Celeb dataset show that TF-DCAM reduces EER and minDCF by 14.98% and 10.98% respectively, compared with the baseline. The model also demonstrates strong cross-lingual generalization on the VoxCeleb1 dataset. Results indicate that the proposed method significantly improves speaker verification performance and robustness while maintaining model efficiency.

Keywords: deep learning; speaker verification; temporal-frequency multi-scale attention; dilated convolution; contrastive loss

0 引 言

说话人确认技术旨在判断语音是否来自某个特定说话人[1],广泛应用于身份验证、生物识别和智能交互等场景。传统方法通常通过分析语音信号的时域与频域特征进行说话人识别。近年来,随着深度学习的迅速发展,基于神经网络的说话人确认模型在准确性和鲁棒性方面取得了显著进展。主流模型多采用卷积神经网络(convolutional neural networks, CNN)[2]、时延神经网络(time delay neural network, TDNN)[3]及残差网络(residual networks,

收稿日期:2025-03-27
*基金项目:认知无线电与信息处理教育部重点实验室项目(CRKL230103)资助

ResNet)[4]等架构,通过多层次特征提取和信息融合,有效提升了语音嵌入的判别能力。

现有说话人确认系统一般包括声学特征提取器、嵌入编码器和相似性计算模块[5-6]。特征提取常用梅尔频率倒谱系数(mel-frequency cepstral coefficient,MFCC)[7]或对数梅尔滤波器组能量(log-mel filter bank energy,FBank)[8]等表示语音的时频信息。尽管这些方法在静态条件下有效,但在复杂环境中存在鲁棒性差、特征利用不足的问题,难以充分建模语音信号中的结构性信息。

为解决上述问题,深度学习驱动的说话人嵌入模型不断发展,许多新模型相继提出。其中,密集连接时延神经网络(densely connected time delay neural network,D-TDNN)[9]通过引入密集连接机制,降低计算复杂度并提升建模能力;ECAPA-TDNN(emphasized channel attention propagation and aggregation in time delay neural network)[10]进一步引入 Res2Net[11]和通道注意力机制(squeeze-and-excitation network,SE)[12]在增强时间上下文与频率建模方面取得显著进展;CAM++在 D-TDNN 结构基础上引入上下文感知掩蔽(context-aware masking,CAM)[13]模块,并通过多粒度池化优化了计算效率与性能之间的平衡。然而,这些模型在复杂语音条件下仍存在特征提取不足、计算开销较大的问题。

近年来,多尺度特征融合与时频联合建模逐渐成为提升语音建模能力的重要方向[14-16]。国外研究者在图像识别、语音分离和语音识别等任务中广泛应用空洞卷积、金字塔结构和多尺度注意力机制以增强模型的感受野与特征选择能力;国内学者则提出将时间与频率维度上的信息融合,通过多尺度卷积核[17]、频带划分策略、多特征融合[18]等方法提升语音特征的丰富性与鲁棒性。与此同时,特征重组机制如通道重聚焦、自适应掩码和动态加权等也被用于去除冗余信息、突出关键成分,显示出良好的应用前景。然而,现有方法在多尺度信息建模的系统性、时频注意力融合的充分性,以及深度时序建模的稳定性方面仍存在一定不足,难以兼顾识别精度与计算效率。

基于此,本文提出了一种基于时频空洞卷积与自适应掩码机制模型(temporal-frequency dilated convolution and adaptive masking model,TF-DCAM)。首先,采用空洞卷积模块(dms-convNet,DMS),利用空洞残差卷积(dilated-resblock)扩展感受野,并结合时频自适应重聚焦单元(time-frequency adaptive refocusing unit,TF-ARU)提升语音特征提取质量并有效抑制冗余信息;其次,引入时频多尺度注意力模块(temporal-frequency multi-scale attention module,TF-MAM),结合通道注意力、时间注意力及跨纬度交互机制,实现时频信息的高效融合;此外,为优化深度时序卷积网络,改进时序建模结构,引入自适应掩码时序卷积模块(adaptive masking temporal convolution,AMTC),通过多尺度池化与动态加权机制强化对长时依赖的建模;最后,采用对比损失函数联合优化策略,在主损失函数 AAMLoss 的基础上,引入 SupConLoss 和 NT-XentLoss 作为对比损失函数,以提升嵌入空间的判别能力与泛化能力。

实验在 CN-Celeb 数据集上验证了所提方法的有效性,所提 TF-DCAM 模型等错误率(equal error rate,EER)和最小代价函数(minimum detection cost function,minDCF)上分别降低 14.98%和 10.98%,在提升识别性能的同时保持较低计算开销。此外,为进一步验证模型在跨语言、跨场景条件下的泛化能力与鲁棒性,本文还在英文公开测试集 VoxCeleb1 上进行了迁移评估实验,结果表明所提方法在不同语言仍保持良好性能,展现出较强的通用性与适应性。

1　模　　型

所提 TF-DCAM 模型如图 1 所示,由空洞卷积模块,时频多尺度注意力模块及主干网络 D-TDNN 组成,旨在提升语音嵌入的表达能力。在特征提取阶段,采用空洞卷积模块进行建模,结合时频重聚焦增强特征选择能力;在注意力融合阶段,引入时频多尺度注意力模块,自适应优化时间和频率维度的特征权重,提升模型在复杂环境下的鲁棒性;在主干网络阶段,构建 D-TDNN 进行深度时序建模,并通过统计池化和全连接层生成高判别力的说话人嵌入。

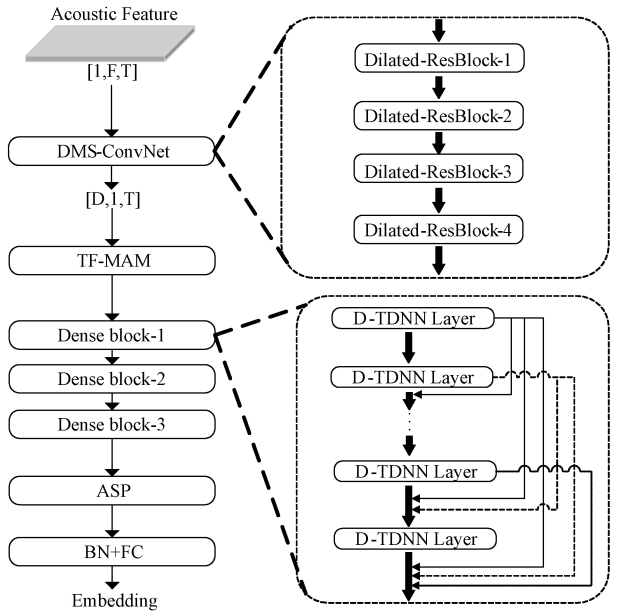


图 1　TF-DCAM 模型整体结构图

Fig. 1　Overall structure diagram of the TF-DCAM model

1.1　空洞卷积模块

在说话人确认任务中,语音信号往往包含大量冗余信息。传统卷积神经网络受限于固定感受野,难以兼顾局部与全局信息。为此,DMS 模块结合空洞残差卷积和时频自适应重聚焦单元进行优化。前者通过不同扩张率的空洞卷积扩展感受野,高效建模时频依赖;后者参考通道重组策

