面向 TidyVoice 2026 挑战赛的语言不变的多语言说话人验证

ZeLi[1,4], *Xiaoxiao Miao*[4], *Juan Liu*[2], *MingLi*[2,3,4,**]

[1] 武汉大学计算机科学学院，中国武汉[2] 武汉大学人工智能学院，中国武汉[3] 香港中文大学（深圳）人工智能学院，中国深圳[4] 昆山杜克大学数字创新研究中心，中国昆山

lize389@whu.edu.cn, xiaoxiao.miao@dukekunshan.edu.cn, liujuan@whu.edu.cn, ming.li.cuhksz@gmail.com

arXiv:2603.08092v1 [eess.AS] 9 Mar 2026

摘要

由于跨语言数据有限且说话人嵌入中包含依赖语言的信息，多语言说话人验证（SV）仍面临挑战。本文提出了一种面向 TidyVoice 2026 挑战赛的语言不变多语言说话人验证系统。我们采用多语言自监督 w2v-BERT 2.0 模型作为骨干网络，并通过层适配器和多尺度特征聚合加以增强，以更好地利用多层表示。我们采用结合梯度反转层的语言对抗训练策略，以促进生成语言不变的说话人嵌入。此外，我们使用多语言零样本文本转语音系统合成多种语言的语音，从而提高语言多样性。实验结果表明，对大规模预训练模型进行微调可取得有竞争力的性能，而语言对抗训练进一步增强了鲁棒性。此外，在训练数据有限的情况下，合成语音增强还能带来额外收益。源代码见 https://github.com/ZXHY-82/ LI-MSV-TidyVoice2026。

索引词：说话人验证、跨语言、语言不变

1. 引言

说话人验证（SV）旨在通过分析说话人的语音样本来验证其身份。近年来，随着深度神经网络的快速发展以及大规模标注语音数据集的出现，基于深度学习的说话人验证系统 [1–4] 在各种声学条件下均取得了显著的性能。然而，在语言不匹配的条件下，说话人验证系统的性能会显著下降，而该领域对以英语为中心的数据集的依赖进一步加剧了这一问题。此外，单个说话人的多语言语音数据往往有限，导致说话人嵌入将身份信息与语言特定特征纠缠在一起，从而降低跨语言泛化能力和鲁棒性。

大规模自监督预训练模型（预训练模型），如 WavLM [5], 、wav2vec 2.0 [6], 、HuBERT [7] 和 w2v-BERT 2.0 [8], ，在数十万甚至数百万小时的未标注语音数据上训练，并提供丰富的语音表征。这些模型在研究中被越来越多地采用，以提升包括说话人验证任务在内的各类下游任务的性能。特别是，w2v-BERT 2.0 预训练模型在涵盖 143 种语言的 450 万小时未标注语音上训练，因此非常适合跨语言说话人验证任务。Li 等人 [9] 基于 w2v-BERT 2.0 构建了最先进的说话人验证系统。在其方法中，层适配器 [10] 被应用于每个 Conformer 层的输出，以降低维度并促进领域自适应。随后，利用多尺度特征聚合（MFA） [11] 框架对所有层的自适应特征进行聚合，以生成说话人嵌入，并在训练过程中采用低秩适应（LoRA） [12] 来高效微调模型。

从多语言数据中提取的说话人嵌入往往不仅包含身份信息，还包含语言特定特征，这可能会降低跨语言泛化能力和鲁棒性。为缓解这一问题，有必要推动语言不变的说话人表示学习。类似方法已应用于其他场景，例如年龄不变的说话人表示学习 [13], , 其中采用对抗训练来抑制嵌入中的年龄相关信息。

