arXiv:2603.11841v1 [eess.AS] 12 Mar 2026

ReDimNet2：通过时间池化维度重塑扩展说话人验证

Ivan Yakovlev[1],*Anton Okhotnikov*

[1] Palabra AI，伦敦，英国

i.yakovlev@palabra.ai, ant.okhotnikov@gmail.com

摘要

我们提出了 ReDimNet2，这是一种改进的神经网络架构，用于提取话语级说话人表征，构建于 ReDimNet维度重塑框架之上。ReDimNet2 的关键改进是在一维处理路径中引入沿时间维度的池化操作。由于无论时间分辨率如何，一维特征始终是二维特征重塑后的形式，因此该操作保留了一维特征空间的特性，同时能够大幅提升通道维度，而无需相应增加计算量。我们提出了包含七种模型配置（B0-B6）的一系列模型，参数量从 1.1M 到 12.3M，计算量从 0.33 到 13 十亿次乘加运算。在 VoxCeleb1 基准测试上的实验结果表明，与 ReDimNet 相比，ReDimNet2 在所有规模下均改善了计算成本与准确率之间的帕累托前沿；在 Vox1-O 上，该模型以 12.3M 个参数和 13 十亿次乘加运算实现了 0.287% 的等错误率。

索引词：说话人识别、说话人验证、语音处理、ReDimNet

1. 引言

说话人识别是一个专业领域，旨在通过个体独特的语音特征识别或验证身份。深度神经网络已成为提取说话人嵌入的主流技术，这些嵌入用于说话人验证（SV）及相关任务。通过开发新数据集 [1–4], 、架构设计 [5–18],和损失函数 [19, 20]，大量研究推动了该领域的发展。

目前已出现多种架构，包括 1D[5–7, 9, 10] 和 2D [14–18] 卷积神经网络、二者的混合架构 [8, 11, 13],以及自注意力网络 [12]。每种方法各有优势：1D 模型效率高，可直接进行时间分析；2D 架构具有频率平移不变性 [21], ；混合系统则旨在兼顾二者。近来，WavLM [22]和 W2V-BERT 2.0 [23, 24] 等大型自监督模型取得了出色的效果，但计算成本显著更高。

ReDimNet [25] 提出了一种基于二维和一维表征之间维度重塑的新方法，使这两种类型的模块能够通过残差连接在整个网络中无缝集成。其紧凑且准确的

此后，说话人嵌入已应用于各种下游任务：用于零样本文本转语音合成的说话人条件控制 [26], 语音增强中的说话人相似度评估, 以及语音生成 [27, 28], 作为用于个人语音活动检测的紧凑型说话人编码器 [29]。尽管 ReDimNet 取得了最先进的成果，但其设计要求在整个网络中保留时间分辨率，这限制了通道维度的扩展性：在不缩短时间的情况下增加通道数会导致

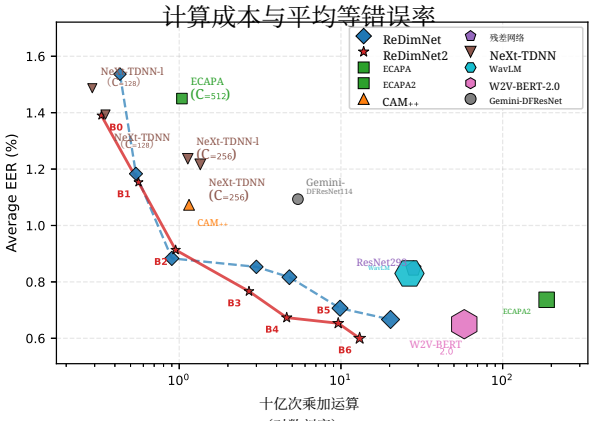


图 1 计算成本与平均等错误率。 等错误率取 Vox1-O、Vox1-E 和 Vox1-H 协议上的平均值。标记面积与参数量成正比，颜色表示模型系列。虚线：ReDimNet；实线：ReDimNet2。十亿次乘加运算数以 2 秒输入为基准测量。

一维路径中的计算量呈二次增长。

本文提出 ReDimNet2，通过在一维路径中引入沿时间维度的池化来解决这一局限。其核心洞见是，时间池化不会从根本上改变一维特征空间——一维特征仍是二维特征经过重塑后的形式，因此即使时间分辨率降低，残差连接和维度重塑逻辑依然有效。这一简单改动使得通道缩放可以更加激进，并且在每种计算预算下都能持续提升准确率（图 1）。

