面向流式文本输入的基于大语言模型的韵律边界感知流式生成

Changsong Liu[1,*], *Tianrui Wang*[2,*], *Ye Ni*[3], *Yizhou Peng*[1], *EngSiongChng*[11] 新加坡南洋理工大学[2] 中国天津大学[3] 中国东南大学

changsong.liu@ntu.edu.sg, wangtianrui@tju.edu.cn, niye@seu.edu.cn, peng.yizhou@ntu.edu.sg, ASESChng@ntu.edu.sg

arXiv:2603.06444v1 [cs.SD] 6 Mar 2026

摘要

接收流式文本的流式文本转语音对于交互系统至关重要，但这种方案面临两大挑战：缺乏前瞻导致韵律不自然，以及无界上下文导致长篇生成崩溃。我们提出一种韵律边界感知后训练策略，使用弱时间对齐数据对预训练的基于大语言模型的文本转语音模型进行适配。具体而言，通过提供有限的未来文本，使模型学会在指定的内容边界处提前停止。推理时，滑动窗口提示会延续先前的文本和语音标记，从而确保上下文有限且拼接流畅。评估结果表明，在短篇和长篇场景中，我们的方法均优于 CosyVoice-Style 交错基线。尤其在长文本合成中，该方法使词错误率绝对降低 66.2%（从 71.0% 降至 4.8%），并使说话人相似度和情感相似度分别相对提升 16.1% 和 1.5%，为使用增量文本的流式文本转语音提供了可靠的解决方案。

索引词：流式文本转语音、基于大语言模型的文本转语音、增量文本输入

1. 引言

采用流式文本输入的流式文本转语音（文本转语音）旨在随着文本到达实时生成语音，这在对话系统和语音到语音翻译等应用中需求旺盛 [1, 2]。

此类系统的可用性在很大程度上取决于合成延迟，延迟主要来自两个方面：等待文本片段累积的时间，以及模型将文本转换为音频所需的推理时间。虽然可以通过优化模型规模或引入因果结构 [3, 4], 有效缓解后者，但要实现超低延迟，就必须将文本累积窗口控制得极小。这带来了首要挑战：自然、高质量的语音合成高度依赖充足的上下文信息。模型不仅需要历史文本来保持连贯性，还需要未来文本（前瞻）来准确预测重音和停顿等韵律特征 [5]。因此，受限的感受野会导致韵律不自然。现有方法 [1, 6, 7] 尝试通过局部上下文建模解决这一问题，但通常需要对注意力机制进行复杂的因果性修改，并依赖精确的文本—语音强制对齐 [8, 9]。

与此同时，随着大语言模型（LLM）架构在语音领域得到广泛应用，现代文本转语音系统日益采用交错文本与语音标记的跨模态建模方式 [10, 11, 12, 13, 14, 15]。尽管这类架构实现了最先进的合成质量，但要在流式文本输入下实现流式生成，还面临第二个挑战：生成历史不断增长，导致长篇性能崩溃。例如，Cosyvoice 系列 [16, 17] 采用固定数量的文本和语音标记交错排列。类似地，Qwen3-TTS [18] 的流式版本也采用了相近的交错式结构，不过是将文本和语音标记并行排列。然而，在长期连续输入的真实场景中，由于单个文本标记所对应的语音长度差异极大，文本与其关联语音标记之间的物理距离会逐渐增大。这最终会导致生成失败，使系统难以支持长期流式交互。Speakstream [19]通过字符级对齐和历史截断缓解了这一问题，但它高度依赖精确的对齐标注。这些局限引出了以下研究问题：**在不修改架构、仅使用弱时间对齐数据的情况下，能否在基于大语言模型的文本转语音中实现稳健的长篇流式生成，并支持流式文本输入?**

为解决这一问题，我们提出一种新颖的韵律边界感知适配后训练策略，仅使用弱时间对齐数据，即可使现有的基于大语言模型的文本转语音模型实现稳健的流式生成。我们的贡献如下:

- 我们提出一种结合窗口化前瞻机制的韵律边界感知适配方法，使模型能够预先处理后续文本，从而改善韵律，且无需进行复杂的因果性修改。
- 我们设计了一种声学提示方法，利用前一文本块的音频尾段，实现无缝拼接，并缓解长篇跨模态连续流式生成中的生成崩溃问题。
- 我们证明，仅使用弱时间对齐的开源数据，即可实现最先进的流式稳定性和鲁棒性，并在实时部署中显著优于现有的交错基线。音频演示可在此处查看：[1]