在这项工作中，我们遵循 [9] 中的方法，构建基于 w2v-BERT 2.0 预训练模型的说话人验证系统。为进一步提升跨语言鲁棒性，我们引入语言对抗训练策略，以促进语言不变的说话人表示的学习，并抑制嵌入空间中与语言相关的变化。此外，尽管官方 TidyVoiceX 训练集是多语言的，但每位说话人通常只包含两到三种语言，这限制了鲁棒跨语言说话人建模所需的语言多样性。近年来，零样本文本转语音（ZS-TTS） [14, 15] 技术取得了进展，使得无需针对说话人进行专门训练，仅凭数秒的参考音频，就能合成保留目标说话人声音特征的高质量语音。这使得跨不同语言和文本内容进行灵活的语音克隆成为可能。利用这一能力，我们采用多语言 ZS-TTS 系统 Qwen3-TTS [15], 为每位说话人合成其他语言的语音。我们的目标是丰富每位说话人的多语言语音数据，从而促进学习更具语言不变性的说话人表示。

2. 方法

2.1. w2v-BERT 2.0 预训练模型的微调

W2v-BERT 2.0 是一种大规模多语言自监督语音表示模型，作为 Seam- lessM4T 框架的一部分开发而成 [8]。它在原始 w2v-BERT 架

**表示通讯作者。

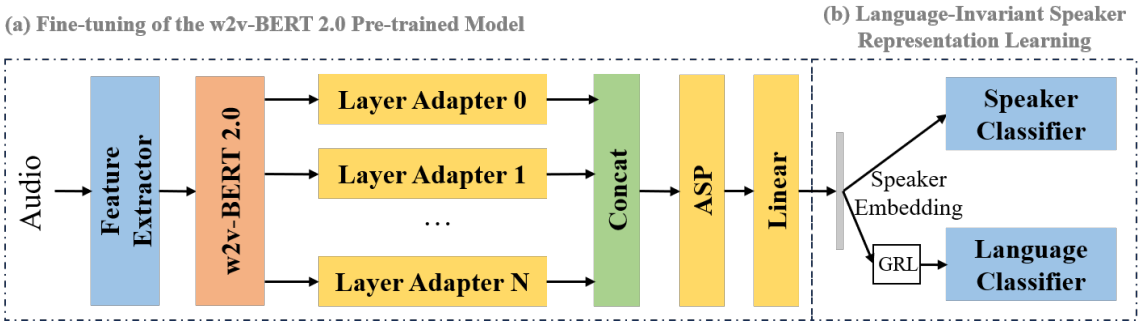


图1： 基于 *w2v-BERT 2.0* 的说话人验证系统概览，采用语言不变学习 g

构 [16] 的基础上，采用 24 层 Conformer 编码器，并联合优化对比学习和掩码预测目标。该模型使用来自 143 种语言、约 450 万小时的未标注语音数据进行训练，从而能够学习丰富且稳健的多语言语音表示。

我们的方法遵循 [9] 的方法。如图1(a)所示，给定输入语句 x，我们首先提取 Fbank 特征，并将其输入预训练的 w2v-BERT 2.0 模型。随后，将每个 Conformer 层的隐藏表示 h_i 传入各自的层适配器 [10], ，以降低特征维度，并促进说话人验证任务的领域自适应。将所有层的适配后特征 h_i' 拼接起来，再通过注意力统计池化（ASP）[17] 模块进行聚合，最后经线性投影层得到最终的说话人嵌入 e。训练期间，我们采用低秩适应（LoRA）[12]，以高效微调预训练模型。

$$[h_0, h_1, \ldots, h_L] = \text{w2v-BERT-2.0}(\text{Fbank}(x)) \quad (1)$$

$$h_i' = \text{Layer Adapter}_i(h_i), \quad i = 0, 1, \ldots, L \quad (2)$$

$$e = \text{Linear}(\text{ASP}(\text{Concat}(h_0', h_1', \ldots, h_L'))) \quad (3)$$

2.2. 语言不变的说话人表示学习

语言失配会给说话人嵌入带来不必要的变化，可能降低多语言场景下说话人验证系统的性能。为解决这一问题，我们旨在通过显式移除与语言相关的信息，学习语言不变的说话人表示，如图1(b)所示。