2. 模型架构

2.1. ReDimNet 回顾

ReDimNet [25] 以二维和一维特征图之间的维度重塑为核心。该架构将所有二维特征图的体积限制为恒定值 $V = C \cdot F \cdot T$，其中C、 F和 T 分别表示通道、频率箱和时间步。这通过以下方式实现：(1) 消除时间轴上的所有步幅；(2) 在各阶段同步频率步幅与通道增长率。由于体积恒定，任何二维特征图（C_i、 F_i、 T）都可以重塑为统一的一维表示（$C_i \cdot F_i$、 T）=（$C_0 \cdot F_0$、 T），从而能够在任意阶段之间建立残差连接。联合一维+2D块通过交错的一维卷积子模块（基础 ResNet 模块）和一维时间上下文子模块（ConvNeXt-

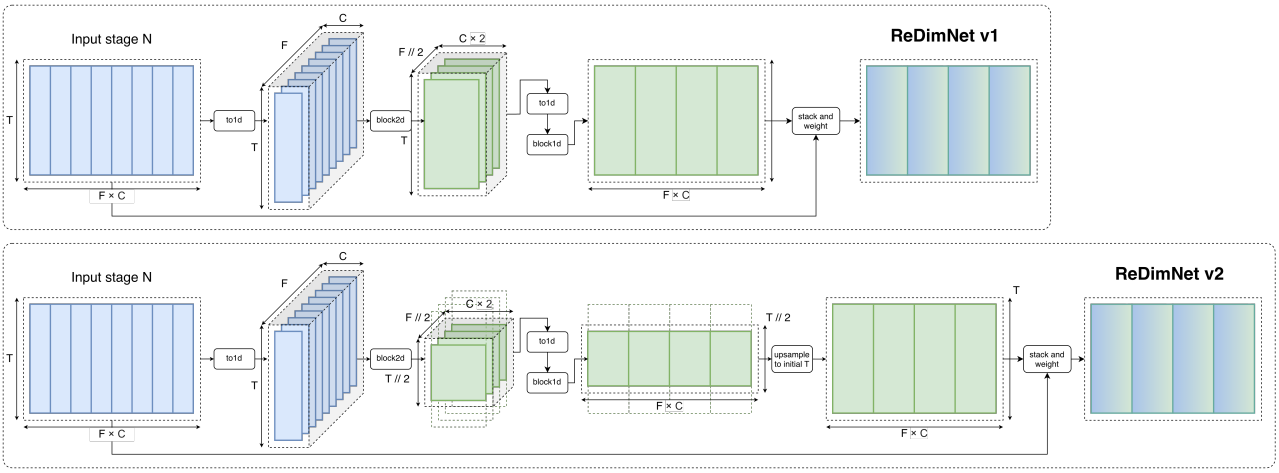


图2 ReDimNet2 单阶段架构概览 上： ReDimNet v1 对一维输入执行 to2d重塑操作、二维块（仅采用频率步长）、to1d重塑操作，随后通过一维块和逐阶段加权聚合（“堆叠并加权”）进行处理，并在整个过程中保持时间维度 T不变。下：ReDimNet2 在二维块内部沿时间轴额外采用一个步幅，使 F和 T均减半。随后，一维块对较短的序列（$T/2$）进行处理，再通过上采样至初始 T 步幅，在聚合前恢复原始时间分辨率。

表1 B3 v2配置的特征图尺寸。 S_f, S_t：频率步幅和时间步幅。时间池化阶段（3、5）将 T减半，但不调整 C，从而适度放宽 $C{\cdot}F{\cdot}T$ 约束。

块编号	输入形状	S_f	S_t	通道	输出形状	音量
1	(C, F, T)	1	1	C	(C, F, T)	$C{\cdot}F{\cdot}T$
2	(C, F, T)	2	1	$2C$	$(2C, F/2, T)$	$C{\cdot}F{\cdot}T$
3	$(2C, F/2, T)$	1	2	$2C$	$(2C, F/2, T/2)$	$C{\cdot}F{\cdot}T/2$
4	$(2C, F/2, T/2)$	2	1	$4C$	$(4C, F/4, T/2)$	$C{\cdot}F{\cdot}T/2$
5	$(4C, F/4, T/2)$	1	2	$4C$	$(4C, F/4, T/4)$	$C{\cdot}F{\cdot}T/4$
6	$(4C, F/4, T/4)$	2	1	$8C$	$(8C, F/8, T/4)$	$C{\cdot}F{\cdot}T/4$

