利用文本转语音数据增强改进语码转换语音识别

Yue Heng Yeo*†、Yuchen Hu†、Shreyas Gopal†、Yizhou Peng†、Hexin Liu†和Eng Siong Chng†
* 资讯通信研究院（I²R），新加坡科技研究局，新加坡
† 南洋理工大学计算与数据科学学院，新加坡 电子邮箱：yueheng001@ntu.edu.sg

arXiv:2601.00935v1 [eess.AS] 2 Jan 2026

摘要—**会话语码转换语音的自动语音识别（ASR）仍面临挑战，原因在于缺乏真实且高质量的标注语音数据。本文探讨多语种文本转语音（TTS）模型作为一种有效的数据增强技术，以解决这一数据短缺问题。具体而言，我们在 SEAME 数据集上微调多语言 CosyVoice2 TTS 模型，生成合成会话式中英语码转换语音，显著增加可用训练数据的数量和说话人多样性。实验表明，使用合成语音增强真实语音，可将 DevMan 上的混合错误率（MER）从 12.1% 降至 10.1%，并将 DevSGE 上的混合错误率从 17.8% 降至 16.0%，表明性能有所提升。这些结果证实，多语种 TTS 是提升低资源会话语码转换场景中自动语音识别鲁棒性的有效实用工具。**

I. 引言

语码转换是多语言使用者在日常交流中的常见做法，他们会在单句内或句与句之间将两种或更多语言混合在同一段对话中，并选择最符合其交际意图的词语或语法结构 [1], [2], [3]。在自动语音识别中，语码转换尤其具有挑战性，因为说话人在切换语言时往往会调整语调、节奏和发音，要求自动语音识别系统实时追踪这些变化 [4], [5], [6], [7]。尽管已有相关进展 [8], [9], [10], [11], [12], ，真实且准确转录的语码转换数据集仍然匮乏，这是语码转换自动语音识别面临的一大障碍，严重限制了模型性能。

解决语码转换数据短缺的一种常见方法是音频拼接 [13]。这种技术将不同单语录音中的音频片段拼接起来，合成语码转换语音，无需额外的数据收集即可创建合成双语话语。实证研究表明，使用音频拼接数据训练的自动语音识别系统能够降低错误率并减少单语偏差。然而，拼接音频片段通常会导致韵律不自然和明显的声学不连续性，并产生可能造成模型过拟合的协同发音伪影。因此，尽管音频拼接数据适用于初步实验，但与先进TTS模型生成的合成语音相比，它在真实感和语言覆盖范围方面存在固有局限，尤其是在 SEAME 等对话数据集上 [10], [14]。

另一种可能的方法是采用TTS数据增强。尽管早期的TTS模型因难以模拟自然韵律、说话人变异性和复杂的语言切换模式，而被认为不适用于代码切换自动语音识别 [10]–[15], [16], ，Chou等人近期的研究 [17] 表明，先进TTS模型生成的合成语音能够显著提升自动语音识别性能。他们利用TTS合成数据构建的自优化框架，显著降低了自动语音识别任务中的错误率，凸显了TTS数据增强在语码转换场景下微调自动语音识别系统的实际效果。

研究确定，成功进行数据增强的关键因素包括充足的文本多样性、适度的说话人变化，以及真实语音和合成语音之间的适当平衡 [8], [18]。利用此类通用文本转语音模型生成合成数据是一种有效且经济高效的解决方案，因为它仅依赖于易于抓取的文本和自动生成的语音嵌入来增强真实世界的语码转换数据集，从而绕过说话人招募、录音棚录音和人工语码转换转写等成本高昂的环节，直接应对数据稀缺问题 [10], [18]。近期研究表明，使用先进的多语种文本转语音模型生成的合成语音训练自动语音识别系统时，性能可获得显著提升，并大幅缩小与真实数据相比的性能差距 [18]–[19]。

II. 相关工作

近期自动语音识别领域的进展越来越多地利用文本转语音系统生成的合成语音，以缓解低资源多语言场景中的数据稀缺挑战 [8], [15]–[20]。例如，Yang 等 [18],通过在多种低资源领域中使用多语言 CosyVoice-base 文本转语音模型，展示了自动语音识别性能的显著提升，这些领域包括带口音语音、少数民族语言（韩语、汉语方言）以及专业词汇（例如汽车领域热词）。他们的结果强调了充足的文本多样性和适度的说话人多样性对于有效开展 TTS数据增强的重要性。