具体而言，我们引入了一个辅助语言分类器，并通过梯度反转层（GRL）[18]将其连接到说话人嵌入提取器。给定提取出的说话人嵌入 e，语言分类器预测语言标签，而GRL在反向传播过程中反转梯度。这种对抗学习策略促使嵌入提取器生成既能区分说话人身份、又不会泄露语言身份信息的表示。最终损失定义为：

$$\mathcal{L}_{\text{spk}}(e) = l_{\text{spk}}\left(C_{\text{spk}}(e), y_{\text{spk}}\right) \quad (4)$$

$$\mathcal{L}_{\text{lang}}(e) = l_{\text{lang}}\left(C_{\text{lang}}\left(\text{GRL}_{\lambda_{\text{GRL}}}(e)\right), y_{\text{lang}}\right) \quad (5)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spk}} + \lambda_{\text{lang}}\,\mathcal{L}_{\text{lang}} \quad (6)$$

其中 $C_{\text{spk}}(\cdot)$ 和 $C_{\text{lang}}(\cdot)$ 分别为说话人分类器和语言分类器， $l_{\text{spk}}(\cdot)$ 和 $l_{\text{lang}}(\cdot)$ 为相应的损失函数，$\text{GRL}_{\lambda_{\text{GRL}}}(\cdot)$ 是缩放系数为 λ_{GRL} 的梯度反转层，λ_{lang} 是控制语言

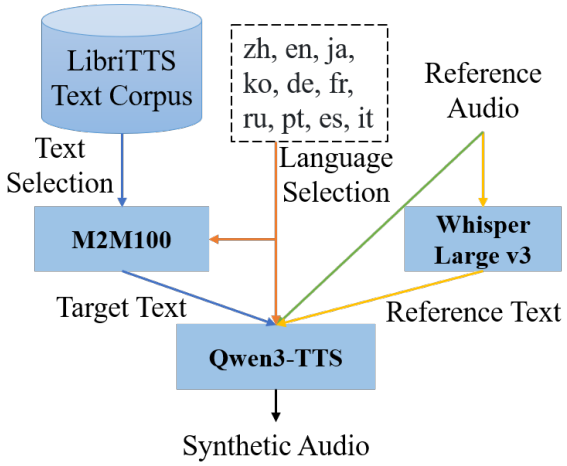


图2： 语音合成流程

损失权重的超参数， y_{spk} 和 y_{lang} 分别为说话人和语言的标签。

2.3. 多语言合成语音数据增强 ion

近期，零样本文本转语音技术的发展使得仅凭几秒钟的参考音频即可实现跨多种语言的高质量语音合成。借助这一能力，我们为每位说话人生成多种语言的合成语音，旨在提升说话人嵌入的语言不变性，并增强跨语言说话人验证性能。

在本研究中，我们采用Qwen3-TTS [15] 进行合成语音生成。Qwen3-TTS是一种能够进行多语言语音克隆的模型，可以生成十种语言的语音，包括中文、英语、日语、韩语、德语、法语、俄语、葡萄牙语、西班牙语和意大利语。如图2所示，对于文本语料库，我们使用来自LibriTTS [19], 的英语句子，然后利用M2M100 [20] 多语言翻译模型将其翻译成目标语言。参考音频通过Whisper-large-v3 [21]进行处理，以获取相应的转录文本；随后，将参考音频和文本一同输入Qwen3-TTS，以生成指定语言和目标文本对应的合成语音。

3. 实验设置

3.1. 数据集

为了训练一个稳健的多语言说话人验证系统，除 TidyVoice 提供的官方 TidyVoiceX 训练集外

2026 挑战赛 [22], 我们还纳入了多个公开数据集，包括 VoxCeleb2 [23], VoxBlink2 [24], 3D-Speaker [25], KeSpeech [26], 和 CN-Celeb1&2 [27,28], , 以增加说话人和语言的多样性。说话人验证性能在本次挑战赛提供的官方开发集和评测集上进行评估。评测集进一步分为两个子集：tv26 eval-A，包含官方训练集中出现过的语言；以及 tv26 eval-U，由 38 种未出现过的语言组成。