略,优化时频特征分布,减少冗余并提升特征表达能力。该模块为后续的深度时序建模提供高质量特征表示。本文使用的空洞残差卷积块如图 2 所示。

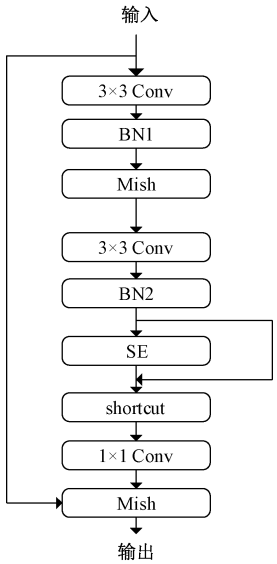


图 2 空洞卷积残差块

Fig. 2 Dilated convolution residual block

空洞卷积通过扩张因子(dilation rate)扩大感受野,在保持参数量稳定的同时提高建模能力。在 DMS 模块中,为增强多尺度建模能力,采用多个具有不同扩张率的空洞残差卷积(dilated-resblock)进行堆叠,其中单层计算公式如下:

$$\boldsymbol{X}_{L+1} = f(\boldsymbol{W}_L *_d \boldsymbol{X}_L + \boldsymbol{B}_L) \tag{1}$$

其中,$\boldsymbol{X}_{L+1}$ 和 $\boldsymbol{X}_L$ 分别表示第 L 层的输入和输出特征,$\boldsymbol{W}_L$ 为卷积核,$*_d$ 表示空洞卷积操作,$\boldsymbol{B}_L$ 为偏置项,$f(\cdot)$ 代表非线性激活函数。

为了增强特征表达能力,此处选择使用 Mish 作为激活函数,其平滑特性有助于梯度稳定传播,相较于 ReLU 能更好地保留负区间信息,以减少特征损失,提高 DMS 模块地时频建模能力,其表达式如下:

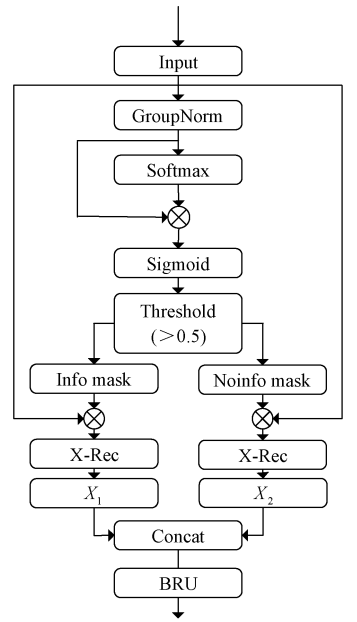
$$\mathrm{Mish}(x) = x \cdot \tanh(\ln(1+\mathrm{e}^x)) \tag{2}$$

DMS 模块采用不同扩张率的空洞卷积进行计算,并通过堆叠多个空洞残差块构建时频感受野金字塔结构,以捕获不同尺度的时频依赖信息:

$$d \in \{1,2,4,8\} \tag{3}$$

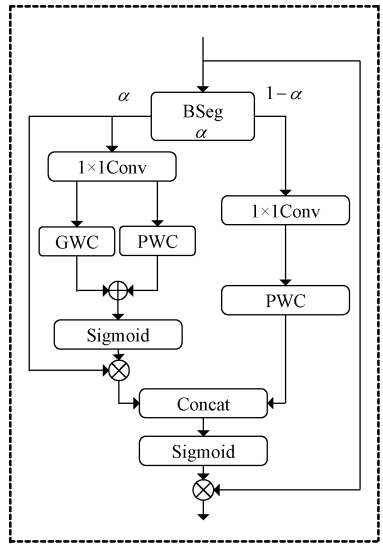
这种策略既能提取短时局特征,又能建模长期依赖,在适度提升计算复杂度的前提下提高模型在说话人确认任务中的鲁棒性。

传统卷积神经网络在处理语音数据时,通道信息往往存在冗余,影响特征表达的准确。借鉴计算机视觉领域的通道重组策略,DMS 模块引入时频自适应重聚焦单元(TF-ARU),以减少时频特征中的冗余信息,并增强关键特征的表达能力。TF-ARU 主要由时频动态门控单元(time-frequency gating unit,TFGU)和多尺度频带重建单元(band reconstructing unit,BRU)组成,分别负责动态调整信息流和频带特征的优化重构,有效地优化了时频信息的选择和组织方式,提高了语音嵌入的质量及声纹特征的区分能力,TF-ARU 结构如图 3 所示。



(a) 时频动态门控单元

(a) Time-frequency gating unit



(b) 多尺度频带重建单元

(b) Band reconstructing unit

图 3 TF-ARU 结构图

Fig. 3 Structure diagram of TF-ARU

TFGU 采用“分离-重建”策略,通过动态门控机制筛选输入特征,增强关键信息并抑制冗余,其主要流程如下:

对输入特征组归一化,使不同通道的特征分布更加稳定:

$$\boldsymbol{X}_{\mathrm{GN}}=\frac{\boldsymbol{X}-\mu}{\varepsilon} \tag{4}$$

其中，$\boldsymbol{X}$ 为输入特征；μ，ε 分别为均值和标准差。

归一化后的特征通过 softmax 和 sigmoid 计算权重，生成门控掩码：

$$\boldsymbol{M}=\sigma(\mathrm{Softmax}(\boldsymbol{X}_{\mathrm{GN}})) \tag{5}$$

