胡鸿淋, 熊淑华. 基于层级注意力机制的 Res2Net 说话人确认算法[J]. 智能计算机与应用,2025,15(6):190-195. DOI:10.20169/j.issn.2095-2163.25010806

基于层级注意力机制的 Res2Net 说话人确认算法

胡鸿淋, 熊淑华
(四川大学 电子信息学院, 成都 610065)

摘 要: 针对说话人确认任务中网络难以有效利用全局信息的问题,本文提出基于层级注意力机制的 Res2Net 说话人确认算法,通过融合多分辨率的层级输出结果,并依次经过通道注意力机制和空间注意力机制,确保可以有效提取出全局信息。此外,根据说话人确认任务设计了与传统注意力机制不同的局部特征融合算法,提取出更细节的局部特征并有效保留上下文信息。实验结果表明,本文算法比基线系统在等错误率(*EER*)和最小检测代价函数(*minDCF*)上分别提高了 41.7%和 29.7%,与 Res2Net 的其他变体 Res2Net-26w8s 和 ECAPA-TDNN 对比,等错误率分别提高了 39.3%和 12.9%,最小检测代价函数分别提高了 27.9%和 16.5%,由此可见本文算法在说话人确认的任务上有更好的性能。

关键词: 说话人确认; 深度学习; 注意力机制; 多分辨率
中图分类号: TP18 **文献标志码**: A **文章编号**: 2095-2163(2025)06-0190-06

Res2Net based on layer attention mechanism for speaker verification

HU Honglin, XIONG Shuhua
(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Aiming at the problem that it is difficult for the network to effectively use the global information in the speaker verification task, this paper proposes Res2Net speaker verification algorithm based on hierarchical attention mechanism, which can effectively extract the global information by fusing the multi-resolution hierarchical output results, and passing through the channel attention mechanism and spatial attention mechanism in turn. In addition, according to the speaker verification task, a local feature fusion algorithm which is different from the traditional attention mechanism is designed to extract more detailed local features and effectively retain the context information. Experimental results show that compared with the baseline system, the proposed algorithm improves the Equal Error Rate (*EER*) and minimum Detection Cost Function (*minDCF*) by 41.7% and 29.7%, respectively. Compared with other variants of Res2Net, Res2Net-26w8s and ECAPA-TDNN, the proposed algorithm improves the Equal Error Rate by 39.3% and 12.9%, and the minimum Detection Cost Function by 27.9% and 16.5%, respectively. It can be demonstrated that the proposed algorithm has better performance in the task of speaker verification.

Key words: speaker verification; deep learning; attention mechanism; multi-resolution

0 引 言

说话人确认在近年来已经成为一种新型的生物识别技术,旨在通过分析个体的声音来确认说话人的身份。目前,说话人确认也进入高速发展阶段,这项技术也已经开始应用于金融服务、智能设备、安全监控等领域,因此说话人确认技术开始得到更多人的重视。在传统说话人确认研究领域,高斯混合-通用背景模型[1-3](GMM-UBM)的提出,被视为是说话人识别领域的重要奠基算法,是通过引入概率模型来描述说话人的独特特征,也引领了后续联合因子分析[4-7](JFA)、i-vector[8-11]、高斯混合-支持向量机[12-15](GMM-SVM)等算法的提出。

得益于深度学习的兴起,其强大的学习能力能够很好地提取说话人的声学特征,d-vector[16]和 x-vector[17]是基于 DNN 模型来研发声学特征提取模型的开创性工作。d-vector 可以提取帧级音频的声学特征,并将说话人确认任务转化为分类模型进行训练;x-vector 利用多层时延神经网络结构 TDNN 和统计池化层将帧级音频特征连接为段级来作为说

作者简介: 胡鸿淋(1999—),男,硕士研究生,主要研究方向:深度学习,时序信号处理。
通信作者: 熊淑华(1969—),女, 博士,副教授,硕士生导师,主要研究方向:图像视频通信,图像处理和信息论。Email:xiongsh@scu.edu.cn。