3.2. 多语言合成语音生成

我们使用 Qwen3-TTS-12Hz-1.7B-Base [15] 模型进行多语言语音克隆。从 TidyVoiceX 训练集中，我们为每位说话人最多选取 10 条时长超过 3 秒的语句作为参考音频，共计 3,495 条参考音频。对于每条参考音频，我们使用 Qwen3-TTS 支持的 10 种语言合成语音，并在每种语言中生成 10 条语句。总计生成 349,500 条合成语句，用于多语言数据增强。

3.3. 训练细节

我们的训练过程分为两个阶段，包括 (i) 大规模说话人模型预训练，以及 (ii) 基于语言不变学习的 TidyVoiceX 训练集微调。

3.3.1. 大规模说话人模型预训练

在此阶段，我们使用多个大型公开说话人数据集（包括 VoxCeleb2、VoxBlink2、3D-Speaker、KeSpeech 和 CN-Celeb1&2）对预训练的 w2v-BERT 2.0 模型进行微调，以获得一个鲁棒的大规模预训练说话人模型。

在初始训练阶段，w2v-BERT 2.0 参数保持冻结。输入声学特征为 80 维滤波器组系数，帧长为 25ms，帧移为 10ms。将特征输入模型前，会进行均值和方差归一化。通过添加背景噪声或卷积混响噪声进行在线数据增强 [29] 。MUSAN [30] 和 RIR Noise [31] 数据集分别用作噪声源和房间脉冲响应函数。使用权重衰减为 1e-4 的 Adam W [32] 优化器，并采用每 5 个 epoch 衰减一次的 StepLR 调度器。前 5 个 epoch 采用线性预热调度以稳定训练，随后使用衰减因子为 0.1 的 StepLR 调度器，将学习率从 1e-4 降至 1e-5。采用 ArcFace [33] 损失作为说话人分类目标，边距和缩放系数分别设为 0.2 和 32。输入帧长在 200 至 300 帧之间随机采样。

收敛后，解除对 w2v-BERT 2.0 参数的冻结，以进行进一步微调。在此阶段，学习率从 1e-5 开始，采用余弦衰减策略在 2 个周期内逐渐降至 5e-6，微调共进行 4 个周期。

3.3.2. 在 *TidyVoiceX* 训练集上进行语言无关学习微调

在此阶段，将 TidyVoiceX 训练集纳入训练流程以进一步进行领域自适应，并引入语言对抗学习策略，以促进语言不变的说话人表示学习。具体而言，在说话人嵌入上方新增了一个由两个线性层组成的语言分类器。为确保稳定的语言监督，首先冻结所有其他模块，仅训练语言分类器直至收敛。随后，解冻说话人编码器及其他模块，并在说话人嵌入与语言分类器之间引入 GRL。在反向传播过程中，GRL 会反转来自语言分类目标的梯度，迫使编码器抑制语言相关信息，同时保留具有说话人区分性的特征。系数 λ_{GRL} 和 λ_{lang} 均设为 0.1，以控制对抗学习的强度。模型采用余弦衰减学习率调度，在 2 个训练轮次内将学习率从 1e-5 逐渐降至 5e-6，总训练轮次为 4。数据增强策略与前一阶段相同。由于 ArcFace 损失应用广泛，因此将其作为默认损失函数。作为比较，我们还在此阶段使用 SphereFace2 (SF2) [34] 损失函数进行了若干实验。

3.4. 分数校准

为进一步提高分数的可靠性，我们采用质量度量函数（QMF） [35] 来校准说话人验证分数。QMF 旨在补偿由语音时长、信号质量和嵌入可靠性差异造成的分数变化。QMF 模型使用从 TidyVoiceX 训练集随机生成的试验进行训练。如 [36],所述，我们采用以下 QMF 质量集合 q 来校准分数：

$$\mathbf{q} = [\log(d_e), \log(d_t), \|e\|, \|t\|, \text{SNR}_e, \text{SNR}_t, s] \quad (7)$$

