长上下文语音合成与上下文感知内存

Zhipeng Li[1,2], *Xiaofen Xing*[1,*], *Jingyuan Xing*[1], *Hangrui Hu*[2], *Heng Lu*[2], *Xiangmin Xu*[1,3]

[1]South China University of Technology, China

[2]Speech Lab, Alibaba Group, China

[3]Pazhou Lab, China

eeleezp@mail.scut.edu.cn, xfxing@scut.edu.cn

Abstract

在长文本语音合成中，当前的方法通常以句子级别将文本转换为语音，并将结果拼接起来形成伪段落级别的语音。这些方法忽略了段落的上下文连贯性，导致生成的长篇幅语音自然度降低以及风格和音色上的不一致。为了解决这些问题，我们提出了一种基于上下文感知记忆（CAM）的长上下文化文本到语音（TTS）模型。CAM 模块集成了长期记忆和局部上下文细节，并在长段落内进行动态内存更新和传递以指导句子级别的语音合成。此外，前缀掩码通过允许前缀标记上的双向注意力同时保持单向生成来增强上下文中学习的能力。实验结果表明，所提出的方法在韵律表现力、连贯性和段落级别语音的上下文推理成本方面优于基线和最先进的长上下文化方法。音频样本可在 https://leezp99.github.io/长上下文-CAM-文本到语音/获取。

Index Terms: 文本转语音，长上下文，内存压缩

1. 介绍

近年来，随着生成模型 [1, 2, 3, 4]、声码器 [5, 6] 以及非自回归模型 [7, 8, 9, 10, 11] 和自回归模型 [12, 13, 14] 的进步，语音生成技术已经达到能够产生具有人声质量的自然语音的水平。得益于大型语言模型（LLM）[15, 16] 所展现的高度可扩展性，近期的研究 [17, 18, 19] 已采用 LLM 作为 TTS 任务中文本到语义标记建模的核心模块，展示了卓越的自然语义建模能力。

随着语音助手、有声读物和新闻广播等应用需求的增长，文本转语音任务的目标已逐渐从高质量的句子级合成转向连贯且富有表现力的段落级语音。在这些长上下文场景中，历史文本与语音之间存在显式和隐式的上下文依赖关系。然而，当前主流方法通常将段落级别的文本分割成句子级别的文本，并分别合成句子级的语音。这种方法忽视了段落在内部及跨段落之间的上下文关联，导致以下问题：1）减弱了韵律的表现力；2）在风格、音色和语速上的一致性差，特别是语音连贯性，严重影响听众体验。

一些长上下文建模方法最近被提出。Xin 等人。[20] 利用前面的语音（1 句话）和双向文本上下文（2-3 句话）来改进语音韵律；Xiao 等人。[21] 提出了一种基于固定长度前向语音的记忆缓存递归机制，以及一个上下文文本编码器；Xue 等人。[22] 提出了一个多模态上下文增强的 Q-Former，它压缩前面的文本和语音（5 句话）以利用更长的上下文信息；Xue 等人。[23] 提出利用 CA-CLAP 通过上下文检索来增强生成，选择整个语音-文本提示（1-2 句话）作为前缀令牌来引导语音生成。尽管这些方法为长上下文 TTS 提供了有价值的见解，但仍有一些需要改进的部分。1) 语音韵律依赖于文本，因此需要一个精确的句子级别语音合成上下文检索机制；2) 过长的指导提示会导致模型不稳定，需要更简洁的指导提示；3) 历史上下文中的一部分会在每次推理中重复使用，这将导致高计算成本；4) 无法有效利用远处的上下文信息。

受 Google 研究团队 [24] 提出的具有长期压缩记忆的无限注意机制启发，我们提出了一个上下文感知内存（CAM）模块（图 1）。该 CAM 块利用 Perceiver Resampler 来压缩目标文本，并分别从历史文本和语音的长期记忆及局部上下文细节中检索关键依赖信息。它动态更新内存以指导当前的语音合成并将之传递到不同长度的后续句子。

*Corresponding author.