其中，σ 表示 sigmoid 函数。

根据设定阈值（本文设定为 0.5），将特征划分为信息保留部分 $\boldsymbol{X}_{\mathrm{inf}}$ 和冗余部分 $\boldsymbol{X}_{\mathrm{noninf}}$：

$$\boldsymbol{X}_{\mathrm{inf}}=\boldsymbol{M}\odot\boldsymbol{X},\boldsymbol{X}_{\mathrm{noninf}}=(1-\boldsymbol{M})\odot\boldsymbol{X} \tag{6}$$

其中，$\odot$ 代表逐点乘法（element-wise multiplication）。

筛选后对信息保留部分和冗余部分进行交叉重构，对数据进行部分交换：

$$\boldsymbol{X}_1=\mathrm{Concat}(\boldsymbol{X}_{\mathrm{inf}}[:L/2,/:],\boldsymbol{X}_{\mathrm{noninf}}[L/x:L,:]) \tag{7}$$

$$\boldsymbol{X}_2=\mathrm{Concat}(\boldsymbol{X}_{\mathrm{noninf}}[:L/2,/:],\boldsymbol{X}_{\mathrm{inf}}[L/x:L,:]) \tag{8}$$

交换后的特征拼接得到 TFGU 的输出：

$$\boldsymbol{X}_{\mathrm{TFGU}}=\mathrm{Concat}(\boldsymbol{X}_1,\boldsymbol{X}_2) \tag{9}$$

经过 TFGU 处理后的特征 $\boldsymbol{X}_{\mathrm{TFGU}}$ 作为多尺度频带重建单元（BRU）的输入进行后续处理，如图 3(b)所示。

多尺度频带重建单元（BRU）通过带通分割对输入特征进行划分，以增强模型对关键语音特征的感知能力，其主要流程如下：

带通分割将输入特征 $\boldsymbol{X}_{\mathrm{TFGU}}$ 分为两条不同频带路径，以增强不同频带信息的建模能力：

$$\boldsymbol{X}_{\alpha}=\alpha\cdot\boldsymbol{X}_{\mathrm{TFGU}},\boldsymbol{X}_{1-\alpha}=(1-\alpha)\cdot\boldsymbol{X}_{\mathrm{TFGU}} \tag{10}$$

其中，α 是可学习参数（初始化为 0.5），$\boldsymbol{X}_{\alpha}$ 侧重高能量频带特征，$\boldsymbol{X}_{1-\alpha}$ 侧重低能量频带信息。

高能量频带路径上，通过 1×1 卷积进行线性映射，结合分组卷积（GWC）和逐点卷积（PWC）分别进行特征分组处理和全局通道建模，再将两部分特征融合：

$$\boldsymbol{X}_{\mathrm{high}}=\boldsymbol{X}_{\alpha}\odot\sigma(\mathrm{GWC}(\boldsymbol{X}_{\alpha})+\mathrm{PWC}(\boldsymbol{X}_{\alpha})) \tag{11}$$

低能量频带路径仅使用 1×1 卷积和逐点卷积进行优化：

$$\boldsymbol{X}_{\mathrm{low}}=\mathrm{PWC}(\boldsymbol{X}_{1-\alpha}) \tag{12}$$

将两路径输出进行拼接，通过 sigmoid 计算最终权重，对输入特征进行重加权，得到最终输出：

$$\boldsymbol{X}_{\mathrm{out}}=\boldsymbol{X}_{\mathrm{TFGU}}\odot\sigma(\mathrm{Concat}(\boldsymbol{X}_{\mathrm{high}},\boldsymbol{X}_{\mathrm{low}})) \tag{13}$$

该模块在提升模型性能的同时控制计算开销，生成优化后的特征表示，输入至 TF-MAM 模块进行进一步处理。

1.2　时频多尺度注意力模块

时频多尺度注意力模块（temporal-frequency multi-scale attention module，TF-MAM）是 TF-DCAM 说话人确认模型中的核心组件之一，旨在优化时频特征表达，提高模型的鲁棒性和准确性。该模块结合时间注意力机制、通道注意力机制和跨纬度交互（cross-dimensional interaction，CDI）机制，在时域和频域上动态调整特征权重，以实现更高效的时频特征融合和信息表达。

TF-MAM 采用双支路结构，如图 4 所示。左支路侧重全局信息，右支侧重局部信息提取，最终输出将两条支路特征通过加权融合形成综合的特征表示，为说话人确认任务提供更加精确的语音特征。

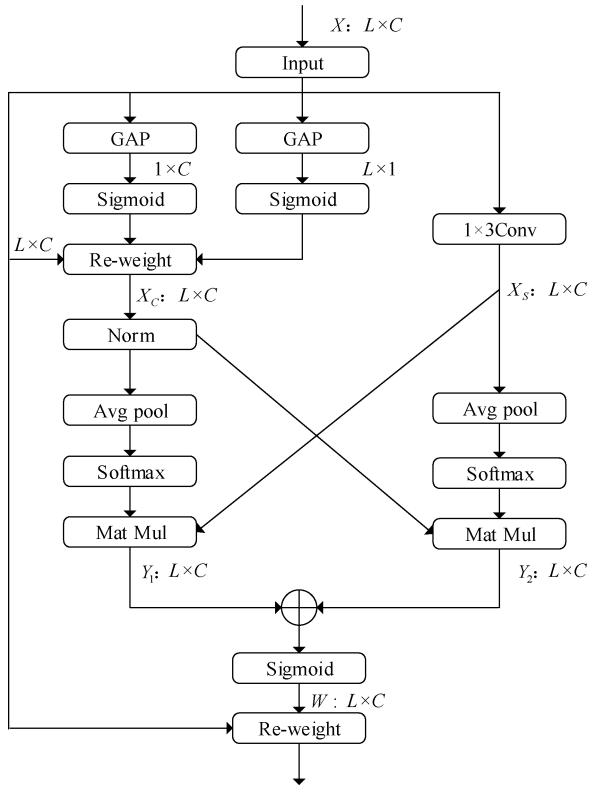


图 4　TF-MAM 结构图

Fig. 4　Structure diagram of TF-MAM

左支路通过全局平均池化（global average pooling，GAP）提取时间和频率维度的全局特征，并通过 sigmoid 计算通道注意力权重，得到每个通道的权重系数 $\boldsymbol{A}_{\mathrm{C}}$，计算过程如下：

$$\boldsymbol{A}_{\mathrm{C}}=\sigma(\boldsymbol{W}_1(\frac{1}{L}\sum_{i=1}^{L}\boldsymbol{X}_i)) \tag{14}$$