收稿日期: 2025-01-08

话人嵌入，两者都取得了不错的说话人识别准确率，但是当遇到与训练数据分布差异较大的域外数据时，两者的识别准确率都会显著下降，因此，说话人确认研究逐渐倾向于更为复杂的网络结构开发，以增强对于域外信号的处理能力。

近年来，不断有学者对残差网络 ResNet 进行改进，并在图像处理等领域取得不俗成果，因此吸引研究者们将 ResNet 应用于说话人确认的研究中来。Gao 等学者[18]提出 Res2Net 来增强残差网络提取多尺度特征的能力。Desplanques 等学者[19]提出了 ECAPA-TDNN 模型，通过引入强调通道注意力机制和自适应统计池化，有效提升了语音特征的提取和表示能力，是当前取得先进结果的模型之一。Jung 等学者[20]提出用特征金字塔模块改进多尺度聚合，以实现鲁棒说话人验证。但是这些方法却都着重关注了语音信号的短时局部特征，或只是简单运用全局特征，缺乏对于层级特征的处理与运用，使其在面对复杂场景时的识别准确率仍有待提升。

针对上述问题，本文提出基于融合多分辨率层级注意力机制的改进 Res2Net 算法。基于 Res2Net 的总体框架，在全局特征和局部特征提取时引入不同的注意力机制模块，在全局特征提取上使用多分辨率层级特征融合的卷积块注意力模块（Convolutional Block Attention Module，CBMA）模块，在局部特征上使用局部特征注意力融合模块（Attetion Feature Fusion，AFF）模块，对比原模型，在说话人确认任务的准确率上取得明显的效果提升。

1　系统结构

本文系统在 Res2Net 的基线系统上添加了层级注意力特征机制和局部特征注意力融合机制，总体结构如图 1 所示。首先提取出语音信号的 FBank 特征，然后依次经过 4 层改进的 Res2Net，每层网络由一定数量的改进 Res2Net Block 堆叠而成。将第 3 层和第 4 层的输出进行层级特征融合，输出经过池化和全连接层得到一个说话人嵌入，最后经过分类器判定这段语音是否是目标说话人。

1.1　改进的 Res2Net Block

Res2Net 的提出主要是为了在更细粒度级别提取多尺度特征，并在后续网络中逐步融合这些不同尺度的特征，Res2Net Block 如图 2（a）所示。虽然 Res2Net 引入多尺度特征提取，但同时也巧妙地控制了计算成本，保持了较高的计算效率。Res2Net 将前一个尺度的卷积结果直接与下一个尺度的输入相加，这种不同尺度特征融合的设计在一定程度上实现了多尺度特征的聚合，但是简单的相加运算缺乏对多尺度局部特征的有效交互，因此本文使用改进的 Res2Net Block，如图 2（b）所示。