其中，$\log(d_e)$ 和 $\log(d_t)$ 分别表示注册语音和测试语音时长的对数，$\|e\|$ 和 $\|t\|$ 分别表示注册嵌入和测试嵌入的幅度，SNR_e 和 SNR_t 分别表示注册语音和测试语音的信噪比，s 表示验证分数。只有 SNR_e 和 SNR_t 采用最大—最小归一化。

采用逻辑回归在生成的试验样本上训练 QMF 模型：

$$\mathbf{s}' = \sigma(\mathbf{w}^\top \mathbf{q} + b) \quad (8)$$

其中 w 和 b 是学习得到的参数，s' 是校准后的分数。

推理时，使用训练好的 QMF 模型生成最终的验证分数。

4. 结果

表 1 展示了基于 w2v-BERT 2.0 的说话人验证系统在 TidyVoice 2026年开发集和评估集上的性能。与官方基线系统 SimAM-ResNet34 [22, 37], 相比，微调预训练 w2v-BERT 2.0 模型的说话人验证系统表现出明显优势；SimAM-ResNet34 [22, 37], 同样是在大量数据上预训练的。即使不使用 TidyVoiceX 训练集，仅使用预训练数据进行微调的模型，在 tv26 开发集上的等错误率也达到 2.74%，相较于基线值 3.07%，等错误率相对降低了 11%。

此外，我们还探究了不同损失函数和微调数据配置的效果。对于损失函数，除了广泛使用的 ArcFace 损失外，我们还评估了采用 A 和 C 配置的 SphereFace2 损失。结果表明，使用 SphereFace2 进行微调的模型显著优于使用 ArcFace 的模型，其中 SphereFace2-C 的表现最佳。这主要归因于

表 1： *w2v-BERT 2.0* 说话人验证系统在 *TidyVoice 2026*年开发集和评估集上的性能。*w2v- BERT 2.0 Based*、*w2v-BERT 2.0 Based*$_{SF2-A}$ 和 *w2v-BERT 2.0 Based*$_{SF2-C}$ 分别表示使用默认 *ArcFace* 损失、*SphereFace2-A* 损失和 *SphereFace2-C* 损失训练的模型。

模型	预训练数据	微调数据	tv26 开发集		tv26 eval-A		tv26 eval-U	
			等错误率(%)	最小检测代价函数$_{0.01}$	等错误率(%)	最小检测代价函数$_{0.01}$	等错误率(%)	最小检测代价函数$_{0.01}$
官方基线 [22]	VoxBlink2 +VoxCeleb2	TidyVoiceX 训练	3.07	0.82	9.058	0.65	11.59	0.60
基于 w2v-BERT 2.0	VoxBlink2 + VoxCeleb2 + 3D-Speaker + KeSpeech + CnCeleb1&2	None	2.740	0.79	-	-	-	-
基于 w2v-BERT 2.0		预训练数据 + TidyVoiceX 训练	1.466	0.66	-	-	-	-
基于 w2v-BERT 2.0$_{SF2-A}$			1.089	0.63	-	-	-	-
基于 w2v-BERT 2.0$_{SF2-C}$			1.065	0.62	3.061	0.24	**4.338**	**0.29**
基于 w2v-BERT 2.0		TidyVoiceX 训练	1.191	0.63	-	-	-	-
基于 w2v-BERT 2.0$_{SF2-A}$			0.966	0.61	-	-	-	-
基于 w2v-BERT 2.0$_{SF2-C}$			0.950	0.61	-	-	-	-
+GRL			0.937	0.60	2.964	0.23	5.020	0.30
++QMF			**0.893**	**0.60**	**2.458**	**0.21**	4.451	0.29
基于 w2v-BERT 2.0$_{SF2-C}$		全部合成数据	1.022	0.60	-	-	-	-
基于 w2v-BERT 2.0$_{SF2-C}$		TidyVoiceX 训练 + 全部合成数据	0.999	0.61	-	-	-	-
基于 w2v-BERT 2.0$_{SF2-C}$		TidyVoiceX 训练 + 子合成数据	0.954	0.61	-	-	-	-