其中，$\boldsymbol{W}_1$ 是一个可训练的权重矩阵。

计算所得权重用于重加权输入特征，得到增强后的全局信息特征 $\boldsymbol{X}_{\mathrm{C}}$：

$$\boldsymbol{X}_{\mathrm{C}}=\boldsymbol{A}_{\mathrm{C}}\odot\boldsymbol{X} \tag{15}$$

右支路通过在时间维度上执行 1×3 卷积聚合局部时间窗口信息，得到局部表示 $\boldsymbol{X}_{\mathrm{S}}$：

$$\boldsymbol{X}_{\mathrm{S}}=\mathrm{Conv}_{1\times3}(\boldsymbol{X}) \tag{16}$$

为了获得更丰富的特征表达，提出一种不同时间感受野的跨时间聚合算法，对获得的全局表示和局部表示进行信息编码，左支路的全局信息与右支路的局部信息分别通过 Softmax 归一化生成通道描述符，并交叉作用，以实现多尺度信息融合：

$$\boldsymbol{A}_1=\mathrm{Softmax}(\mathrm{GAP}(\boldsymbol{X}_{\mathrm{C}})) \tag{17}$$

$$\boldsymbol{A}_2 = \mathrm{Softmax}(\mathrm{GAP}(\boldsymbol{X}_S)) \tag{18}$$

$$\boldsymbol{X}_{att} = \boldsymbol{A}_1 \odot \boldsymbol{X}_S + \boldsymbol{A}_2 \odot \boldsymbol{X}_C \tag{19}$$

通过 Sigmoid 门控函数计算最终加权系数,并对输入特征进行校准:

$$\boldsymbol{X}_{TF-MAM} = \sigma(\boldsymbol{X}_{att}) \odot \boldsymbol{X} \tag{20}$$

该模块在全局和局部特征之间建立动态联系,使模型能够更有效地捕捉语音信号的关键时频信息,从而提升说话人确认任务的表现。

1.3 深度时序卷积网络

深度时序卷积网络(D-TDNN)作为 TF-DCAM 的主干部分,采用多层膨胀卷积以建模语音的时序结构,相比传统 TDNN,D-TDNN 通过扩展感受野增强了对长时依赖的建模能力,在保证计算效率的前提下更有效地捕捉语音序列的动态变化。为增强对时序信息的调控能力同时提升通道信息的自适应性,本文在单层卷积模块中采用了改进的自适应掩码时序卷积(adaptive masking temporal convolution,AMTC),该模块融合多尺度时序上下文,通过动态加权机制增强关键时间区域的表示能力,提高语音嵌入的判别性和鲁棒性,其结构如图 5 所示。

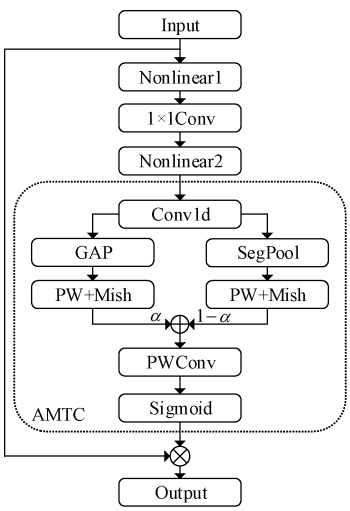


图 5 单一时序卷积层结构

Fig. 5 Single temporal convolutional layer structure

单一时序卷积层过程如下:

对输入特征 $\boldsymbol{X}$ 先经过两次非线性变换与 1×1 卷积,以增强特征的非线性表达能力并调整通道维度:

$$\boldsymbol{X}_{NL2} = F_2(\mathrm{Conv}_{1\times1}(F_1(\boldsymbol{X}))) \tag{21}$$

其中,$F_1(\cdot)$ 和 $F_2(\cdot)$ 分别表示两次非线性变换。

随后,通过膨胀卷积提取长时间依赖特征,接着引入多尺度池化策略,分别通过全局平均池化和分段池化提取不同时间尺度的信息,随后分别经过降维操作降低复杂度和 Mish 激活函数进行增强:

$$\boldsymbol{X}_T = \mathrm{Conv1d}(\boldsymbol{X}_{NL2}) \tag{22}$$

$$\boldsymbol{A}_G = \mathrm{Mish}(\mathrm{PWC}(\mathrm{GAP}(\boldsymbol{X}_T))) \tag{23}$$

$$\boldsymbol{A}_S = \mathrm{Mish}(\mathrm{PWC}(\mathrm{SegPool}(\boldsymbol{X}_T))) \tag{24}$$

其中,GAP 提取全局时间尺度的信息,SegPool 通过对时间轴划分不同区段并计算平均值,更关注局部时间特征。

为自适应融合不同尺度的信息,引入可学习自适应系数 α(初始值为 0.5)进行权重调节:

$$\boldsymbol{A}_{fused} = \alpha\boldsymbol{A}_G + (1-\alpha)\boldsymbol{A}_S \tag{25}$$

融合后的注意力分数先升维还原,再通过 Sigmoid 激活,生成上下文掩码权重:

$$\boldsymbol{A}_{AMTC} = \sigma(\mathrm{PWC}(\boldsymbol{A}_{fused})) \tag{26}$$

将该掩码作用于输入特征,完成加权:

$$\boldsymbol{X}_{OUT} = \boldsymbol{A}_{AMTC} \odot \boldsymbol{X} \tag{27}$$

多个上述卷积层堆叠构成 D-TDNN,每层通过膨胀卷积学习不同粒度的时间信息,逐层提取更高层次的特征表示。假设网络的第 L 层卷积层的输出为 $\boldsymbol{X}_L$,那么该层卷积操作可简化表示为:

$$\boldsymbol{X}_L = \mathrm{Conv}_L(\boldsymbol{X}_{L-1}) \tag{28}$$

为了增强特征流动和信息复用,D-TDNN 采用密集连接(dense connection)机制,将每一层的输出会与之前层的信息进行拼接:

$$\boldsymbol{X}_L = \mathrm{Concat}(\boldsymbol{X}_{L-1}, \mathrm{Conv}_L(\boldsymbol{X}_{L-1})) \tag{29}$$

密集连接不仅保留了低层细节特征,也提升了梯度传播效率,增强训练稳定性。在经过一个密集连接块后,网络会接入过渡层(transit layer)进行通道数调整,确保不同特征通道的适应性,同时还会用于批归一化和非线性激活函数,增强模型的稳定性和非线性表达能力:

$$\boldsymbol{X}_{BN} = \frac{\boldsymbol{X}_L - \boldsymbol{\mu}}{\sqrt{\sigma^2 + \varepsilon}} \cdot \gamma + \boldsymbol{\beta} \tag{30}$$