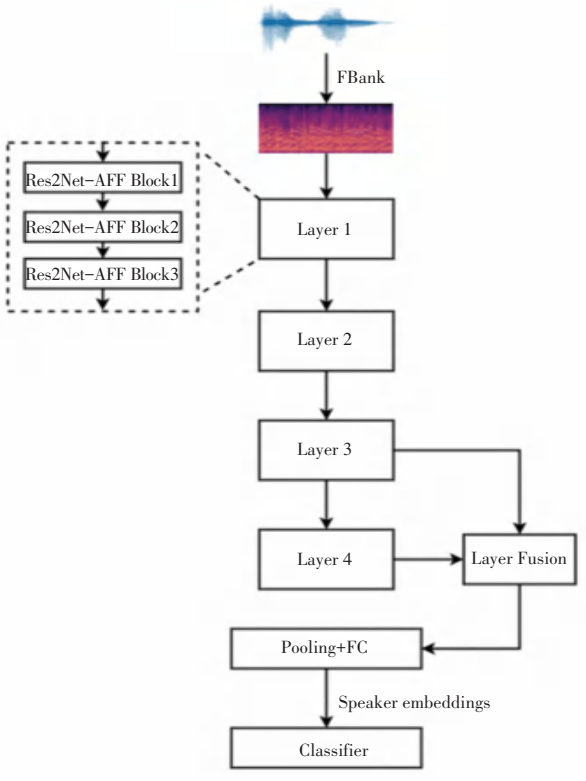


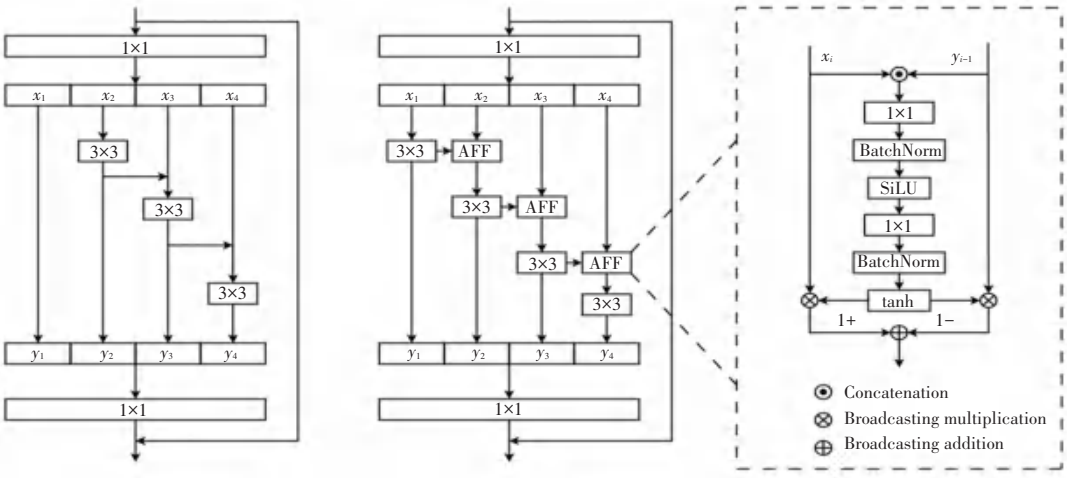
图 1　本文网络总体结构

Fig. 1　Overall network structure

注意力特征融合机制能够明显加强网络的感受野，并从通道维度整合局部信息，其运算流程如图 2（c）所示。首先将输入特征图分组 x_i 和前一组输出 y_{i-1} 按通道维度拼接在一起，经过 1×1 卷积层 C_1 通道进行降维计算，本文降维系数设置为 4，再经过批归一化（BatchNorm）和激活函数 SiLU，此后经过 1×1 卷积层 C_2 将通道数升回特征图 x_i 的通道大小，接下来又经过批归一化和 tanh 函数，获得通道的注意力权重 Att，Att 可表示为：

$$Att = \tanh(BN(C_2 \cdot SiLU(BN(C_1 \cdot [x_i, y_{i-1}])))) \tag{1}$$

获取到注意力权重后，将当前组的输入特征图 x_i 与 $(1+Att)$ 进行元素积，将 y_i 与 $(1-Att)$ 进行元素积，在此基础上进行元素相加操作。这一设计目的是考虑到激活函数 tanh 的输出范围是 $(-1,1)$，两者相加可以根据特征的重要程度进行动态加权和特征组合，提高模型从输入信号中提取相关信息的能力。因此，改进的 Res2Net Block 的输出 y_i 可表示为：



(a) Res2Net Bolck 结构 (b) 改进后的 Res2Net Block 结构 (c) 注意力特征融合机制

图 2 Res2Net Block 与改进后的结构对比

Fig. 2 Comparison between Res2Net Block and improved structure

$$y_i = \begin{cases} K_i(x_i), & i = 1 \\ K_i((Att([x_i, y_{i-1}])) + 1) \times x_i + \\ \quad (1 - Att([x_i, y_{i-1}])) \times y_{i-1}), & i > 1 \end{cases} \tag{2}$$

1.2 多分辨率层级注意力特征融合机制

除了关注局部特征,语音数据的全局特征对于说话人识别任务也是非常重要的信息。为了更广泛地捕获全局上下文信息,部分研究会引入多分辨率的层级特征。一般的多分辨层级特征表示为将各层输出做直接拼接,缺乏对不同层输出特征差异化的关注,使得层级间的特征交互过于简单,难以更高效地获取上下文信息,且缺乏对空间域的关注,因此本文提出多分辨率层级注意力特征融合机制,通过下采样,实现将不同分辨率的层级特征进行拼接,然后依次通过通道注意力模块(CAM)和空间注意力模块(SAM),实现层级特征融合,如图 3 所示。