然而，尽管合成文本转语音语音增强已被证明在各种语言环境中行之有效，但大多数既有研究都忽视了会话语码转换，而这会带来独特的挑战，例如句内快速语言

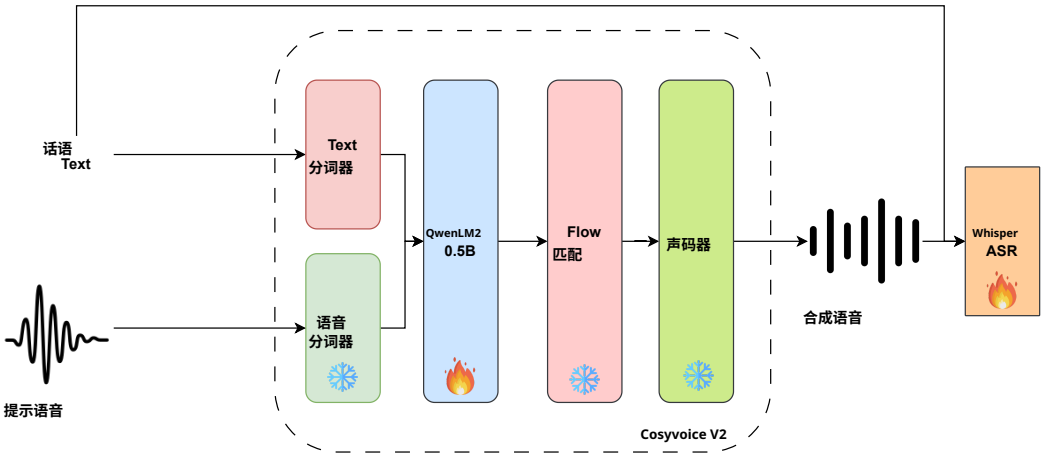


图 1. 端到端合成数据流水线。真实文本和语音经过分词后，输入 Qwen-2 语言模型、流匹配解码器和 HiFT声码器，生成合成音频，随后用于微调 Whisper。

切换、非正式词汇使用以及复杂的韵律模式 [2], [10]。常见的既有方法（例如通过拼接单语音频片段来构成合成双语话语的音频拼接增强）取得了有限的改进，但往往会产生不自然的韵律和声学伪影，从而限制了其在真实会话场景中的有效性 [8], [19]。

为解决这些局限性，我们的研究明确探讨了使用现代多语言 CosyVoice 文本转语音模型，并针对 SEAME 等中英语码转换语料进行微调 [9]。通过强调真实的会话结构、自发的韵律变化以及多样化的说话人特征，我们旨在大幅提升自动语音识别系统在复杂语码转换会话场景中的鲁棒性，从而拓展现有的多语言数据增强框架。

A. 我们的贡献

本文的贡献总结如下：

- 我们证明，多语种文本转语音模型，尤其是基于语码转换数据集微调的 CosyVoice，能够有效捕捉真实的会话韵律、非正式词汇用法以及快速的句内语言切换。此外，我们还表明，文本转语音数据可用于微调语音基础模型。
- 我们确定了实现文本转语音数据增强的关键因素，包括增加说话人变化、增加数据量，以及在合成语音数据与真实语音数据之间实现最佳平衡。
- 我们通过将经过 SEAME 微调的 CosyVoice 模型成功迁移到另一个代码切换语料库（ASCEND），大幅缩小其与真实数据性能之间的差距，证实了我们的 TTS数据增强流程具有良好的适应性和有效性。