其中,$\boldsymbol{\mu}$ 和 σ^2 分别表示批次特征的均值和方差;γ 和 $\boldsymbol{\beta}$ 为可训练的缩放和偏移参数,ε 为防止除零错误的小常数。

D-TDNN 结构通过层层堆叠与特征融合,有效建模多层次的时间依赖信息,为最终嵌入表示提供丰富的时序基础。

1.4 对比损失函数

在说话人确认任务中,损失函数的选择直接决定了嵌入空间的分布形式,从而影响模型的区分能力和泛化能力。原模型使用的损失函数为 AAMLoss(additive angular margin loss),是一种改进型的 Softmax 损失函数,旨在通过在角度空间引入附加间隔,提高类间可分性和类内聚合性,其损失形式如下:

$$L_{AAM} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{S(\cos(\theta_{Y_i})-m)}}{\mathrm{e}^{S(\cos(\theta_{Y_i})-m)} + \sum_{j\neq y_i}\mathrm{e}^{S\cos(\theta_j)}} \tag{31}$$

其中,θ 为缩放因子,用于扩大角度分布;m 为角度间隔,用于提高不同类之间的分离度;θ_{y_i} 表示第 i 个样本与其真实类别中心之间的角度,N 为批次大小。

虽然 AAMLoss 在增强说话人类间间隔、提高判别能力方面表现优越,但仍然存在不足之处,首先 AAMLoss 主要依赖单个类别中心的角度信息进行优化,局限于单点决

策，仅考虑类别中心和样本之间的距离，未充分利用样本间的信息结构；其次，AAMLoss 不能直接优化同类别内部的样本结构，容易出现类内分布不均匀的问题。基于上述 AAMLoss 的局限性，本文引入两种对比损失函数，用于进一步优化嵌入空间[19-20]。

1）SupConLoss

SupConLoss（supervised contrastive loss）引入标签监督，鼓励同类样本在嵌入空间中聚集，增强类内一致性，其形式如下：

$$L_{SC}=\sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i\cdot z_a/\tau)} \tag{32}$$

其中，$P(i)$ 表示与样本 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，用于调整样本间的对比度，通常设为 0.07。

相比于 AAMLoss 仅考虑类别中心，SupConLoss 还强调了同一类别样本之间的聚合度，优化了嵌入空间的类内分布，提高对边界样本的适应能力。

2）NT-XentLoss

NT-XentLoss（normalized temperature-scaled cross entropy loss）是一种无监督对比学习损失，在本文中用于增强类间分离，其表达形式为：

$$L_{NT}=-\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(z_i,z_j)/\eta)}{\sum_{k=1}^{2N}1_{[k\neq i]}\exp(\mathrm{sim}(z_i,z_k)/\eta)} \tag{33}$$

其中，$\mathrm{sim}(z_i,z_j)$ 代表样本 i 和 j 之间的相似度，一般用余弦相似度计算；η 为温度参数，通常设为 0.5；$1_{[k\neq i]}$ 指排除自身样本的负样本对。

NT-XentLoss 通过拉远不同类别样本间距离，以优化类间边界，使不同类别的样本更加分离；且由于它计算的是样本两两之间的相似度，因此不会受到单一类别重心偏移的影响，能够更加灵活地优化嵌入空间结构。

3）多损失函数融合

本文最终采用 AAMLoss 作为主损失函数，引入 SupConLoss 和 NT-XentLoss 作为对比损失函数进行联合训练，总损失函数表达式如下：

$$L=L_{AAM}+\lambda_1 L_{SC}+\lambda_2 L_{NT} \tag{34}$$

其中，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 $\lambda_1+\lambda_2$ 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。

2　实验结果与性能分析

2.1　数据集

在本次实验中，为验证所提 TF-DCAM 说话人确认模型的有效性，选用公开数据集 CN-Celeb 作为实验数据来源。CN-Celeb 数据集是目前最大的中文开源说话人识别数据集之一，由 CN-Celeb1 和 CN-Celeb2 两部分组成，其中 CN-Celeb1 包含约 1 000 名说话人的语音样本，CN-Celeb2 为其扩展版本，新增约 2 000 名说话人，总计约 3 000 名说话人、650 000 条语音，总时长超过 800 h。所有音频均为单通道录音，采样率为 16 kHz，位深为 16 bit，涵盖了多种真实环境下的语音场景，具有良好的多样性和泛化能力。实验中将 CN-Celeb1 与 CN-Celeb2 整体作为训练集，测试集使用 CN-Celeb 官方提供的预设测试集，包含 200 名说话人、约 18 000 组验证对，用于评估模型在中文复杂环境中的说话人识别性能。

此外，为进一步验证所提模型的跨语种泛化能力与鲁棒性，本文在英文公开数据集 VoxCeleb1 上进行额外测试。VoxCeleb1 是由牛津大学 Deep Learning 组发布的国际主流说话人识别数据集，包含 1 251 名不同说话人、超过 15 万段语音片段，全部来源于 YouTube 视频，涵盖多种语言、口音和背景噪声环境，具有高度真实场景复杂性。本文采用 VoxCeleb1 标准测试划分，作为跨语种验证集，以评估 TF-DCAM 在非中文语料下的表现，进一步验证其语种适应性和实际应用潜力。

为提升模型在真实环境中的鲁棒性和泛化能力，在训练过程中采用以下数据预处理和增强策略：

1）音量归一化：将音频统一归一化至 −20 dB，降低录音设备差异对模型训练的干扰；

2）音频裁剪：过滤小于 0.3 s 的短音频，截断超过 3 s 的长音频，保证训练语音片段的稳定性；

3）批量采样策略：默认采用随机采样方式提升样本多样性，同时支持 PK-Sampler，确保说话人类别均衡，有利于训练收敛和泛化性能提升。

2.2　实验设置

本文所用实验环境：处理器为 Intel Core i5-12600KF（3.7 GHz），GPU 为 GeForce RTX 4060Ti，操作系统为 Windows11，Pytorch 版本为 2.3.0，Python 版本为 3.10.14，CUDA 版本为 11.8。