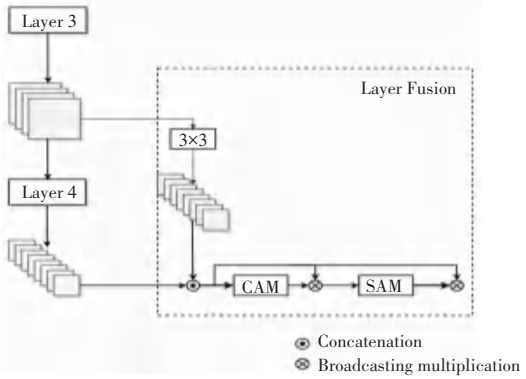


图 3 多分辨率层级注意力特征融合机制

Fig. 3 Multi-resolution layer attention feature fusion mechanism

通道注意力模块原理如下。首先对特征图分别进行空间维度的全局最大池化和全局平均池化,获得 2 个池化结果,然后分别送入同一个多层感知机(MLP)中进行学习。MLP 由 2 层卷积层组成,分别得到 2 个 $1 \times 1 \times C$ 的特征图,在按元素将两者相加后计算 *Sigmoid* 值得到通道注意力权重,可表示为:

$$M_C = Sigmoid(MLP(MaxPool(x_f)) + MLP(AvgPool(x_f))) \tag{3}$$

$$MLP = W_1 \cdot ReLU(W_0(x)) \tag{4}$$

空间注意力机制原理如下。首先对特征图分别进行通道维度的全局最大池化和全局平均池化,得到 2 个大池化特征,将 2 个池化结果按通道拼接得到特征图,再经过 1 个 7×7 的卷积层,压缩通道大小为 1,得到 $H \times W \times 1$ 的特征图,经过 *Sigmoid* 函数后即为空间注意力权重,可表示为:

$$M_S = W([MaxPool(x_f), AvgPool(x_f)]) \tag{5}$$

因此,多分辨率层级注意力特征融合结果可表示为:

$$F_i = [L_{i+1}, W_{\mathrm{di}}(L_i)] \tag{6}$$

$$LF_i = F_i \times M_C(F_i) \times M_S(F_i) \tag{7}$$

其中,LF_i 表示层级特征融合的输出;L_i 表示每一层的直接输出;F_i 表示拼接特征;W_{di} 表示下采样层。

通过层级注意特征融合的引入,网络通过训练来学习对不同分辨率层级输出的最优融合结果,优化每一层级输出对最终结果的表达影响,更高效率地提取出对结果有用的上下文信息,增强网络对全局特征的感知能力,减少冗余信息的干扰,进一步提高模型的泛化能力。

2　实验与分析

2.1　数据集

本文采用 VoxCeleb2[21] 作为训练数据，其包含了 5 994 名说话人的 1 092 009 条音频数据，平均音频长短在 8 s 左右、最短为 3 s，短语音较多，总计语音时长超过 2 000 h。测试集采用 VoxCeleb1[22]，总计包含 1 251 名说话人的 153 516 条语音数据，用数据集官方公布的 3 个测试任务进行测试评估：VoxCeleb1－O（original）、VoxCeleb1－E（extended）、VoxCeleb1－H（hard）。

2.2　评价指标

说话人确认任务一般使用等错误率（*EER*）和最小检测代价函数（*minDCF*）作为性能评价指标。为了兼顾便利性和安全性，常用错误接受率（*FAR*）和错误拒绝率（*FRR*）相等时所对应的阈值作为最终阈值，并得到等错误率。等错误率越低，代表算法性能越好，反之算法性能则越差。

minDCF 在 EER 的基础上考虑了先验概率和不同代价，其计算公式为：

$$minDCF = C_{\mathrm{fa}} \times FAR \times (1 - p_{\mathrm{target}}) + C_{\mathrm{fr}} \times FRR \times p_{\mathrm{target}} \tag{8}$$