III. 方法

A. *CosyVoice* 文本转语音

CosyVoice [21] 是一种可扩展的多语言零样本文本转语音（TTS）系统，依赖从多语言自动语音识别编码器中提取的监督式语义词元。如图1所示，该架构由四个紧密耦合的模块组成。(1) 文本编码器：语言无关的BPE前端将输入句子转换为词元，并将其与语音时间线对齐。(2) 语音标记器：该模块基于自动语音识别编码器上的向量量化构建，将训练音频离散化为低速率语义代码。(3) 大型语言模型（LLM），即文本转语音系统的骨干：将TTS视为自回归序列生成任务，基于Transformer的LLM接收文本词元和语音词元组成的混合流，并根据输入词元以自回归方式预测下一个语音词元。(4) 条件流匹配解码器：生成的词元序列经过上采样后，被送入流匹配网络，转换为梅尔频谱图，再由轻量级声码器渲染为波形。

CosyVoice 分两个阶段进行训练：语音标记器使用约 20 万小时的中英对齐音频进行学习，而完整的文本转语音模型则使用额外的 16.7 万小时数据进行训练，涵盖中文、英语、日语和韩语。这种规模和多样性使 CosyVoice 能够生成自然、说话人一致的语音，包括流畅的句内语码转换，因此可作为实用的合成数据来源，用于增强多语言自动语音识别系统。

B. *Whisper* 自动语音识别

Whisper 是 OpenAI[1],推出的大规模 Transformer 架构自动语音识别（自动语音识别）模型，使用约 68 万小时的多语言音频进行训练。由于其训练数据涵盖了广泛的声学条件和口语语言，Whisper 往往

[1]模型，网址：https://github.com/openai/whisper。

擅长处理不同的说话人、口音和嘈杂录音。

然而，语码转换仍然是个难题。尽管 Whisper 的多语言方法通常能够分别处理多种语言，但当这些语言快速交替出现时，它可能难以应对，导致转录错误或语言标识错误。

C. 数据生成流程

我们的方法分为三个阶段，用于扩充参考语音数据，并在说话人多样性和数据量方面生成更多数据。

1) 文本转语音调优：我们首先针对 SEAME 领域对 CosyVoice 2 进行适配。我们只更新 QwenLM 语言模型组件，而语音分词器、流匹配解码器和声码器均保持不变。在微调过程中，使用 SEAME 文本令牌训练 QwenLM，使其自回归地预测语音令牌序列，从而让它掌握语料中快速的普通话–英语交替、非正式表达方式和会话韵律。这种单模块更新在计算方面开销较小，却足以引导文本转语音系统生成自然的语码转换语音，同时保留原始 CosyVoice 技术栈的声学保真度。

2) 合成语音生成：适配后，使用从大型说话人库中采样的不同 x-vector 说话人嵌入，对每条 SEAME 转录文本进行多次重新合成。由此得到的合成语料库具有丰富多样的说话人特征，在保留原始文本的同时，也丰富了音色、音高范围和语速变化。

3) 自动语音识别调优：将合成语音与 100 小时 SEAME 真实标注音频混合，用于微调 Whisper-small 模型。我们比较三种条件：(i) 仅使用真实标注数据，(ii) 真实标注数据 +文本转语音（所提出的混合方案），以及 (iii) 仅文本转语音。保持 Whisper 架构和数据增强方案不变，即可单独评估额外的多说话人合成数据对语码转换识别准确率的影响。

IV. 实验设置

A. SEAME 数据集

SEAME（东南亚普通话-英语）语料库是一个语音数据集，专门用于捕捉新加坡和马来西亚双语说话人之间自发的会话语码转换。该语料库包含约 192 小时的自然对话和访谈音频，参与者共 156 名说话人。对话涵盖日常话题，展现了语言之间频繁切换的现象，甚至常常在单个句子或短语中切换。每段话语都经过仔细转录，并以清晰标签标明语言边界，因此 SEAME 非常适合训练和评估必须应对真实世界双语互动的自动语音识别系统。由于其自发性、真实的语言混用情况以及详细的标注，SEAME 被广泛用作语码转换研究与开发的标准基准。

表 I 不同真实语音、原说话人文本转语音（TTS-O）和随机说话人文本转语音（TTS-R）混合方案下，Whisper-Largev3 在 DevMan 和 DevSGE 上的混合错误率（MER）。粗体标示各测试集中的最低 MER。

模型	时长（小时）			MER（%）	
	Real	TTS-O	TTS-R	DevMan	DevSGE
Whisper-Largev3	100	-	-	12.1	17.8
	-	100	-	12.5	18.6
	-	-	100	17.7	22.4
	100	100	-	11.1	17.0
	100	-	100	**10.1**	**16.0**
	-	-	200	12.2	18.5