在本次实验中，对于模型的设置如下：音频采样率固定为 16 kHz，batch_size 设置为 64，优化方法采用 Adam，初始学习率设为 0.001，学习率衰减由 WarmupCosineSchedulerLR 进行管理，预热周期设置为 5 轮，最大学习率设为 0.001，最小学习率为 1×10^{-5}，对比损失权重设定为 0.01，实验采用 60 轮训练，保存效果最佳的模型。

2.3　评价指标

在说话人确认任务中，主要采用等错误率（equal error rate，EER）和最小检测代价函数（minimum detection cost function，minDCF）作为核心性能评价指标，其中 EER 表示假接受率和假拒率相等时的错误率，取值范围为[0，1]，其数值越低，表示模型区分不同说话人的能力越强；minDCF

通过引入错误检测代价评估模型在不同应用场景下的决策表现,数值越低代表模型在复杂环境下具有更强的鲁棒性。此外,为更全面地衡量模型性能,本文还引入了以下 3 项补充指标:参数量(Params)反映模型结构的复杂度;浮点运算量(FLOPs)衡量模型在推理阶段的计算成本;实时因子(real-time factor,RTF)衡量模型在实际部署中的推理速度,越低代表实时性越好。

2.4 对比损失函数权重优化实验

为进一步优化嵌入空间结构,提升说话人类内聚合度与类间区分度,本文在主损失函数 AAMLoss 的基础上,引入 SupConLoss 和 NT-XentLoss 两种对比损失函数进行联合训练。由于对比损失的权重对模型收敛性和最终性能影响显著,本文设计了一组权重组合实验,评估其对识别性能(EER 和 minDCF)的影响。

实验结果如表 1 所示,可以观察到:1)当总权重设为 0.005(两种损失各 0.002 5)时,模型性能相较基线明显提升,表明对比损失有助于优化特征空间结构;2)将总权重进一步提升至 0.01 后,EER 降至 14.72%,minDCF 降至 0.672 9,达到最佳性能,说明适当增加对比信号可增强判别能力;3)当权重增至 0.015 后,性能略有下降,表明过强的对比约束可能破坏主损失的优化目标,导致嵌入结构失衡;4)当权重达到 0.02 时,训练出现梯度爆炸现象,推测是因对比损失占比过高削弱了 AAMLoss 对类别边界的有效约束,致使训练过程失控。综上考虑,本文最终选择 SupConLoss 与 NT-XentLoss 权重各设为 0.005,在保证最优性能的同时,兼顾训练稳定性与收敛效率。

表 1 在 CN-Celeb 数据集上的对比损失函数权重优化实验

Table 1 Experiment on the optimization of the weight of the contrastive loss function on the CN-Celeb dataset

模型	SupConLoss 权重	NT-XentLoss 权重	对比损失总权重	EER/%	MinDCF
CAM++	0	0	0	15.55	0.715 8
CAM++	0.002 5	0.002 5	0.005	14.91	0.690 3
CAM++	0.005 0	0.005 0	0.010	14.72	0.672 9
CAM++	0.007 5	0.007 5	0.015	15.03	0.688 5
CAM++	0.010 0	0.010 0	0.020	—	—
TF-DCAM	0	0	0	13.88	0.659 0
TF-DCAM	0.002 5	0.002 5	0.005	13.46	0.640 9
TF-DCAM	0.005 0	0.005 0	0.010	13.22	0.637 2
TF-DCAM	0.007 5	0.007 5	0.015	13.65	0.648 1

2.5 CN-Celeb 数据集实验对比

为全面评估所提 TF-DCAM 模型的性能,本文在 CN-Celeb 数据集上与近年来具有代表性的说话人确认模型进行了对比实验,选取的对比模型包括 ResNet34、ERes2Net、ECAPA-TDNN 和 CAM++,在相同的预处理流程和训练设置下,使用等错误率(EER)与最小检测代价函数(minDCF)作为主要评价指标。

实验结果如表 2 所示,从实验结果可以看出,TF-DCAM 在 EER 和 minDCF 两项指标上均优于其他模型,体现出更强的判别能力与鲁棒性。具体分析如下:1)与 ResNet34 和 ERes2Net 相比,TF-DCAM 通过引入空洞卷积与自适应掩码时序卷积,提升了长时依赖建模能力,使嵌入特征更加紧凑,EER 和 minDCF 显著下降,在保持计算效率的同时有效提升了识别性能。2)与 ECAPA-TDNN 和 CAM++相比,尽管三者均基于 TDNN 框架构建,TF-DCAM 进一步引入多尺度注意力机制与对比损失联合优化,EER 分别下降 3.83% 和 2.09%,minDCF 分别降低 0.118 1 和 0.078 6,验证了本模型在保持轻量化的同时具备更强的特征表达能力和泛化性能。

表 2 各模型在 CN-Celeb 数据集的对比实验

Table 2 Comparative experiments of each model on the CN-Celeb dataset

模型	Params/M	EER/%	minDCF
ResNet34	6.7	15.99	0.721 2
ERes2Net	6.6	15.18	0.669 4
ECAPA-TDNN	14.7	17.29	0.755 3
CAM++	7.2	15.55	0.715 8
TF-DCAM	**7.6**	**13.22**	**0.637 2**

2.6 VoxCeleb1 验证集实验对比

为进一步验证所提模型在跨语种与复杂环境条件下的泛化能力与鲁棒性,本文在 VoxCeleb1 测试集上进行了跨数据集验证实验。所有模型均仅使用中文语料 CN-Celeb 数据集进行训练,不引入 VoxCeleb1 中的任何训练样本,以保证测试结果的客观性与泛化评估的真实性。测试阶段采用 VoxCeleb1 官方划分的验证集,并沿用 CN-Celeb 实验中相同的评估指标——等错误率(equal error rate,EER)和最小检测代价函数(minimum detection cost

function，minDCF）。实验选取与 2.5 节相同的代表性模型进行对比，包括 ResNet34、ECAPA-TDNN、CAM＋＋和本文所提 TF-DCAM 模型。所有模型在相同训练设置与推理流程下进行测试，验证阶段统一使用余弦相似度打分和固定阈值判决机制。