2. 方法论

2.1. 韵律边界标记符

为在保留韵律自然度的同时实现流式生成，我们通过韵律边界感知标记符 $\text{marker}_{\text{boundary}}$ 将输入构造为分叉序列，从而将声学生成范围与更广泛的韵律 con-

[*]这些作者对本项工作贡献相同。

[1]https://anonymous-demo-168.github.io/Prosodic-Boundary-Aware-Streaming-Text-TTS-demo/

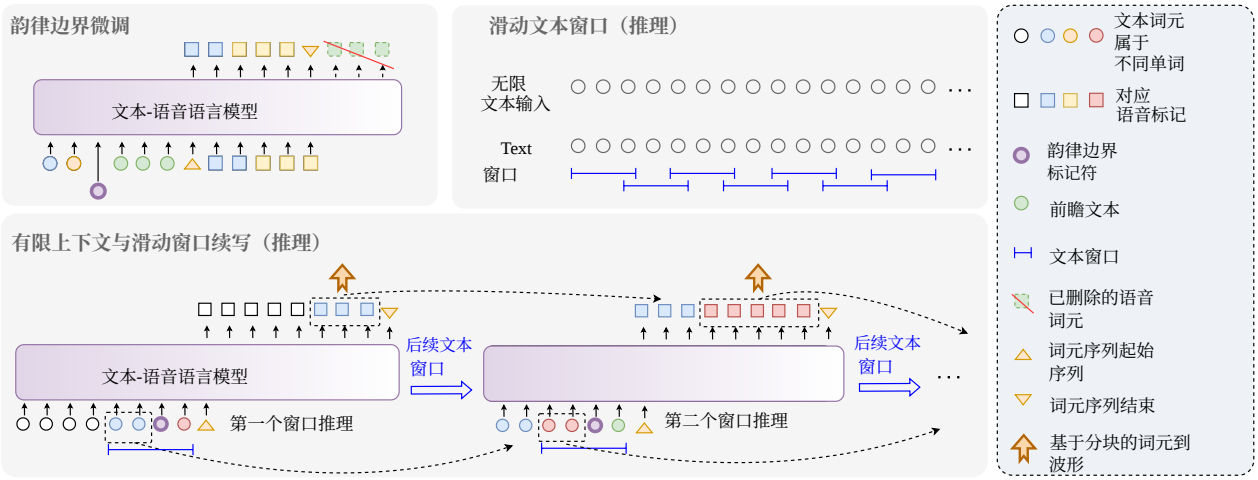

图1：所提　　出的微调与推理流程，用于韵律边界感知的流式基于大语言模型的文本转语音生成　　.

text 解耦。如图 1 所示，模型学会将该标记视为软边界。在推理过程中，每隔 k 个词插入一次标记符，使模型能够利用有限的未来上下文进行韵律规划，同时避免生成上下文无限扩展。

2.2. 基于弱时间对齐监督的训练

该模型利用现成对齐器WhisperX [20]，获得的词级时间戳进行适配，从而在无需人工标注的情况下近似候选韵律边界。给定一个包含文本词元 $x_{1:T}$、语音标记$s_{1:L}$和词级边界的语句（其中词 j 映射到文本词元跨度 $[b_j, e_j]$ 以及音频结束时间 a_j），我们应用**动态边界插入**。在训练期间，我们以概率 p_{full} 随机决定使用完整且未经修改的语句，以保持全局连贯性。否则，我们随机采样一个词索引 m，并将边界标记插入文本序列：

$$x' = (x_{1:e_m}, \texttt{marker}_{\texttt{boundary}}, x_{e_m+1:T}). \tag{1}$$

目标语音序列被截断到词 m的对齐音频位置。令 r_s 表示语音标记帧率， ℓmin表示最小长度。截断后的目标长度 L' 为：

$$L' = \max\big(\ell_{\min}, \lfloor a_m \cdot r_s \rfloor\big). \tag{2}$$

随后对语言模型进行微调，使其在以 x'为条件的情况下预测截断流 $s_{1:L'}$。此过程训练模型将标记符同时视为分段提示和韵律锚点，确保声学输出仅与标记符之前的片段对齐。