B. 模型

a) CosyVoice 微调：我们将 CosyVoice 2 的 QWENLM2（0.5 B 参数）模型适配至目标域。优化采用 Adam，初始学习率为1×10^{-4}。学习率在前 10 000 次更新（预热）期间线性增长，随后在剩余的 200 个训练轮次中保持不变。

b) Whisper 自动语音识别微调：在消融研究中，自动语音识别后端采用已发布的 Whisper-small 检查点作为起点（$\sim$240 M 个参数），并在 ESPnet [2] 中进行微调。输入波形被转换为 80 个频段的对数梅尔滤波器组（24 kHz，20 ms 窗长，12 ms 帧移）；我们采用相同的 SpecAugment 设置：两个频率掩码（宽度为 ≤ 40 个频段）、五个时间掩码（宽度为话语的 $\leq 12\%$）以及一个五帧的时间扭曲窗口。优化使用 AdamW（$\beta = 0.9/0.99$，$\epsilon = 1 \times 10^{-6}$，权重衰减为 0.01）。学习率遵循 ESPnet 的 warmuplr 调度：在前 1 500 次更新中从零线性升至1×10^{-5}，随后按平方根倒数规则衰减。小批次根据频谱图元素总数构建；每次更新限制为约 12 M 个元素，并将梯度累积四步。所有实验中的语言标识均设置为自动。

c) 评估：自动语音识别质量在 ESPNET 工具包的 SEAME 配置中，以 DEVMAN 和 DEVSGE 上的混合错误率（混合错误率）进行报告。

C. 实验结果

表 I 中的结果证实，除了现有真实录音外，加入合成语音后，高质量文本转语音是增强 Whisper-Largev3 数据的有效工具。仅使用 100 小时真实语音对 Whisper 进行微调的基线模型，在 DevMan 和 DevSGE 上的混合错误率分别为 12.1% 和 17.8%。使用 CosyVoice2 重新生成相同的话语，同时保留原始说话人嵌入（TTS-O），并将两组数据按一比一混合后，混合错误率降至 11.1% / 17.0%。当以随机采样的说话人嵌入替换原有嵌入（TTS-R）时，效果提升最为显著：将 100 小时真实语音与 100 小时随机说话人文本转语音合成数据结合后，混合错误率进一步降至 DevMan 上的 10.1% 和 DevSGE 上的 16.0%，优于真实标注

[2]https://github.com/espnet/espnet

表 II 使用由经不同数量目标域数据微调的 ATTS模型生成的合成语音进行训练时，WHISPER-SMALL 在 DEVMAN 和 DEVSGE 上的混合错误率（%）。UTMOS 表示相应合成语音集的平均意见分；粗体表示每列中的最佳值。

时长（小时）	DevMan	DevSGE	UTMOS
真实标注	**13.4**	**19.2**	**3.6**
10	21.8	26.1	2.9
50	15.2	23.2	3.1
100	13.8	20.1	3.2

基线模型在每个测试集上约高出两个百分点。结果表明，关键因素是说话人多样性；仅仅重复原有声音的合成音频，效果不如引入新音色和韵律的音频。同时，仅使用 200 小时随机说话人文本转语音数据进行训练，效果不如仅使用真实数据的模型（DevMan 上为 12.2%，DevSGE 上为 18.5%），这说明合成数据最适合作为补充，而非替代品。综上，这些发现表明，文本转语音能够生成“几乎真实”的训练样例，大幅提升识别准确率，前提是合成数据量至少达到真实数据量的两倍，并且带来新的说话人特征，而非重复已有特征。