实验结果如表 3 所示，从表中结果可以看出，TF-DCAM 在 VoxCeleb1 上取得了显著优于其他模型的性能表现 EER 为 14.55%，minDCF 为 0.891 3，在所有对比模型中均为最低，展现出优异的跨语种泛化能力。具体分析如下：1）对比轻量 CNN 架构（ResNet34、ERes2Net）：两者在跨语言场景中表现受限，EER 分别为 16.95% 和 16.85%，minDCF 超过 0.90，说明浅层特征提取器在语种转移中难以捕捉稳定的说话人特征。相比之下，TF-DCAM 在 EER 上降低 2.4%和 2.3%，表明本文模型通过引入空洞卷积与深度特征融合，有效增强了对发音风格变化与背景噪声的适应能力。2）对比主流 TDNN 架构（ECAPA-TDNN、CAM＋＋）：虽然这两种方法具备一定的上下文建模能力，但由于未引入多尺度调控和深度重聚焦机制，在语种和语境发生变化时，模型的特征表征稳定性下降。TF-DCAM 相比 ECAPA-TDNN，EER 降低 1.86%，minDCF 降低 0.051 2，表现出更强的抗干扰与判别能力。简而言之，TF-DCAM 在保持轻量化的前提下，展现出显著优于其他模型的跨语言鲁棒性与特征判别能力。

表 3　各模型在 VoxCeleb1 验证集上的对比实验

Table 3　Comparative experiments of each model on the VoxCeleb1 validation set

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	16.36	0.930 6
TF-DCAM	**7.6**	**14.55**	**0.891 3**

2.7　消融实验

为验证各个改进模块在 TF-DCAM 模型中的有效性，本文设计了消融实验，分别在 Baseline 模型上逐步引入关键组件，包括：空洞卷积模块（DMS）、时频多尺度注意力模块（TF-MAM）、自适应掩码时序卷积（AMTC）以及对比损失函数，实验在 CN-Celeb 数据集上进行，评估指标为 EER 和 minDCF。

实验结果如表 4 所示，结果表明：1）在原始模型中引入 DMS 模块后，EER 和 minDCF 降至 14.91%和 0.689 2，说明空洞卷积有助于提升特征提取的全局感受野，增强时频依赖建模能力；2）在此基础上加入 TF-MAM 模块，EER 进一步下降至 14.25%，表明该模块能有效融合时频特征，提升模型对多维信息的表达能力；3）继续引入 AMTC 模块后，EER 和 minDCF 降至 13.88%和 0.659 0，表明该模块优化了时序建模过程，增强了特征的上下文鲁棒性；4）最终加入对比损失函数，模型性能进一步提升，EER 降至 13.22%，minDCF 降至 0.637 2，验证了多任务监督优化对嵌入空间判别性的显著促进作用。综上，消融实验证明了本文提出的各个模块均能有效提升模型性能，且组合使用时具备良好的协同增益效果。

表 4　在 CN-Celeb 数据集上的消融实验

Table 4　Ablation experiments on the CN-Celeb dataset

模型	空洞卷积模块（DMS）	时频多尺度注意力（TF-MAM）	自适应掩码时序卷积（AMTC）	对比损失函数	EER/%	minDCF
Baseline	×	×	×	×	15.55	0.715 8
＋DMS	√	×	×	×	14.91	0.689 2
＋TF-MAM	√	√	×	×	14.25	0.663 8
＋AMTC	√	√	√	×	13.88	0.659 0
＋对比损失	√	√	√	√	13.22	0.637 2

2.8　学习率敏感性分析

在深度神经网络模型训练过程中，初始学习率的设置对模型的收敛速度与最终性能具有关键影响。尽管文献中普遍建议将初始学习率设定在 0.1～0.001 之间，且 0.01 被广泛认为是兼顾收敛速度与稳定性的常用值，但在本研究所涉及的任务与网络结构中，该设置未能取得预期效果。具体而言，当初始学习率设定为 0.01 时，模型在训练初期出现较大的梯度波动，导致损失函数异常增长，甚至引发数值不稳定，进而触发程序的自动中止机制。为此，本文将初始学习率下调至 0.001，以增强训练过程的数值稳定性。该设置不仅有效缓解了训练初期的震荡问题，避免了程序非正常终止的现象，也为模型提供了更平稳的优化路径。实验结果表明，在初始学习率为 0.001 的条件下，模型能够顺利完成训练，并在验证指标上取得更优表现，验证了该设定的合理性与实用性。

同时为了选择最佳的最终学习率，对最终学习率进行敏感性训练分析。根据已有工作中对学习率设置的范围，通常学习率被控制在 1×10^{-4}～2×10^{-3} 之间，以实现较平

滑的收敛曲线和稳定的优化路径。使用本文模型在 CN-Celeb 数据集上,参考上述范围,设计五组不同的最终学习率方案,为 0.000 1、0.000 5、0.001、0.001 5、0.002,在保持其他超参数不变的条件下,每组均进行 60 轮训练,并以验证集上的 EER 作为性能评估指标。

实验结果如图 6 所示,不同学习率设置下模型性能呈现出一定的差异性,通过验证结果可知,随着学习轮数增加,不同学习率下的模型验证指标 EER 都会向着一个极小值靠近,当选择最终学习率为 0.001 时,验证指标 EER 达到最小,说明本文算法选择最终学习率为 0.001 时能过够生产性能最佳的模型。因此选择初始学习率为 0.001,最终学习率为 0.001 作为实验训练指标。

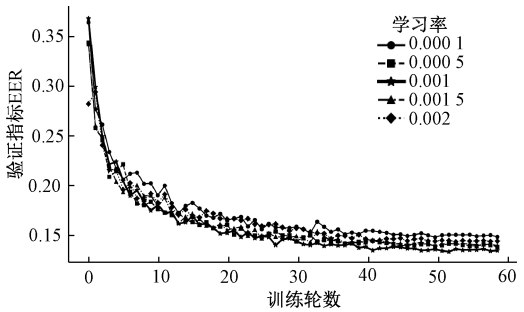


图 6 不同最终学习率下模型性能变化

Fig. 6 The variation of model performance under different final learning rates

2.9 复杂性分析

为评估 TF-DCAM 模型在计算资源上的开销,本文将其与 ECAPA-TDNN、ResNet34 及 CAM++三种主流模型进行复杂度对比,评估指标包括参数量(Params)、浮点运算次数(FLOPs)和实时因子(RTF)。所有模型均在单线程 CPU 环境下评估推理速度,实验结果如表 5 所示。结果表明:1)相比 ECAPA-TDNN,TF-DCAM 参数量减少约 48.1%,FLOPs 降低 53.3%,RTF 降至 0.014,推理速度提升超过 57%,在显著降低计算复杂度的同时保持优秀的识别性能;2)相比 ResNet34,TF-DCAM 的 FLOPs 仅为其约 27.1%,RTF 略低,显示出更高的计算效率,而 ResNet34 受较高参数依赖影响,可能存在较大的内存访问开销,导致推理速度受限;3)相比 CAM++,TF-DCAM 参数略有增加,FLOPs 增加轻微,但仍远低于其他两种主流模型,表明在保持轻量级特性的同时,进一步提升了建模能力和性能表现。综合来看,TF-DCAM 在模型规模、计算量和实时性之间实现了良好平衡,更适用于资源有限的实际应用场景。