与将说话人验证建模为多类分类问题的 ArcFace 不同，SphereFace2 在超球空间中采用二元分类目标。由于说话人验证中的训练和评估都依赖成对相似度比较，这种建模方式有效缩小了训练目标与评估协议之间的差异。

关于微调数据策略，我们比较了仅使用 TidyVoiceX 训练集与将其和大规模预训练数据集混合使用的效果。结果表明，仅使用 TidyVoiceX 训练集进行微调，在 tv26 开发集和 tv26 eval-A 上表现更好，因为这些语言在训练期间出现过；但在包含未见语言的 tv26 eval-U 上表现较差。这主要是领域专门化与泛化之间的差异所致。仅使用 TidyVoiceX 训练集进行微调，可以让模型更加专注于目标领域，因此在语言和声学条件与训练期间所见情况相似的 tv26 开发集和 tv26 eval-A 上表现更好。相比之下，加入大规模多语言预训练数据可以让模型接触更多样的语言和声学变化，从而提升其泛化能力。此外，预训练数据可能包含与 tv26 eval-U 中语言重叠的语言。

图 3 展示了说话人 $id011337$ 的真实语音嵌入与合成语音嵌入的 t-SNE 可视化。合成嵌入与真实嵌入高度一致，表明 Qwen3-TTS 有效保留了说话人身份。特别是，与真实语音语言相同的合成嵌入（以红色标出）非常靠近相应的真实嵌入，而来自其他语言的合成语音也形成了聚类良好且分离清晰的嵌入。然而，如表 1 所示，用合成语音增强训练数据并未带来进一步的性能提升。有趣的是，仅使用合成数据进行训练在 tv26 开发集上取得了 1.022% 的等错误率，接近使用真实数据获得的 0.95% 等错误率。值得注意的是，这些合成样本仅使用约十分之一的真实数据作为参考音频生成。这表明，在训练数据充足的条件下，合成数据与真实数据之间的领域失配可能会降低性能。相比之下，在低资源场景中，合成数据增强可以成为提升跨语言说话人验证的可行策略。

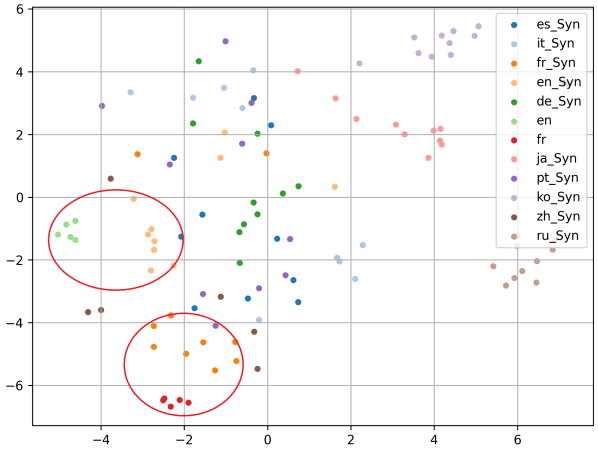


图 3： 说话人 $id011337$ 的真实与合成语音嵌入的 *t-SNE* 可视化。

5. 结论

本文介绍了我们为 TidyVoice2026 挑战赛开发的说话人验证系统。在本工作中，我们提出了一种基于大规模自监督 w2v-BERT 2.0 模型的多语言说话人验证系统，利用层适配器和 MFA 框架提取稳健的说话人嵌入。为提升跨语言泛化能力，我们引入了使用 GRL 的语言对抗训练策略，以促进语言不变表示的学习。此外，我们采用多语言零样本文本转语音系统 Qwen3-TTS，为每位说话人合成额外语音，从而增强训练数据中的语言多样性。在 TidyVoice 2026 开发集和评测集上的实验结果表明，在大规模预训练模型上进行微调可显著提升说话人验证性能。此外，使用 SphereFace2 损失比 ArcFace 损失取得了更好的结果，而语言对抗训练在抑制语言特定变化方面带来了适度改进，合成数据增强在数据有限的条件下也是有效的。