2.3. 有限上下文与滑动窗口续写

推理期间，输入文本按每块 k 个词进行处理，并前瞻 f 个未来词。对于块索引 t，令 $x_{t_{\text{cur}}}$ 表示当前片段， $x_{t_{\text{fut}}}$ 表示前瞻文本。输入序列构造如下：

$$X_t = \big(x_{t_{\text{cur}}}, \texttt{marker}_{\texttt{boundary}}, x_{t_{\text{fut}}}\big). \tag{3}$$

为保持跨块连续性，我们采用**滑动窗口提示**。第一个块以参考话语（x_{ref},(s_{ref})为条件，用于零样本语音克隆。对于后续块，提示（p_t^x, p_t^s）将替换为上一步合成的文本和语音标记：

$$(p_t^x,\, p_t^s) = \begin{cases} (x_{\text{ref}},\, s_{\text{ref}}), & t = 1, \\ (\hat{x}_{t-1},\, \hat{s}_{t-1}), & t \geq 2. \end{cases} \tag{4}$$

其中，$\hat{x}_{t-1}$ 表示与当前块对应的先前生成的文本片段（不包括前瞻文本），$\hat{s}_{t-1}$ 表示对应的合成语音。无论序列总长度如何，这种设计都将键值（KV）缓存限制在 $\mathcal{O}$（$k + f$），从而防止延迟增长和长篇内容不稳定。最后，将生成的语音标记传递给流式声码器，以增量方式进行波形合成，从而实现各块之间的无缝拼接。

3. 实验

3.1. 训练数据集

我们使用从 CommonVoice 13. 中采样的英语子集训练模型，该子集包含约 93 万条话语（约 ~1,000 小时）。为确保零样本评估的完整性，我们移除了所有与 Seed-TTS-Eval 测试集 [22] 重叠的训练话语。

训练前，我们会预先提取语音标记、梅尔频谱图、文本词元和说话人嵌入。我们使用 WhisperX [20] 获取词级时间戳，以近似确定候选韵律边界。微调期间，使用完整语句的概率设为 $p_{\text{full}} = 0.15$。语音分词器的帧率为 $r_s = 25$Hz，最小截断长度为 $\ell_{\min} = 5$ 帧。

表 1：流式处理效率评估。所提方法的块大小为 $k = 5$ 个词，前瞻为 $f = 2$ 个词。

系统	RTF↓	首令牌生成时间（毫秒）↓
交错式	0.843	<u>1414</u>
滑动窗口	**0.718**	2588
边界感知	<u>0.782</u>	**1296**

表2：客观质量评估：*Seed-TTS-Eval*。块大小 $k = 5$词；前瞻 $f =$ 2 词，用于所提方法。

系统	词错误率 (%) ↓		说话人相似度 ↑		情感相似度 ↑	
	标准	长篇	标准	长篇	标准	长篇
交错式	7.48	70.97	0.53	0.56	0.899	0.899
滑动窗口	6.03	7.83	0.57	0.22	0.912	0.857
边界感知（我们的方法）	**4.03**	**4.77**	**0.64**	**0.65**	**0.918**	**0.912**

表 3：*Seed-TTS-Eval* 上的主观平均意见分（*MOS*）评估。分数越高表示感知质量越好。

系统	平均意见分 ↑		语音自然度平均意见分 ↑		情感相似度评分 ↑	
	标准	长篇	标准	长篇	标准	长篇
交错式	3.99 ± 0.16	3.18 ± 0.23	4.05 ± 0.15	3.24 ± 0.22	4.03 ± 0.15	3.21 ± 0.18
滑动窗口	3.43 ± 0.20	1.60 ± 0.18	3.18 ± 0.20	1.68 ± 0.18	3.25 ± 0.18	1.67 ± 0.17
边界感知（我们的方法）	**4.28 ± 0.14**	**4.13 ± 0.13**	**4.25 ± 0.13**	**4.24 ± 0.13**	**4.38 ± 0.12**	**4.19 ± 0.13**

3.2. 评估数据集

我们从两个互补层级评估系统，以同时考察句子级质量和长篇场景下的鲁棒性。

标准形式评估：我们使用 Seed-TTS-Eval 基准测试集 [22], ，其中包含带有参考提示的短篇朗读式句子。该数据集用于评估标准句子级流式合成。

大语言模型扩展的长篇评估：为了对长篇稳定性进行压力测试，我们使用 DeepSeek-V3 [23]，将每个 Seed-TTS-Eval 句子扩展为一段连贯的 280–320 词段落，同时保留原句作为开头部分。该设置用于评估扩展独白过程中的韵律一致性和合成稳定性。