风格的一维卷积与多头注意力相结合）处理特征，并使用注意力统计池化 [30] 进行话语级聚合。我们沿用通过 [25] 中的消融研究确定的性能最佳的一维和二维块类型。

2.2. ReDimNet2：时间池化维度重塑

ReDimNet2 的核心改动是在一维处理路径中引入**沿时间维度的池化**（图2）。在原始 ReDimNet 中，整个网络的时间轴 T都保持输入分辨率。虽然这样能够保留完整的时间信息，但一维子块的计算成本会随着通道 C 和时间步 T 的增加而线性增长，因此，大型-C 配置的成本会越来越高。

池化机制。ReDimNet2 在中间阶段采用时间池化，使用与原始架构中实现频率下采样的同一个步幅二维卷积层——该层通过在 F 减半时将通道数加倍来保持空间体积。沿时间方向应用时，通道维度不会调整，从而适度放宽了恒定体积约束 $C \cdot F \cdot T$。表1展示了B3配置下对应的特征图形状。频率池化阶段（2、4、6）精确保留体积，而时间池化阶段（3、5）将 T减半，同时不改变 C 或 F。

残差连接。 在原始 ReDimNet 中，所有阶段输出共享相同的一维形状（$B,\ C{\cdot}F,\ T$），因此可以通过学习得到的阶段级权重直接进行聚合。在 ReDimNet2 中，各阶段会生成时间长度不同的一维特征图：（$B,\ C{\cdot}F,\ T$）、（$B,\ C{\cdot}F,\ T/2$）、（$B,\ C{\cdot}F,\ T/4$）等。为保持残差连接，我们在逐阶段加权聚合之前，对所有特征图进行最近邻上采样，使其与输入的时间分辨率 T^* 对齐：

$$[(B,\, C{\cdot}F,\, T), (B,\, C{\cdot}F,\, \tfrac{T}{2}), (B,\, C{\cdot}F,\, \tfrac{T}{4}), \ldots] \\ \rightarrow [(B,\, C{\cdot}F,\, T^*), (B,\, C{\cdot}F,\, T^*), \ldots]$$

其中 $T^* = T$。每个阶段仍以降低后的时间分辨率处理自身特征；仅在聚合点进行上采样，因此每个阶段内部的计算量节省得以保留。

双重效率优势。 时间池化同时惠及两条处理路径。一维子块可直接受益于更短的序列，其成本按比例降低。二维子块的受益更为显著：由于一维到二维重塑生成的二维特征图，其空间范围取决于序列长度，因此降低的 T 也会压缩二维表示，从而进一步叠加节省效果。这种双重缩减使我们能够在相同的十亿次乘加运算预算内，更激进地扩展通道维度 C。

释放出的计算预算被重新分配给更宽的模型（更高的 C），在计算成本相当的情况下，实验结果表明这能带来更好的说话人判别能力。

2.3. 模型配置

遵循 [31],，我们定义了七种配置 B0–B6，其计算复杂度以十亿次乘加运算为界：*B0 < 0.5, B1 ≈ 0.5, B2 < 1.5, B3 < 3.0, B4 < 5.5, B5 < 10.0, B6 > 10.0.* ReDimNet2 的规模从 B0 的 110 万参数量 / 0.33 十亿次乘加运算扩展到 B6 的 1230 万参数量 / 13.0 十亿次乘加运算。

3. 实验设置

我们使用 WeSpeaker [32]训练流程，通过两阶段训练方法在 VoxCeleb2 [1] 上训练 ReDimNet2。

表 2 VoxCeleb1 清洗协议上的评估结果（等错误率，%）。除非另有说明，所有模型均在 VoxCeleb2-dev 上训练。十亿次乘加运算量以 2 秒音频片段进行测量。* = 估算值。来自 WeSpeaker 或 ECAPA2 代码库的模型在我们的环境中重新测试。† = 在 VoxBlink2 + VoxCeleb2 上训练。