6. 生成式人工智能使用披露

大语言模型（LLM）仅用于语言编辑，包括改写和语法润色，以提高清晰度和可读性。大语言模型未参与想法的形成、方法设计、实验流程、数据分析或结果解释。所有科学内容均由作者开发并核实。

7. 参考文献

[1] W. Cai、J. Chen 和 M. Li，《探索端到端说话人与语言识别系统中的编码层和损失函数》，载于 Odyssey 论文集，2018年，第 74–81 页。[2] B. Desplanques、J. Thienpondt 和 K. Demuynck，《Ecapa-tdnn：基于 TDNN 的说话人验证中的强调通道注意力、传播与聚合》，载于 Interspeech 论文集，2020年，第 3830–3834 页。[3] H. Wang、S. Zheng、Y. Chen、L. Cheng 和 Q. Chen，《Cam++：一种利用上下文感知掩码进行说话人验证的快速高效网络》，载于 Interspeech 论文集，2023年，第 5301–5305 页。[4] I. Yakovlev、R. Makarov、A. Balykin 等人，《用于说话人识别的 Reshape Dimensions Network》，载于 Interspeech 论文集，2024年，第 3235–3239 页。[5] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao 等人，《Wavlm：面向全栈语音处理的大规模自监督预训练》，IEEE信号处理选题期刊，第 16 卷，第 6 期，第 1505–1518 页，2022年。[6] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，《wav2vec 2.0：语音表征自监督学习框架》，载于 NeurIPS 论文集，2020年，第 12 449–12 460 页。[7] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai 等人，《Hubert：通过对隐藏单元进行掩码预测实现自监督语音表征学习》，IEEE/ACM音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021年。[8] L. Barrault、Y.-A. Chung、M. C. Meglioli 等人，《Seamless：多语言表现力流式语音翻译》，arXiv 预印本 arXiv:2312.05187，2023年。[9] Z. Li、M. Cheng 和 M. Li，《结合 w2v-BERT 2.0 与知识蒸馏引导的结构化剪枝增强说话人验证》，载于 ICASSP 论文集，2026年。[10] D. Cai 和 M. Li，《通过迁移学习和知识蒸馏利用预训练自动语音识别 Conformer 进行说话人验证》，IEEE/ACM音频、语音与语言处理汇刊，第 32 卷，第 3532–3545 页，2024年。[11] Y. Zhang、Z. Lv、H. Wu、S. Zhang、P. Hu、Z. Wu 等人，《MFA-Conformer：用于自动说话人验证的多尺度特征聚合 Conformer》，载于 Interspeech 论文集，2022年，第 306–310 页。[12] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang 等人，《Lora：大语言模型的低秩适应》，载于 ICLR 论文集，2022年。[13] X. Qin、N. Li、W. Chao、D. Su 和 M. Li，《跨年龄说话人验证：学习年龄不变说话人嵌入》，载于 Interspeech 论文集，2022年，第 1436–1440 页。[14] Z. Du、C. Gao、Y. Wang、F. Yu、T. Zhao、H. Wang、X. Lv、H. Wang、C. Ni、X. Shi 等人，《Cosyvoice 3：通过扩展和后训练实现面向真实场景的语音生成》，arXiv 预印本 arXiv:2505.17589，2025年。[15] H. Hu、X. Zhu、T. He、D. Guo、B. Zhang、X. Wang、Z. Guo、Z. Jiang、H. Hao、Z. Guo 等人，《Qwen3-tts 技术报告》，arXiv 预印本 arXiv:2601.15621，2026年。[16] Y.-A. Chung、Y. Zhang、W. Han 等人，《W2v-bert：结合对比学习与掩码语言建模进行自监督语音预训练》，载于 ASRU 论文集，2021年，第 244–250 页。[17] K. Okabe、T. Koshinaka 和 K. Shinoda，《用于深度说话人嵌入的注意力统计池化》，载于 Interspeech 论文集，2018年，第 2252–2256 页。[18] Y. Ganin 和 V. Lempitsky，《通过反向传播实现无监督领域自适应》，载于 ICML 论文集，2015年，第 1180–1189 页。