3.3. 评估指标

我们使用首音频时间（TTFA）和实时因子（RTF）来评估流式处理效率。TTFA 衡量初始合成请求与首个可解码音频块之间的延迟，反映用户感知到的响应速度。RTF < 1.0 表示系统具备实时合成能力。为避免硬件预热的影响，我们在每条语句进行两次预热运行后，对 50 次试验的结果取平均值。延迟指标在单个 NVIDIA A40 GPU（48 GB 显存）上测量。

在合成质量评估方面，我们遵循 Seed-TTS-Eval 协议 [22]。使用 Parakeet-TDT-0.6B-v2 模型 [24]，测量语言准确度，指标为词错误率（WER）；该模型支持超出通常 30 秒限制的长篇转录。使用 WavLM-Large 计算话语嵌入的余弦相似度，以评估声学保真度：说话人相似度（SPK-SIM）和基于 emotion2vec+ [25] 的情感相似度（EMO-SIM）。此外，我们还进行了主观听测，由 20 名评估者采用 5 分制李克特量表，对生成语音的可懂度（MOS）、说话人相似度（S MOS）和情感相似度（EMOS）进行评分。相似度评分通过将生成样本与参考提示进行比较得出。所有主观评分均报告为平均分，并附上通过标准 t 检验计算的 95% 置信区间。

3.4. 实验设置与基线

3.4.1. 基础架构

我们采用 CosyVoice2 [16] 作为基础模型。该架构由一个基于 Qwen 的大语言模型组成，用于生成语音标记；其后接流匹配模块和预训练的 HiFi-GAN 声码器 [26]，用于波形合成。训练期间，仅使用所提方法（第 2 节）对大语言模型进行微调；流匹配模块和声码器保持冻结。

3.4.2. 基线与所提方法

本节介绍基线方法。所有系统在流式输入时均采用固定的 $k = 5$ 词块大小。

交错基线：我们采用 CosyVoice2 的原生流式实现（推理双流模式），在单个序列中按 5:15 的比例交错排列文本和语音标记。KV 缓存会在生成过程中不断增长，并使用流式声码进行增量波形生成。我们使用未经边界感知适配的原始预训练权重。

滑动窗口基线：我们实现了一种简化的滑动窗口提示策略，其中每个块都以先前块生成的标记为条件。与我们的方法不同，该策略不使用边界标记或前瞻上下文，因此生成完全依赖历史信息。我们采用离线批量声码来合成每个块。

所提方法（边界感知）：我们的方法将滑动窗口续写与所提韵律边界标记以及长度为 $f = 2$ 个词的前瞻窗口相结合。我们采用流式声码进行增量音频合成。通过纳入有限的未来上下文，同时将有效上下文长度限制在 $\mathcal{O}(k + f)$，系统能够实现稳定的长篇流式生成。

4. 结果与讨论

4.1. 流式延迟与效率

流式延迟与效率结果汇总于表 1。所提方法的首令牌生成时间最低（1296 毫秒），优于交错基线和滑动窗口基线。这一改进源于边界感知提示机制，它能够比原生流式实现更早地输出音频。对于实时因子，滑动