其中，C_{fa} 和 C_{fr} 分别表示错误接受样本和错误拒绝样本的风险系数；p_{target} 表示正例对的先验概率，$1 - p_{\mathrm{target}}$ 表示负例对的先验概率。本文中设置 $C_{\mathrm{fa}} = C_{\mathrm{fr}} = 1$，$p_{\mathrm{target}} = 0.01$。*minDCF* 越小，代表系统总体代价更低，性能也越好。

2.3　实验条件与参数设置

本文实验的操作系统为 Ubantu20.04.4，硬件采用 CPU 是 Intel（R）Xeon（R）Platinum 8481C，GPU 是 RTX 4090D（24 GB），开发环境为 Pytorch1.11，Cuda11.3，Python3.8。

实验时，说话人特征提取采用帧长为 25 ms、步长为 10 ms 的 80 维滤波器组特征 FBank。训练中 *Epoch* 设置为 70，*batchsize* 设置为 200，采用随机梯度下降优化器（SGD），损失函数采用 AAM－Softmax，输出的说话人嵌入大小设置为 192，每层残差网络分别包含 3、4、6、3 个残差块。

2.4　实验结果及分析

为了说明本文方法的有效性，首先对多个已有模型在 VoxCeleb1－O 测试集上的效果表现进行对比。实验结果见表 1。本文方法的 *EER* 和 *minDCF* 比基线系统 Res2Net[7] 分别高出 41.7% 和 29.7%，比变体残差网络 ResNet34－FPM[9] 高 55.6% 和 48.3%，比 Res2Net 其他变体 Res2Net－26w8s[15] 高 39.3% 和 27.9%，比 TDNN 变体网络 ECAPA－TDNN[8] 和 D－TDNN[16] 分别高 12.8%、16.5% 和 22.1%、7.8%，结果显示本文方法在准确率上取得领先。

表 1　各模型 VoxCeleb1－O 测试实验结果对比

Tabel 1　Comparison of VoxCeleb1－O test results of various models

模型	*EER*/%	*minDCF*
Res2Net[7]	1.51	0.148
ResNet34－FPM[9]	1.98	0.205
Res2Net－26w8s[15]	1.45	0.147
ECAPA－TDNN[8]	1.01	0.127
D－TDNN[16]	1.13	0.115
本文	**0.88**	**0.106**

分析可知，这说明了局部注意力特征融合机制与层级注意力特征融合机制两者相结合的有效性，两者的结合使得模型不仅对局部特征提取得更加准确，也能从语音信号整体构建出完整且准确的特征表达，使得该模型在复杂多变的现实场景中依然能表现出鲁棒性，显著提高了在复杂环境下说话人确认的准确率。为深入探究上述各部分对模型性能的具体贡献，明确其在整个算法中的不可或缺性，本文还开展了系列消融实验，在基线系统 Res2Net 上依次添加局部特征融合模块和层级注意力融合机制，依然是在 VoxCeleb1 上进行了训练，在 VoxCeleb2 上进行测试，测试结果还包括了 VoxCeleb1－E 和 VoxCeleb1－H 另外 2 个测试集，这 2 个测试集相较于 VoxCeleb1－O，包含更多的语音序列，涉及更丰富应用场景与环境影响，测试结果应该更为贴近现实情况。消融实验结果见表 2。