[19] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen 和 Y. Wu，《LibriTTS：源自 LibriSpeech 的语音合成语料库》，载于 Interspeech 论文集，2019年，第 1526–1530 页。[20] A. Fan、S. Bhosale、H. Schwenk、Z. Ma、A. El-Kishky、S. Goyal、M. Baines、O. Celebi、G. Wenzek、V. Chaudhary 等人，《超越英语中心的多语言机器翻译》，《机器学习研究期刊》，第 22 卷，第 107 期，第 1–48 页，2021年。[21] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，《通过大规模弱监督实现鲁棒语音识别》，载于 ICML 论文集，2023年，第 28 492–28 518 页。[22] A. Farhadipour、J. Marquenie、S. Madikeri、T. Vukovic、V. Dellwo、K. Reid、F. M. Tyers、I. Siegert 和 E. Chodroff，《Tidyvoice 2026 挑战赛评测计划》，arXiv 预印本 arXiv:2601 21960，2026年。[23] J. Chung、A. Nagrani 和 A. Zisserman，《Voxceleb2：深度说话人识别》，载于 I nterspeech 论文集，2018年。[24] Y. Lin、M. Cheng、F. Zhang 等人，《VoxBlink2：一个包含 100K+ 说话人识别语料库及开放集说话人识别基准》，载于 Interspeech 论文集，2024年，第 4263–4267 页。[25] S. Zheng、L. Cheng、Y. Chen、H. Wang 和 Q. Chen，《3d-speaker：用于语音表示解耦的大规模多设备、多距离、多方言语料库》，arXiv 预印本 arXiv:2306.15354，2023年。[26] Z. Tang、D. Wang、Y. Xu、J. Sun、X. Lei、S. Zhao、C. Wen、X. Tan、C. Xie、S. Zhou 等人，《KeSpeech：一个开放源代码的普通话及其八种次方言语音数据集》，载于第三十五届神经信息处理系统大会数据集与基准赛道（第二轮），2021年。[27] Y. Fan、J. Kang、L. Li 等人，《Cn-celeb：一个具有挑战性的中文说话人识别数据集》，载于 ICASSP 论文集，2020年，第 7604–7608 页。[28] L. Li、R. Liu、J. Kang 等人，《Cn-celeb：多体裁说话人识别》，《语音通信》，第 137 卷，第 77–91 页，2022年。[29] W. Cai、J. Chen、J. Zhang 和 M. Li，《用于说话人与语言识别的即时数据加载器和语句级聚合》，IEEE/ACM音频、语音与语言处理汇刊，第 1038–1051 页，2020年。[30] 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。[31] T. Ko、V. Peddinti、D. Povey、M. Seltzer 和 S. Khudanpur，《面向鲁棒语音识别的混响语音数据增强研究》，载于 ICASSP 论文集，2017年，第 5220–5224 页。[32] I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，载于 ICLR 论文集，2019年。[33] 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》，载于 CVPR 论文集，2019年，第 4690–4699 页。[34] B. Han、Z. Chen 和 Y. Qian，《探索用于说话人验证的二分类损失》，载于 IEEE ICASSP 论文集，2023年，第 1–5 页。[35] J. Thienpondt、B. Desplanques 和 K. Demuynck，《idlab VoxSRC-20 提交方案：基于大间隔微调和质量感知分数校准的 DNN 说话人验证》，载于 IEEE ICASSP 论文集，2021年，第 5814–5818 页。[36] Z. Li、Y. Lin、X. Qin、N. Jiang、G. Zhao 和 M. Li，《DKU-MSXF 说话人验证系统：用于 VoxCeleb说话人识别挑战赛 2023》，arXiv 预印本 arXiv:2308.08766，2023年。[37] X. Qin、N. Li、C. Weng、D. Su 和 M. Li，《基于迭代噪声标签检测的简易注意力模块说话人验证》，载于 IEEE ICASSP 论文集，2022年，第 6722–6726 页。