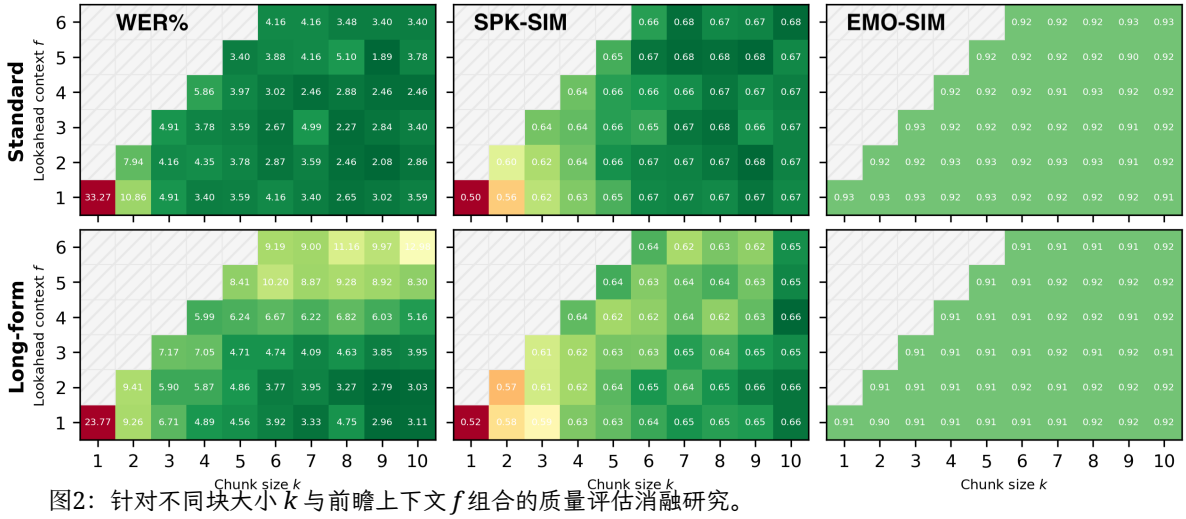

图2：针对不同块大小 k 与前瞻上下文 f 组合的质量评估消融研究。

窗口基线的数值最低（0.718）。然而，如第 3.4.2 节所述，这一结果归因于其采用离线批量声码，尽管这能最大限度地提高 GPU 吞吐量，却不支持增量输出。在采用流式声码的系统中，我们的方法的实时因子（0.782）低于交错基线（0.843）。这表明，通过限制上下文长度来防止 KV 缓存无限增长，可以提高计算效率。

4.2. 合成质量与语言保真度

表2和表3报告了两个评估层级的客观结果与主观评估结果。所提方法在所有指标上均持续取得最佳表现。

4.2.1. 客观评估

在**标准形式**集合上，我们的方法取得了 4.03% 的词错误率，优于交错式（7.48%）和滑动窗口（6.03%）。这表明，即使在短篇合成中，边界标记与前瞻上下文也能稳定语言生成。该模型还取得了最高的说话人相似度（0.64）和情感相似度（0.918），说明未来上下文提升了说话人一致性与情感表达。

性能差距在**长篇**评估中更为显著。交错基线出现灾难性失败，词错误率升至 70.97%。该行为源于无界上下文增长：随着 KV 缓存不断累积，模型出现语义漂移和幻觉，导致“含混不清”的语音，并因严重删除错误而过早终止。滑动窗口基线保持了稳定的词错误率（7.83%），但说话人相似度严重下降（说话人相似度从 0.57 降至 0.22）。在没有边界标记或前瞻上下文的情况下，模型无法正确预测句子边界，从而在各片段之间造成韵律漂移，并导致说话人一致性与情感一致性下降（情感相似度降至 0.857）。

相比之下，我们的边界感知方法在长篇合成中保持了稳定的性能，取得了 4.77% 的词错误率、0.65 的说话人相似度和 0.912 的情感相似度。通过限定上下文长度并引入前瞻窗口进行韵律条件控制，该方法有效缓解了交错架构中的幻觉问题以及朴素滑动窗口方法产生的韵律漂移。

4.2.2. 主观评估

主观听评结果证实了客观评估的发现。尽管交错基线在标准数据集中的质量尚可（3.99±0.16），但由于语言不稳定性，其在长篇场景中的表现骤降至 3.18±0.23。滑动窗口基线的性能进一步下降，平均意见分达到 1.60±0.18，因为听众经常察觉到片段之间存在韵律不连续。

我们的边界感知方法在所有指标上均取得了最高的感知评分。尤其是在长篇合成中，该方法仍能保持较强的说话人身份（4.24±0.13 语音自然度平均意见分）和情感一致性（4.19±0.13 情感相似度评分）。这些结果表明，所提出的边界感知条件能够在保持流式生成稳健性的同时，成功维持韵律连续性。

4.3. 消融研究

如图2所示，我们通过改变块大小 $k \in \{1, \ldots, 10\}$ 和前瞻上下文 $f \in \{1, \ldots, 6\}$，在约束条件 $f \leq k$下分析流式延迟与合成质量之间的权衡。评估时，我们从标准层级和长篇层级中各抽取固定数量的 50 条话语。