D. 微调 *CosyVoice* 所需的数据量

为了确定需要多少数据才能微调 CosyVoice2，使其生成更多域内数据，我们运行了 UTMOS22 [22], ，这是一种开源模型，可预测从 1（差）到 5（优秀）的平均意见分（MOS）。真实录音（真实标注）的得分为 3.6。当 CosyVoice 使用 10 小时的目标语音进行微调时，MOS 达到 2.9。将微调数据扩展到 50 小时后，得分升至 3.1；使用全部 100 小时数据时，得分则小幅升至 3.2。MOS 曲线仅略有上升，因为 CosyVoice 已经通过数百小时的多说话人数据进行了预训练。大规模预训练使模型掌握了大部分清晰的发音、平滑的音高和较低的噪声，因此，即使是使用 10 小时数据微调的版本，起点也相当不错。额外的域内时长主要帮助文本转语音复制我们语料库的对话风格和节奏，而大型 MOS 模型对此给予的奖励有限。相比之下，自动语音识别指标的变化要显著得多：从 10 小时增至 100 小时后，混合错误率绝对值降低了约 9 个百分点。因此，尽管 MOS 的提升不大，规模更大的微调数据集仍然很有价值，因为它们能让合成语音更接近多轮对话语码转换语音的目标域，从而以切实的方式提升 Whisper-small 的识别准确率。 [3]

E. 微调 *Whisper* 所需的合成数据量

表III显示，将合成数据集扩大到超过 100 小时真实标注基线后，混合错误率（MER）会持续下降，但每增加一批数据所带来的边际收益都在缩小。将训练时长翻倍至 200 小时，带来了

[3]https://github.com/sarulab-speech/UTMOS22

表III Whisper-small 使用不断增加的合成语音数据进行微调时，在 DevMan 和 DevSGE 上的混合错误率（%）。最后一列显示每增加 100 小时数据块所带来的混合错误率相对降幅（对两个测试集取平均）。

合成数据（小时）	DevMan	DevSGE	相对增益（%）
真实标注	13.4	19.2	–
100	19.0	23.7	–
200	13.3	19.9	**23.04**
300	11.7	18.2	10.29
400	11.2	17.5	4.06
500	**10.9**	**17.0**	2.54

表 IV 使用两种数据增强技术微调 300 小时后 Whisper-small 的 MER（%）（越低越好；最佳分数以粗体标示）。

技术	DevMan	DevSGE
真实标注 + 速度扰动	13.4	19.2
真实标注 + 文本转语音	**12.6**	**18.7**

最显著的收益：与基线相比，混合错误率（MER）下降了约四分之一，这使这一步成为整个系列中幅度最大的一次提升。进一步增加到 300 小时仍有帮助，但额外收益仅约为上一次增量的一半。此后，曲线迅速趋于平缓。400 小时和 500 小时的设置分别只降低了 0.4 和 0.3 个百分点，却都需要承担再进行一轮 100 小时合成数据生成的全部成本。这些结果表明，使用 Whisper-small 进行微调的理想数据量是真实数据量的两到三倍：这样既能获得大部分性能提升，又不会陷入收益迅速递减的区间。

F. 数据增强比较

另一种常用于训练自动语音识别系统的语音增强技术是速度扰动。速度扰动是一种广泛使用的音频增强方法，通过以一组固定的小幅缩放系数对每段波形进行重采样，从而合成地改变语音速度，这些系数通常为 0.9 ×（较慢）、1.0 ×（原速）和 1.1 ×（较快）。由于该操作在基本保持频谱包络不变的同时拉伸或压缩时间轴，因此既保留了说话人音色和语言内容，又引入了真实的语速变化。应用这三个系数，可有效将训练数据量扩大三倍，无需额外转录，并为声学模型提供更广泛的时间动态覆盖，从而增强模型在测试时应对语速不匹配的能力。

表 IV 对比了两种相似的数据增强微调技术在 *Whisper-small* 上的表现。与传统的三档速度扰动（0.9/1.0/1.1×）相比，使用 300 小时的 CosyVoice文本转语音数据增强，可将 DEVMAN 上的混合错误率从 13.4% 降至 12.6%，并将 DEVSGE 上的混合错误率从 19.2 % 降至 18.7%。两种对比均将 100 小时的真实标注

表 V 领域专用训练 ∼9 小时后，Whisper-small 在 ASCEND-Test 集上的 MER（%），比较真实标注微调与两种 CosyVoice 生成数据变体。粗体表示最佳（最低）MER。