模型	参数量	十亿次乘加运算	大间隔	Vox1-O	Vox1-E	Vox1-H
ReDimNet-B0 [25]	1.0M	0.43	✓	1.16	1.25	2.20
ReDimNet2-B0	**1.1M**	**0.33**	✓	**1.04**	**1.16**	**1.97**
NeXt-TDNN-l (C=128) [7]	**1.6M**	**0.29***	✗	1.10	1.24	2.12
NeXt-TDNN（C=128） [7]	1.9M	0.35*	✗	1.03	1.17	1.98
ReDimNet-B1 [25]	2.2M	0.54	✓	0.85	0.97	1.73
ReDimNet2-B1	2.1M	0.56	✓	**0.78**	**0.96**	**1.72**
ECAPA (C=512) [6, 17]	6.2M	1.04	✗	0.94	1.21	2.20
NeXt-时延神经网络-l (C=256) [7]	6.0M	1.13*	✗	0.81	1.04	1.86
CAM++ [11, 32]	7.2M	1.15	✓	0.71	0.85	1.66
NeXt-TDNN (C=256) [7]	7.1M	1.35*	✗	0.79	1.04	1.82
ReDimNet-B2 [25]	4.7M	**0.90**	✓	**0.57**	**0.76**	**1.32**
ReDimNet2-B2	**3.6M**	0.95	✓	**0.57**	**0.76**	1.41
ReDimNet-B3 [25]	**3.0M**	3.00	✓	0.50	0.73	1.33
ReDimNet2-B3	4.1M	**2.70**	✓	**0.42**	**0.66**	**1.22**
Gemini-DFResNet114 [18]	6.5M	5.42	✗	0.77	0.91	1.60
ReDimNet-B4 [25]	**6.3M**	4.80	✓	0.51	0.68	1.26
ReDimNet2-B4	6.6M	**4.62**	✓	**0.37**	**0.58**	**1.07**
ReDimNet-B5 [25]	9.2M	9.87	✓	0.43	0.61	1.08
ReDimNet2-B5	**8.9M**	**9.62**	✓	**0.33**	**0.56**	**1.07**
残差网络293 [32, 33]	28.6M	28.10	✓	0.53	0.71	1.30
ECAPA2 [13]	27.1M	187.0*	✓	0.44	0.62	1.15
WavLM [22]	324M	26.53	✓	0.52	0.63	1.34
W2V-BERT 2.0 [24]	587M	57.90	✓	0.38	**0.51**	1.06
ReDimNet-B6 [25]	15.0M	20.27	✓	0.40	0.55	1.05
ReDimNet2-B6	**12.3M**	**13.05**	✓	**0.29**	0.52	**0.99**
在扩展数据（*VoxBlink2 + VoxCeleb2*）上训练的模型：						
SimAM-残差网络34† [34]	25.2M	18.20	-	0.42	0.62	1.12
SimAM-残差网络100† [34]	50.2M	57.24	-	0.23	0.46	0.87
W2V-BERT 2.0† [24]	587M	57.90	-	0.14	0.31	0.73

3.1. 预训练阶段

模型使用带 Nesterov动量（$m = 0.9$）的随机梯度下降进行优化，权重衰减为 2×10^{-5}。我们使用 80 维、均值归一化的梅尔滤波器组对数能量，帧长为 25 ms，帧移为 10 ms。预训练期间，随机裁剪 2 秒片段，并按照 [14]应用 MUSAN [35] 和 RIR [36] 数据集中的增强方法。使用因子为 0.9 和 1.1 的双倍速度扰动。默认损失函数为 SphereFace2-C（SF2-C） [20]，边际在第 20–40 训练轮次期间从 0.0 逐步调整至 0.2，之后保持不变。学习率采用指数衰减策略，并设有 6 个训练轮次的预热，$lr_{\max} = 0.1$、$lr_{\min} = 6\times10^{-5}$。每种模型配置均采用不同的图形处理器设置进行训练，以最大限度地利用资源。我们发现，对于较大型号，多图形处理器训练设置（8 张 H100）效果更好，这是因为较大的有效批量大小会使有效学习率线性增加。

3.2. 大边际微调阶段

对于大边际微调阶段 [37], , 训练语句长度扩展至 6 秒，关闭速度扰动，并将 SF2-C 的边际设为固定值 0.3。学习率通过指数衰减降低至 10^{-4}。此阶段运行 5 个训练轮次。