语言保真度（词错误率）对初始上下文大小高度敏感。在 $k = 1$，$f = 1$时，模型缺乏足够的语义依据，导致标准集和长篇集的词错误率分别达到峰值 33.27% 和 23.77%。随着 k 增大，性能迅速趋于稳定；对于 $k \geq 3$，标准层级的词错误率降至 5% 以下。长篇层级也呈现出类似趋势，这表明相对较小的语义锚点就足以稳定生成。

然而，相对于块大小而言，过度的前瞻会降低性能。例如，在 $k = 10$，将 f从 2 增加到 6，会使长篇词错误率从 3.03% 升至 12.98%。这表明，当未来条件过强并压过当前片段上下文时，可能会使生成过程不稳定。

说话人和情感一致性（SPK-SIM 和 EMO-SIM）更为稳健，但仍受益于适度的上下文。对于 $k \geq 5$，说话人相似度从上下文极少时约 0.50 提升至 0.65 以上，而情感相似度在各种配置下始终保持较高水平（> 0.90）。

5. 结论

本文提出了一种边界感知后训练策略，用于输入流式文本的流式大语言模型文本转语音。该方法通过引入韵律边界标记和有界滑动窗口提示，在防止上下文无限增长的同时，稳定韵律生成。在 Seed-TTS-Eval 上进行的实验表明，与交错式和滑动窗口基线相比，该方法提升了流式延迟和合成质量，同时能够稳定地进行长文本生成，并保持说话人身份和情感的一致性。

未来工作将探索向其他基于大语言模型的文本转语音架构及多语言场景的泛化，以及面向更灵活流式生成的自适应边界预测。

6. 生成式人工智能使用披露

在准备本稿件的过程中，作者使用了生成式人工智能工具，以协助语言润色、清晰度提升以及LaTeX格式排版。这些工具仅用于编辑支持。所有科学内容、实验结果和解释均由作者完成并核实。