Data	ASCEND-Test
真实标注	**17.8**
CosyVoice 零样本生成	25.2
CosyVoice（SEAME 微调）生成	19.1

语料库固定不变；因此，性能差异体现了文本转语音带来的说话人和韵律多样性的收益。速度扰动仅改变时间动态，同时保留单一说话人身份；而我们的文本转语音流程则注入数百个合成说话人嵌入，丰富用于微调 Whisper 的声学信息。因此，这表明与仅对数据进行速度扰动相比，说话人多样性至关重要。

G. 跨领域比较

ASCEND[4] 是一个完整转录的语料库，包含在香港采集的 9 小时自发中英语码转换对话。该语料约有 12,000 条话语，来自 38 位说话人（21 位女性、17 位男性），每条话语平均时长为 2.4 秒，采样率为 16 kHz。与 SEAME 中的对话风格数据类似，ASCEND 也体现了对话轮替。

为展示 CosyVoice 生成数据的可移植性，我们使用适配 SEAME 的 CosyVoice 为 ASCEND 合成语音。ASCEND 是一个独立的中英语码转换对话语料库，具有日常对话中常见的快速轮替和非正式词汇混用特点。我们将 ASCEND 的 9 小时文本转录稿输入经 SEAME 调优的文本转语音系统，并使用文本转语音重新生成完全相同的话语，以观察将适配 SEAME 的 CosyVoice 用于另一领域所产生的效果

如表 V 所示，用未经适配的 CosyVoice 模型（“零样本”）生成的合成语音替代收集到的 ASCEND 音频后，Whisper-small 的混合错误率从 17.8% 上升至 25.2%，相对下降了 42%。这凸显了自动语音识别性能在很大程度上取决于训练和测试领域之间韵律与风格的匹配程度。若先用 SEAME 语料库对同一文本转语音引擎进行微调——该语料库与 ASCEND 一样具有会话式语码转换特征——再用它重新生成 ASCEND 话语，混合错误率便会降至 19.1%。与 SEAME 对齐的合成数据消除了零样本条件所引入误差惩罚的四分之三，使混合错误率相对降低 24%，与真实标注基线的差距仅为 1.3%。这一结果表明，经过一轮风格适配，CosyVoice 就能生成几乎与真实录音同样有效的训练语音，为新的对话语码转换领域提供了一条经济高效的自动语音识别引导训练途径，且无需进一步收集数据。

[4]语料库及许可证：https://github.com/HLTCHKUST/ASCEND。

V. 结论与未来工作

我们研究了多语种文本转语音（TTS）能否作为一种可行的数据增强技术，以应对自动语音识别中会话语码转换数据有限这一挑战。结果表明，对现代多语言文本转语音模型进行微调以生成合成语音，能够有效捕捉真实语码转换对话中常见的多样化说话人、非正式词汇选择和自然韵律。我们的方法生成的合成数据显著提升了训练多样性和真实感，为在低资源会话场景中增强自动语音识别鲁棒性提供了一种实用且经济高效的方式。

然而，当前增强流程受限于现有转录文本本身的多样性不足。作为未来的工作，我们计划探索采用专为多语言和语码转换文本生成而设计或调整的大语言模型（LLMs），以提升文本的多样性。尽管目前尚无可靠且可控的语码转换文本生成器，但开发此类工具将有助于合成更丰富、更多样的训练样例，进一步提升多语种对话领域中的自动语音识别性能。