表 5 各模型复杂度对比实验

Table 5 Experiment on the comparison of the complexity of each model

模型	Params/M	FLOPs/G	RTF
ECAPA-TDNN	14.7	3.96	0.033
ResNet34	6.7	6.84	0.032
CAM++	7.2	1.72	0.013
TF-DCAM	7.6	1.85	0.014

3 结 论

本文提出了一种基于时频多尺度建模与特征重组机制的说话人确认模型 TF-DCAM,针对复杂语音环境下语音特征冗余、建模能力不足等问题进行了系统优化。首先,空洞卷积模块(DMS)利用多尺度空洞残差结构与时频重聚焦机制,有效增强了特征建模能力;其次,时频多尺度注意力模块(TF-MAM)结合通道注意力与跨维度交互,提升了模型对关键信息的感知能力;再次,自适应掩码时序卷积模块(AMTC)通过多尺度池化与上下文门控机制优化了时序特征表达;最后,引入 SupConLoss 与 NT-XentLoss 对比损失,实现嵌入空间的结构优化。在 CN-Celeb 数据集和 VoxCeleb1 验证集上的实验验证表明,所提 TF-DCAM 模型在 EER 和 minDCF 两项指标上均优于多种主流对比模型,在提升识别性能的同时保持了较低的计算开销,展示出良好的跨语言鲁棒性与实际部署潜力。未来工作将进一步探索更高效的时频多尺度融合机制,并在更具挑战性的跨语言、跨信道和真实应用场景中,持续验证和提升模型的泛化能力与应用价值。

参考文献

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

[2] ZHOU T Y, ZHAO Y, LI J Y, et al. CNN with phonetic attention for text-independent speaker verification [C]. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019: 718-725.

[3] 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). IEEE, 2022: 7517-7521.

[4] LIU Y, SONG Y, MCLOUGHLIN I, et al. An effective deep embedding learning method based on dense-residual networks for speaker verification[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6683-6687.

[5] 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).

IEEE，2018：5329-5333.

[6] ZHENG S Q，LIU G，SUO H B，et al. Autoencoder-based semi-supervised curriculum learning for out-of-domain speaker verification[J]. System，2019，3：98.

[7] 毛文青，管业鹏. 基于 LPBMFCC 的文本无关说话人识别[J]. 电子测量技术，2020，43(19)：169-176.
MAO W Q，GUAN Y P. Text-independent speaker recognition based on LPBMFCC [J]. Electronic Measurement Technology，2020，43(19)：169-176.

[8] CHEN ZH Y，CHEN S Y，WU Y，et al. Large-scale self-supervised speech representation learning for automatic speaker verification[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2022：6147-6151.

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

[10] 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.

[11] ZHOU T Y，ZHAO Y，WU J. Resnext and res2net structures for speaker verification [C]. 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE，2021：301-307.

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

[13] YU Y Q，ZHENG S Q，SUO H B，et al. CAM：Context-aware masking for robust speaker verification[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2021：6703-6707.

[14] 李沐原，张兰春，张博源. 基于多尺度空洞融合注意力的车道线检测算法[J]. 电子测量技术，2024，47(23)：84-92.
LI M Y，ZHANG L CH，ZHANG B Y. Lane detection method based on multi-scale dilated fusion attention[J]. Electronic Measurement Technology，2024，47(23)：84-92.

[15] 郑展恒，曾庆宁，韦照川. 基于 DSP/BIOS 的语音信号处理系统设计[J]. 桂林电子科技大学学报，2015，35(6)：454-458.
ZHENG ZH H，ZENG Q N，WEI ZH CH. Design of speech signal processing system based on DSP/BIOS[J]. Journal of Guilin University of Electronic Technology，2015，35(6)：454-458.

[16] 李苗苗，华才健，谢涛，等. 融合多尺度特征及注意力机制的食品图像识别[J]. 电子测量技术，2024，47(18)：164-171.
LI M M，HUA C J，XIE T，et al. Integrating multi-scale features and attention mechanisms for food image recognition[J]. Electronic Measurement Technology，2024，47(18)：164-171.

[17] 贾林锋，吴黎明，温腾腾，等. 多尺度卷积的时频域语音分离方法研究[J]. 电子测量与仪器学报，2022，36(11)：134-140.
JIA L F，WU L M，WEN T T，et al. Speech separation in time-and-frequency domain based on multi-scale convolution [J]. Journal of Electronic Measurement and Instrumentation，2022，36(11)：134-140.

[18] 刘望生，刘艳梅. 多特征优化下室内声源鲁棒跟踪算法[J]. 仪器仪表学报，2024，45(8)：316-325.
LIU W SH，LIU Y M. Robust tracking algorithm for indoor sound source based on Multi-feature optimization[J]. Chinese Journal of Scientific Instrument，2024，45(8)：316-325.

[19] 关健，王敏. 基于深度神经网络和多元损失的说话人识别[J]. 电子测量技术，2019，42(5)：39-43.
GUAN J，WANG M. Speaker verification based on deep learning and beyond triplet loss[J]. Electronic Measurement Technology，2019，42(5)：39-43.

[20] 杨亚萍，张敬源. 整合类内差异与类间关联的隐喻感情预测[J]. 电子测量技术，2024，47(14)：108-120.
YANG Y P，ZHANG J Y. Metaphorical affective prediction integrating intra-class differences and interclass associations [J]. Electronic Measurement Technology，2024，47(14)：108-120.

作者简介

李嘉麒，硕士研究生，主要研究方向为说话人确认。
E-mail：1419371311@qq.com

郑展恒(通信作者)，硕士，高级实验师，硕士生导师，主要研究方向为信号处理。
E-mail：glzzh@guet.edu.cn

曾庆宁，博士，教授，主要研究方向为语音信号处理。
E-mail：qingningzeng@126.com

王健，硕士，副教授，硕士生导师，主要研究方向为智能信号处理。
E-mail：wangjian@guet.edu.cn