3.3. 评估

使用 VoxCeleb1[38]的清洗协议评估性能：Vox1-O（原始）、Vox1-E（扩展）和 Vox1-H（困难），并采用等错误率（EER）作为指标。模型使用余弦相似度进行评分，采用完整长度的测试语句，且不进行分数归一化。GMACs 使用 thop 库计算，输入为采样率 16 kHz、时长 2 秒的原始波形。

表3 域外评估（等错误率，%）。所有模型均在 VoxCeleb2-dev 上训练。基线系统来自 [25]。

模型	SITW	VOiCES	Vox1-B	Avg
CAM++ [11]	1.34	6.30	2.79	3.48
ECAPA（C=1024） [6]	1.67	5.31	3.48	3.49
残差网络293 [33]	1.67	5.14	2.23	3.01
ECAPA2 [13]	3.64	13.26	1.81	6.24
ReDimNet-B6 [25]	0.77	3.19	1.66	1.87
ReDimNet2-B6	**0.75**	**3.13**	**1.63**	**1.84**

4. 结果

所有 ReDimNet2 配置和基线的评估结果见表 2，并在图 1 中可视化。B0–B6 的缩放对比构成了时间池化有效性的主要证据：在几乎所有计算预算匹配的情况下，ReDimNet2 的 EER 都低于相应的 ReDimNet 配置。这项覆盖整个帕累托前沿的七点对比提供了有力证据，因为 GMACs 是决定说话人验证准确率的主要因素。若从 ReDimNet2-B3（2.7 GMACs）中移除所有时间步幅，其计算量将增至约 4.0 GMACs，与 ReDimNet-B4 的 4.8 GMACs 直接可比；表 2 中包含了这一对比点。由于所有 B0–B6 配对在 GMACs 和参数量上均相匹配，这七组预算配对对比也构成了完整的跨尺度消融。

在较大规模下，性能提升最为显著，因为时间池化使得更激进的通道缩放成为可能。ReDimNet2-B6 在 Vox1-O 上实现了 **0.29%** 等错误率（EER），相较于 ReDimNet-B6 相对提升了 28%，同时所需十亿次乘加运算减少了 36%，参数量减少了 18%。即使在最小规模下，ReDimNet2-B0 在 Vox1-O 上的等错误率也从 1.16% 降至 1.04%，证实了这一优势在所有模型规模下均稳定存在。此外，ReDimNet2-B6 的表现优于 WavLM（参数量为 324M），并接近 W2V-BERT 2.0（参数量为 587M），而参数量却少了 48× 。中等规模配置同样表现突出：ReDimNet2-B3 在 Vox1-O 上超越 ECAPA2，且所需十亿次乘加运算少了 69×；从 B3 起的每种配置都优于 ResNet293（参数量为 28.6M）。

域外泛化。我们还按照 [25] 中使用的协议，在域外测试集上评估了 ReDimNet2-B6：SITW core-core 集 [39], 、VOiCES 评估集 [40], 以及 VoxCeleb1-B 困难协议 [41]。结果如表3所示。ReDimNet2-B6 在全部三个域外测试集上的等错误率均低于 ReDimNet-B6，这表明时间池化不会损害泛化能力。

训练稳定性。我们观察到，较小模型（B0–B3）的结果在不同随机种子下较为稳定；B3 在 Vox1-O 上的等错误率均值±标准差为 $0.42 \pm 0.01\%$。然而，对于较大的模型（B4–B6），我们发现结果的波动有所增加：B6 在多次运行中于 Vox1-O 上取得了 $0.32 \pm 0.05\%$ 的等错误率。这表明，尽管所提出的架构具有良好的扩展性，但训练较大模型时，可能需要额外的正则化或谨慎进行超参数调优，以确保稳定性。

5. 结论

本文介绍了 ReDimNet2，这是一种改进的说话人嵌入架构，在 ReDimNet 的基础上为一维处理路径引入了时间池化。我们证明，时间池化与维度重塑框架兼容：用于频率下采样的同一体积保持型二维卷积层也可处理时间池化，而聚合点处的最近邻上采样则保持了完整的残差连接。由此形成的模型系列（B0–B6）在每个相同计算预算下，都比 ReDimNet 实现了更好的准确率与效率权衡。具体而言，ReDimNet2-B6 仅需 12.3M 个参数和 13 十亿次乘加运算，在 Vox1-O 上取得 0.29% 的等错误率，在 Vox1-H 上取得 0.99% 的等错误率，同时保持了强大的域外泛化能力。我们认为，时间池化是一种简单而有效的策略，可用于扩展面向说话人验证的维度重塑架构。模型代码、训练方案和预训练权重已发布于 https://github.com/PalabraAI/redimnet2。