致谢

本文的计算资源由新加坡国家超级计算中心（https://www.nscc.sg）提供，并与资讯通信研究院（I2R，新加坡科技研究局）合作。

参考文献

[1] P. Auer，《对话中的语码转换：语言、互动与身份认同》。Routledge，1999 年。[2] B. E. Bullock 和 A. J. Toribio，《剑桥语言语码转换手册》。剑桥大学出版社，2009 年。[3] H. Liu 等人，“面向双语语码转换语音的端到端语言分离”，载于*Interspeech* 论文集，2021 年，第 1489–1493 页。[4] H. Li、B. Ma 和 K. A. Lee，“多语种环境下的口语处理”，《*IEEE* 信号处理杂志》，第 30 卷，第 3 期，第 82–91 页，2013 年。[5] N. T. Vu、P. K. Gupta、H. Adel、D. Telaar、H. Sch¨utze 和 T. Schultz，“口语理解系统的首遍解码”，载于*Interspeech* 论文集，2012 年，第 1624–1627 页。[6] H. Liu 等人，“通过将语音与语言对齐来增强语码转换语音识别”，《*arXiv* 预印本 *arXiv:2403.05887*》，2024 年。[7] H. Liu、L. P. Garcia、X. Zhang、A. W. H. Khong 和 S. Khudanpur，“利用交互式语言偏置增强语码转换语音识别”，载于*IEEE* 声学、语音与信号处理国际会议论文集，2024 年，第 10 886–10890 页。[8] S. Li、X. Chen 和 J. Yu，“语码转换自动语音识别的数据增强：当前挑战与未来方向”，载于*IEEE* 口语语言技术会议论文集，2021 年，第 528–535 页。

[9] D.-C. Lyu、T.-P. Tan、E. S. Chng 和 H. Li，“SEAME：东南亚普通话-英语代码切换语音语料库”，载于*Interspeech* 论文集，2015 年，第 1982–1986 页。[10] G. I. Winata、Z. Liu 和 P. Fung，“代码切换自动语音识别：进展与挑战”，《*IEEE* 音频、语音与语言处理汇刊》，第 30 卷，第 1–17 页，2022 年。[11] H. Liu、H. Xu、L. P. Garcia、A. W. H. Khong、Y. He 和 S. Khudanpur，“利用词元级语言说话人日志降低语码转换语音识别中的语言混淆”，载于*IEEE* 声学、语音与信号处理国际会议论文集，2023 年，第 1–5 页。[12] X. Zhang 等人，“语音中的 Mamba：探索自注意力的替代方案”，《*IEEE* 音频、语音与语言处理汇刊》，第 33 卷，第 1933–1948 页，2025 年。[13] A. Hussein、D. Zeinali、O. Klejch、M. Wiesner、B. Yan 等人，“语音拼贴：通过拼接单语语料库生成语码转换音频”，《*arXiv* 预印本》，第 2309.15674 卷，2023 年，已被 Interspeech 2024 接收。 [在线]。网址：https://arxiv.org/abs/2309.15674[14] T. Nguyen 和 H.-D. Tran，“没有真实语码转换数据，我们能否用语码转换数据训练自动语音识别系统？新加坡语言案例研究”，2025 年，已被 Interspeech 2025 接收。[15] A. Rosenberg、M. Sun 和 V. Ramanathan，“低资源语言的语音识别挑战”，《*IEEE/ACM* 音频、语音与语言处理汇刊》，第 27 卷，第 9 期，第 1531–1542 页，2019 年。[16] X. Zhang 等人，“在小波域中说话：一种加速语音扩散模型的简单高效方法”，载于*EMNLP* 论文集，2024 年，第 159–171 页。[17] C.-K. Chou 等人，“利用 TTS 合成数据增强自动语音识别的自优化框架”，《*arXiv* 预印本 *arXiv:2506.11130*》，2025 年。[18] G. Yang 等人，“通过通用文本转语音增强低资源自动语音识别：弥合数据鸿沟”，《*arXiv* 预印本 *arXiv:2410.16726*》，2024 年。 [在线]。网址：https :// arxiv.org/abs/2410.16726[19] W. Zhang、H. Sun 和 L. Li，“面向多语言和语码转换自动语音识别的合成语音增强策略”，载于*ICASSP* 论文集，2022 年，第 7292–7296 页。[20] D. S. Park 等人，“SpecAugment：一种简单的自动语音识别数据增强方法”，载于*Interspeech* 论文集，2019 年，第 2613–2617 页。[21] Z. Du 等人，“CosyVoice：一种基于监督语义词元、可扩展的多语言零样本文本转语音合成器”，《*arXiv* 预印本》，第 2403.12345 卷，2024 年。[在线]。网址：https://arxiv.org/abs/2403.12345[22] T. Saeki、D. Xin、W. Nakata、T. Koriyama、S. Takamichi 和 H. Saruwatari，“UTMOS：东京大学 UTokyo-Sarulab 的 VoiceMOS 挑战赛 2022 系统”，《*arXiv* 预印本 *arXiv:2204.02152*》，2022 年。