7. 参考文献

[1] T. Dang、D. Aponte、D. Tran、T. Chen 和 K. Koishida，“连续文本流中的零样本文本转语音”，*arXiv* 预印本 *arXiv:2410.00767*，2024年。[2] A. D´efossez、L. Mazar´e、M. Orsini、A. Royer、P. P´erez、H. J´egou、E. Grave 和 N. Zeghidour，“Moshi：用于实时对话的语音文本基础模型”，*arXiv* 预印本 *arXiv:2410 00037*，2024年。[3] G. Pamisetty、R. A. Easow、K. Gupta 和 K. S. R. Murty，“Stream-tts：一种用于流式语音应用的基于 Kolmogorov-Arnold 网络的低延迟文本转语音”，载于 *ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2025年，第 1–2 页。[4] H. Sun、S. Hu、S. Liu、L. Meng、H. Wang、B. Han、Y. Yang、Y. Liu、S. Zhao、Y. Lu 等，“结合转导器与自回归建模的零样本流式文本转语音合成”，*arXiv* 预印本 *arXiv:2505.19669*，2025年。[5] Z. Sheng、Z. Du、S. Zhang、Z. Yan、Y. Yang 和 Z. Ling，“Syncspeech：基于时间掩码变换器的低延迟高效双流文本转语音”，2025年。 [在线]。可访问：https://arxiv.org/abs/2502.11094[6] A. Dekel、S. Shechtman、R. Fernandez、D. Haws、Z. Kons 和 R. Hoory，“边思考边说话：文本生成期间的流式语音合成”，载于 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024年，第 11 931–11 935 页。[7] T. Dang、D. Aponte、D. Tran 和 K. Koishida，“Livespeech：通过对音频离散编码进行自回归建模实现低延迟零样本文本转语音”，*arXiv* 预印本 *arXiv:2406.02897*，2024年。[8] H. Bai、R. Zheng、J. Chen、M. Ma、X. Li 和 L. Huang，“At：面向语音合成与编辑的对齐感知声学与文本预训练”，载于 国际机器学习会议。PMLR，2022年，第 1399–1411 页。[9] M. Kim、M. Jeong、B. J. Choi、D. Lee 和 N. S. Kim，“转导并说话：用于文本转语音及语义标记预测的神经转导器”，载于 *2023*年*IEEE*自动语音识别与理解研讨会 *(ASRU)*。IEEE，2023年，第 1–7 页。[10] Z. Borsos、R. Marinier、D. Vincent、E. Kharitonov、O. Pietquin、M. Sharifi、D. Roblek、O. Teboul、D. Grangier、M. Tagliasacchi 等，“AudioLM：一种用于音频生成的语言建模方法”，*IEEE/ACM*音频、语音与语言处理汇刊，第 31 卷，第 2523–2533 页，2023年。[11] S. Chen、C. Wang、Y. Wu、Z. Zhang、L. Zhou、S. Liu、Z. Chen、Y. Liu、H. Wang、J. Li、L. He、S. Zhao 和 F. Wei，“神经编解码器语言模型是零样本文本转语音合成器”，IEEE 音频、语音与语言处理汇刊，第 33 卷，第 705–718 页，2025年。[12] T. Wang、H. Wang、M. Ge、C. Gong、C. Qiang、Z. Ma、Z. Huang、G. Yang、X. Wang、E. Chng、X. Chen、L. Wang 和 J. Dang，“零样本文本转语音合成中的词级情感表达控制”，载于第三十九届神经信息处理系统年会，2025年。[13] T. A. Nguyen、B. Muller、B. Yu、M. R. Costa-Jussa、M. Elbayad、S. Popuri、C. Ropers、P.-A. Duquenne、R. Algayres、R. Mavlyutov 等，“Spirit-lm：交错式口语与书面语语言模型”，计算语言学协会汇刊，第 13 卷，第 30–52 页，2025年。[14] H. Wang、Y. Yang、S. Liu、J. Li、L. Meng、Y. Liu、J. Zhou、H. Sun、Y. Lu 和 Y. Qin，“Streammel：通过交错式连续自回归建模实现实时零样本文本转语音”，IEEE Signal Processing Letters，2025年。[15] Y. Yang、S. Liu、J. Li、H. Wang、L. Meng、H. Sun、Y. Liang、Z. Ma、Y. Hu、R. Zhao 等，“用于简单流式文本转语音合成的交错式语音文本语言模型”，arXiv 预印本 arXiv:2412.16102，2024年。[16] Z. Du 等，“CosyVoice 2：利用大语言模型实现可扩展的流式语音合成”，arXiv 预印本 arXiv:2412 10117，2024年。[17] Z. Du、C. Gao、Y. Wang、F. Yu、T. Zhao、H. Wang、X. Lv、H. Wang、C. Ni、X. Shi 等，“Cosyvoice 3：通过扩展规模与后训练迈向野外语音生成”，arXiv 预印本 arXiv:2505.17589，2025年。[18] H. Hu、X. Zhu、T. He、D. Guo、B. Zhang、X. Wang、Z. Guo、Z. Jiang、H. Hao、Z. Guo 等，“Qwen3-tts 技术报告”，arXiv 预印本 arXiv:2601.15621，2026年。[19] R. H. Bai、Z. Gu、T. Likhomanenko 和 N. Jaitly，“Speakstream：采用交错数据的流式文本转语音”，2025年。 [在线]。可访问：https://arxiv.org/abs/2505.19206[20] M. Bain、J. Huh、T. Han 和 A. Zisserman，“WhisperX：长篇音频的时间精确语音转录”，INTERSPEECH 2023，2023年。[21] R. Ardila 等，“Common voice：大规模多语言语音语料库”，载于 LREC 论文集，2020年。[22] P. Anastassiou 等，“Seed-TTS：高质量多功能语音生成模型系列”，arXiv 预印本 arXiv:2406 02430，2024年。[23] DeepSeek-AI，“DeepSeek-V3 技术报告”，2024年。 [在线]。可访问：https://arxiv.org/abs/2412.19437[24]nvidia，“nvidia/parakeet-tdt-0.6b-v2”，2026年，通过 CROVIA 透明度注册库访问。 [在线]。可访问：https: // huggingface.co/nvidia/parakeet-tdt-0.6b-v2[25]Z. Ma、Z. Zheng、J. Ye、J. Li、Z. Gao、S. Zhang 和 X. Chen，“emotion2vec：用于语音情感表征的自监督预训练”，载于计算语言学协会论文集：ACL 2024。计算语言学协会，2024年，第 15 747–15 760 页。 [在线]。可访问：https://aclanthology.org/2024.findings-acl.931/[26] J. Kong、J. Kim 和 J. Bae，“HiFi-GAN：用于高效高保真语音合成的生成对抗网络”，2020年。[在线]。可访问：https://arxiv.org/abs/2010.05646