6. AI 工具使用声明

使用了生成式 AI 助手（Claude，Anthropic）对稿件文本进行编辑和润色。所有科学内容、实验设计、模型架构和分析均完全由作者完成。

7. 参考文献

[1] J. S. Chung, A. Nagrani, 和 A. Zisserman，“VoxCeleb2：深度说话人识别”，发表于*Interspeech2018*。ISCA，2018年9月。[2] Y. Lin, X. Qin, G. Zhao, M. Cheng, N. Jiang, H. Wu, 和 M. Li，“VoxBlink：一个基于摄像头的大规模说话人验证数据集”，2023年。[3] S. Zheng, L. Cheng, Y. Chen, H. Wang, 和 Q. Chen，“3D-Speaker：用于语音表征解耦的大规模、多设备、多距离、多方言语料库”，2023年。[4] I. Yakovlev, A. Okhotnikov, N. Torgashov, R. Makarov, Y. Voevodin, 和 K. Simonchik，“VoxTube：一个多语种说话人识别数据集”，发表于*INTERSPEECH 2023*会议录，2023年，第2238–2242页。[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, 和 S. Khudanpur，“X-vectors：用于说话人识别的鲁棒深度神经网络嵌入”，发表于*2018 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2018年，第5329–5333页。[6] B. Desplanques, J. Thienpondt, 和 K. Demuynck，“ECAPA-TDNN：基于时延神经网络的说话人验证中的强调通道注意力、传播与聚合”，*arXiv*预印本 *arXiv:2005.07143*，2020年。[7] H.-J. Heo, U.-H. Shin, R. Lee, Y. Cheon, 和 H.-M. Park，“Next-TDNN：面向说话人验证的多尺度时间卷积骨干网络现代化”，2023年。[8] T. Liu, R. K. Das, K. A. Lee, 和 H. Li，“MFA：结合多尺度频率-通道注意力的时延神经网络，用于短语音的文本无关说话人验证”，2022年。[9] Z. Zhao, Z. Li, W. Wang, 和 P. Zhang，“PCF：采用渐进通道融合的ECAPA-TDNN说话人验证网络”，2023年。[10] Y.-Q. Yu 和 W.-J. Li，“ 用于说话人验证的密集连接时延神经网络 ”，发表于国际语音通信协会年会（*INTERSPEECH*），2020年，第921–925页。[11] H. Wang, S. Zheng, Y. Chen, L. Cheng, 和 Q. Chen，“CAM++：一种使用上下文感知掩蔽的快速高效说话人验证网络”，发表于*INTERSPEECH 2023*会议录，2023年，第5301–5305页。[12] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, 和 H. Meng，“MFA-Conformer：用于自动说话人验证的多尺度特征聚合Conformer”，2022年。[13] J. Thienpondt 和 K. Demuynck，“ECAPA2：一种用于鲁棒说话人嵌入的混合神经网络架构与训练策略”，2024年。