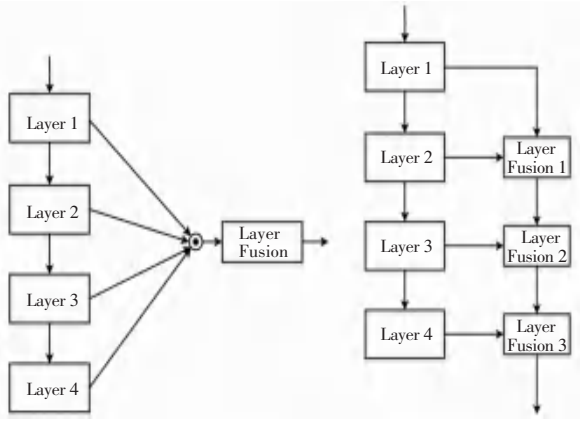
表 2　消融实验

Table 2　Ablation experiment

模型	VoxCeleb1－O		VoxCeleb1－E		VoxCeleb1－H	
	EER/%	*minDCF*	*EER*/%	*minDCF*	*EER*/%	*minDCF*
Res2Net	1.51	0.148	1.38	0.148	2.40	0.224
Res2Net+AFF	1.12	0.108	1.22	0.129	2.07	0.201
Res2Net+LF	1.37	0.129	1.35	0.144	2.36	0.218
Res2Net+AFF+LF	**0.88**	**0.106**	**1.03**	**0.116**	**1.93**	**0.199**

消融实验结果显示,Res2Net 加上局部特征融合模块后,3 个测试集在等错误率上分别提升了 25.8%、11.5%、13.7%, *minDCF* 分别提升了 27.1%、12.8%、10.2%,综上分析说明了本文的局部特征融合模块在局部特征提取上能提取更细粒度的特征,这使得网络能够更准确地计算出说话人嵌入,证明了局部特征提取的有效性。Res2Net 结合层级注意力融合机制后,3 个测试集在等错误率上分别提升了 9.2%、2.1%、1.6%, *minDCF* 分别提升了 12.8%、2.7%、6.7%,这说明了层级注意力特征融合机制能对多分辨率的层级输出结果进行有效分析,将多分辨率的全局特征进行高效融合,使得最终提取的说话人嵌入能够更准确地反映说话人的声学特征。最后将上述 2 个模块结合使用,3 个测试集在等错误率上分别提升了 41.7%、25.3%、19.6%, *minDCF* 分别提升了 29.7%、21.6%、11.1%,对比 2 个模块单独使用的情况,等错误率和 *minDCF* 均取得最佳,局部特征融合与层级注意力融合机制相结合,使得网络对于说话人嵌入提取地更加准确。这种融合方式不仅避免了单纯依赖低层级细节特征可能导致的信息冗余和噪声干扰,也克服了仅关注高层级语义特征而忽略局部细节的问题,使得最终生成的全局特征表示更加全面、准确地反映了说话人的身份特征。

本文还继续在层级注意力机制上进行实验探讨,与另外 2 种层级注意力机制方案进行对比,如图 4 所示。图 4(a)是直接将 4 层的网络输出都保留下来,将不同分辨率的 4 层输出经过下采样后按通道拼接,再通过 CBAM 注意力机制。图 4(b)则是将上一层网络的输出与一层网络依次进行特征融合,本文则是只将 3,4 层的输出进行了特征融合。3 种网络的训练参数设置基本一致,实验结果见表 3。



(a) 4 层输出直接融合 (b) 每层输出依次融合

图 4 2 种不同的层级注意力特征融合机制

Fig. 4 Two different levels of attention feature fusion mechanism

表 3 不同层级注意力机制实验结果对比

Table 3 Comparison of experimental results of different layer attention mechanism

模型	*Train_acc*/ %	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER/%	*minDCF*	*EER*/%	*minDCF*	*EER*/%	*minDCF*
Layer fusion cat1-4 (a)	98.20	1.81	0.185	2.02	0.212	2.85	0.294
Layer fusion 1-4 (b)	**99.07**	2.11	0.230	2.34	0.241	3.25	0.323
Layer fusion 3-4 (本文)	98.28	**0.88**	**0.106**	**1.03**	**0.116**	**1.93**	**0.199**

从实验结果来看,第 1 种层级注意力机制直接融合了 4 层不同分辨率的输出,其训练准确率与直接融合 3、4 层差别不大,但是测试泛化性时表现较差。第 2 种依次进行层级特征融合训练时准确度提高了 0.79%,但是测试泛化性时表现反而降低,这表明本文方法相较于上述 2 种特征融合机制,不仅取得更好的准确率,还降低了计算冗余和参数数量。究其原因是 3、4 层相较于第 1、2 层是更高维度的特征,其中包含更丰富的特征信息以及更抽象的模式表征,在层级特征融合时加入第 1、2 层这种低维度特征,可能引入了额外的连接或者计算逻辑,破坏了原有特征空间的内在结构,使得模型在面对未知数据时难以准确提取有效的特征进行判别,进而导致测试泛化性表现反而下降。