[14] D. Garcia-Romero, G. Sell, 和 A. Mccree，“Magneto：X向量幅度估计网络加偏移，以提升说话人识别性能”，发表于 *Odyssey*，2020年，第1–8页。[15] T. Zhou, Y. Zhao, 和 J. Wu, “用于说话人验证的ResNeXt与Res2Net结构”，2020年。[16] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, 和 J. Qi，“采用局部与全局特征融合的增强型Res2Net说话人验证网络”，发表于 *INTERSPEECH 2023*会议录，2023年，第2228–2232页。[17] B. Liu, Z. Chen, S. Wang, H. Wang, B. Han, 和 Y. Qian，“DF-ResNet：通过深度优先设计提升说话人验证性能”，发表于 *INTERSPEECH 2022*会议录，2022年，第296–300页。[18] T. Liu, K. A. Lee, Q. Wang, 和 H. Li，“Golden gemini：说话人验证的理想选择”，2023年。[19] J. Deng, J. Guo, N. Xue, 和 S. Zafeiriou，“ArcFace：用于深度人脸识别的加性角度间隔损失”，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第4690–4699页。[20] B. Han, Z. Chen, 和 Y. Qian，“探索用于说话人验证的二分类损失”，发表于*ICASSP 2023——2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2023年，第1–5页。[21] J. Thienpondt, B. Desplanques, 和 K. Demuynck，“在时延神经网络中融入频率平移不变性，并在二维ResNet中融入频率位置信息，以增强说话人验证性能”，*arXiv*预印本 *arXiv:2104.02370*，2021年。[22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao等，“WavLM：面向全栈语音处理的大规模自监督预训练”，发表于 *IEEE*信号处理精选主题期刊，第16卷，第6期。IEEE，2022年，第1505–1518页。[23] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Gober, K. Singh, P. von Platen, Y. Saraf, J. Pino等，“大规模多语种自监督语音模型”，发表于国际机器学习会议。PMLR，2022年，第1473–1483页。[24] Z. Li, M. Cheng, 和 M. Li，“利用W2V-BERT 2.0和知识蒸馏引导的结构化剪枝增强说话人验证”，*arXiv*预印本 *arXiv:2510.04213*，2025年。[25] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, 和 N. Torgashov，“用于说话人识别的重塑维度网络”，发表于*INTERSPEECH 2024*会议录，2024年。[26] N. Torgashov, G. E. Henter, 和 G. Skantze，“VoXtream：超低延迟全流式文本转语音”，发表于*ICASSP*会议录，2025年。[27] H. Li, N. Hou, Y. Hu, J. Yao, S. M. Siniscalchi, X. Zhuang, D. Ye, W. Yang, 和 E. S. Chng，“通过感知反馈对齐生成式语音增强”，发表于*ICASSP*会议录，2026年。[28] V. Pankov, A. Gribul, O. Tatanov, V. Proskurov, Y. Korotkova, D. Mylzenova, 和 D. Vypirailenko，“PFluxTTS：结合鲁棒跨语言语音克隆与推理时模型融合的混合流匹配文本转语音”，发表于*ICASSP*会议录，2026年。[29] Y. Lin, J. Chen, W. Li, L. Xiao, 和 C. Weng，“用于缓解领域失配与误接受情形的鲁棒个性化语音活动检测”，发表于 *Interspeech*会议录，2025年，第5783–5787页。[30] K. Okabe, T. Koshinaka, 和 K. Shinoda，“用于深度说话人嵌入的注意力统计池化”，发表于*Interspeech 2018*。ISCA，2018年9月。[31] M. Tan 和 Q. V. Le，“EfficientNet：重新思考卷积神经网络的模型缩放”，2020年。[32] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, 和 Y. Qian，“WeSpeaker：面向研究与生产的说话人嵌入学习工具包”，发表于*ICASSP 2023——2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2023年，第1–5页。[33] K. He、X. Zhang、S. Ren 和 J. Sun，《用于图像识别的深度残差学习》，载于 *2016*年 *IEEE*计算机视觉与模式识别会议 *(CVPR)*，2016年，第770–778页。[34] X. Qin、N. Li、C. Weng、D. Su 和 M. Li，《结合迭代噪声标签检测的简易注意力模块说话人验证》，载于 *ICASSP 2022 - 2022*年 *IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2022年，第6722–6726页。[35] D. Snyder、G. Chen 和 D. Povey，《MUSAN：音乐、语音与噪声语料库》，*arXiv*预印本 *arXiv:1510.08484*，2015年。[36] I. Sz¨oke、M. Sk´acel、L. Mo˘sner、J. Paliesek 和 J. ˇCernock`y，《真实房间脉冲响应数据集的构建与评估》，*IEEE*信号处理精选主题期刊，第13卷，第4期，第863–876页，2019年。[37] Y. Liu、L. He 和 J. Liu，《用于说话人验证的大间隔softmax损失》，*arXiv*预印本 *arXiv:1904.03479*，2019年。[38] A. Nagrani、J. S. Chung、W. Xie 和 A. Zisserman，《VoxCeleb：野外大规模说话人验证》，计算机科学与语言，2019年。[39] M. McLaren、L. Ferrer、D. Castan 和 A. Lawson，《野外说话人 (SITW) 说话人识别数据库》，载于 *Interspeech*，2016年，第818–822页。[40] M. K. Nandwana、J. van Hout、M. McLaren、C. Richey、A. Law- son 和 M. A. Barrios，《2019年远距离语音挑战赛评估方案》，2019年。[41] K. Nam、Y. Kim、J. Huh、H. S. Heo、J. weon Jung 和 J. S. Chung，《面向多语种说话人识别的解耦表示学习》，2023年。