3 结束语

本文针对说话人确认任务,提出了基于注意力融合机制的 Res2Net 说话人确认算法,在全局特征和局部特征提取时引入不同的注意力机制模块,有效地融合了多层次的特征信息。在 VoxCeleb 数据集上的系列实验结果表明,本文算法有效降低了说话人确认算法的等错误率及最小检测代价函数。未来,基于深度学习的语音识别方法依然会是语音相关研究领域的主流方向,后续工作应该进一步探索网络的轻量化与鲁棒性,提高噪声抗性与跨域性能,并与实际应用场景相结合。

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(1-3):19-41.

[2] YU Qiuchen, ZHOU Ruohua. Wake word detection based on Res2Net[J]. arXiv preprint arXiv,2209.15296,2022.

[3] CHEN Yafeng, ZHENG Siqi, WANG Haibo, et al. An enhanced Res2Net with local and global feature fusion for speaker verification[J]. arXiv preprint arXiv,2305.12838, 2023.

[4] KENNY P, BOULIANNE G, OUELLET P, et al. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition[J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(4):1435-1447.

[5] BAI Zhongxin, ZHANG Xiaolei. Speaker recognition based on deep learning: An overview[J]. Neural Networks, 2021, 140: 65-99.

[6] ZHOU Tianyan, ZHAO Yong, WU Jian. ResNeXt and Res2Net structures for speaker verification[C]//Proceedings of 2021 IEEE Spoken Language Technology Workshop (SLT). Piscataway,NJ: IEEE, 2021: 301-307.

[7] YU Yaqi, ZHENG Siqi, SUO Hongbin, et al. Cam: Context-aware masking for robust speaker verification[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 6703-6707.

[8] 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.

[9] CAMPBELL W M, STURIM D E, REYNOLDS D A. Support vector machines using GMM supervectors for speaker verification [J]. IEEE Signal Processing Letters, 2006, 13(5):308-311.

[10] LIU Mengyao, LI Xueqing, WANG Mou, et al. MTBV: Multi-trigger backdoor attacks on speaker verification[C]//Proceedings of 2024 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). Piscataway, NJ: IEEE,2024: 1-5.

[11] LI J, ZHANG Y, WANG X, et al. Deep learning-enhanced speaker recognition in complex acoustic environments[J]. Journal of Audio Engineering Society, 2023, 71(5):345-355.

[12] ZHANG M, LI N, ZHAO Q, et al. Speaker recognition with adversarial training for noise resistance[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10):4815-4826.

[13] CHEN Q, HUANG Y, WU Z, et al. Unsupervised domain adaptation for cross-domain speaker recognition[J]. Pattern Recognition, 2023, 140:109669.

[14] ZHANG R, WEI J G, LU X G, et al. Unsupervised adaptive speaker recognition by coupling-regularized optimal transport[J]. ACM Transactions on Audio, Speech, and Language Processing, 2024,32:3603-3617.

[15] ZHANG Leying, CHEN Zhengyang, QIAN Yanmin. Adaptive large margin fine-tuning for robust speaker verification[C]// Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ:IEEE, 2023: 1-5.

[16] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ:IEEE, 2014: 4052-4056.

[17] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C] // 18th Annual Conference of the International Speech Communication Association (InterSpeech 2017). Stockholm: ISCA, 2017: 999-1003.

[18] GAO Shanghua, CHENG Mingming, ZHAO Kai, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2):652-662.

[19] 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.

[20] JUNG Y, KYE S M, CHOI Y, et al. Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances[C]//Proceedings of 21th Annual Conference of the International Speech Communication Association. Piscataway, NJ:IEEE, 2020: 1501-1505.

[21] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset[J]. arXiv preprint arXiv, 1706.08612, 2017.

[22] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[J]. arXiv preprint arXiv, 1806.05622, 2018.