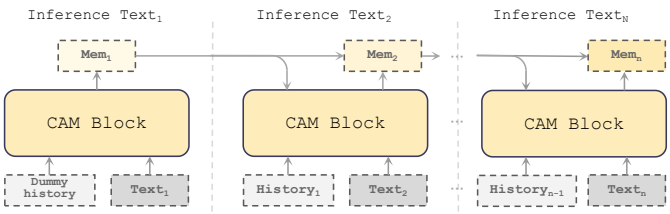

图 1: *CAM* 块的段落推理过程

我们将 CAM 块与大型语言模型（LLM）集成，构建了一个用于长上下文文本到语义建模的模块。为了进一步增强上下文学习能力，我们用前缀掩码替换了传统的因果掩码，允许内存输入和文本输入自由相互注意。相比于采用几段历史句子的方法，我们的解决方案不仅高效且具有创新性，只需要固定长度的长期记忆及上一个上下文，就能将模型视野从句子扩展到段落。我们将贡献总结如下：

- 我们提出了一种基于上下文感知记忆的长上下文 TTS 模型，该模型从长期记忆和局部细节中检索并更新记忆，以指导段落内高质量的句子级语音合成。
- 我们引入前缀掩码来替换大语言模型中的因果掩码，增强理解和上下文学习能力。
- 客观和主观评估表明，我们提出的方法在自然度、连贯性和推理成本方面优于基线和最先进的长上下文 TTS 方法。

2. 方法论

基于上下文感知内存块的长上下文 LM 模型的一般架构如图 2 所示。该模型包括两个核心组件：一个 CAM 模块用于通过压缩、检索和更新来维护上下文记忆，以及一个大型语言模型（LLM）用于文本到语义建模。假设合成的目标话语索引为 n，我们分别使用第 (n-1) 个语音/文本作为历史 - 语音 $_{n\text{-}1}$/文本 $_{n\text{-}1}$。从第 (n-1) 次语音合成效应传递下来的内存表示为 Mem$_{n\text{-}1}$。这些，连同文本 $_n$ 一起，用于指导语音 $_n$ 的生成。

2.1. 上下文感知内存块

如图 2(a) 所示，我们提出的 CAM 块由三个阶段组成：压缩、检索和更新。由于语音和文本之间固有的模态差异，我们设计了专用的 CAM-Speech 和 CAM-Text 模块（作为 CAM-S、CAM-T），它们具有相同的结构但权重独立。同样地，内存被分为语音内存（Mem-S）和文本内存（Mem-T）。为简化起见，在本节中省略了模态注释（-S/-T）。图 1 中的 Dummy History 分别由静音语音和空白文本组成，设计用于首个合成话语的情景，在这种情况下没有先前的上下文可用。

$$Mem_n = CAM(Text_n, Mem_{n\text{-}1}, History_{n\text{-}1})$$

压缩。我们使用 Perceiver Resampler[25, 26] 来执行可变长度文本 $_n$ 隐含表示和固定长度可学习隐含查询向量之间的交叉注意力。Perceiver Resampler 的输出是目标文本的压缩固定长度隐含表示。重采样方法使模型能够从原始特征中提取关键信息。然后，将突出的压缩文本隐含作为查询发送到交叉注意力（在检索阶段）。

检索。由于长期记忆 Mem$_{n\text{-}1}$ 是从之前的发言 Text$_{n\text{-}1}$ 中检索出来的，并且不包含 History$_{n\text{-}1}$，因此需要先进行融合。我们将它们连接并输入基于 Transformer 的历史编码器中。然后，使用从压缩阶段获得的目标文本表示（作为 Query），对融合的上下文信息（作为 Key 和 Value）执行多次检索以捕捉当前发言中最关键的上下文依赖关系 Mem$^*_n$。

更新。检索后，我们更新记忆并获得下一个状态。我们将长期记忆 Mem$_{n\text{-}1}$ 和检索到的内存值 Mem$^*_n$ 通过一个学习到的标量 α 进行聚合，允许在长期和局部上下文之间进行动态权衡。

$$Mem_n = sigmoid(\alpha) \odot Mem^*_n + (1 - sigmoid(\alpha)) \odot Mem_{n\text{-}1}$$

压缩、检索和更新后，获得最新的记忆表示 Mem-S$_n$ 和 Mem-T$_n$，这些表示结合了长期上下文和局部内容。然后将这些表示输入到长上下文语言模型中，以指导文本到语义的建模。

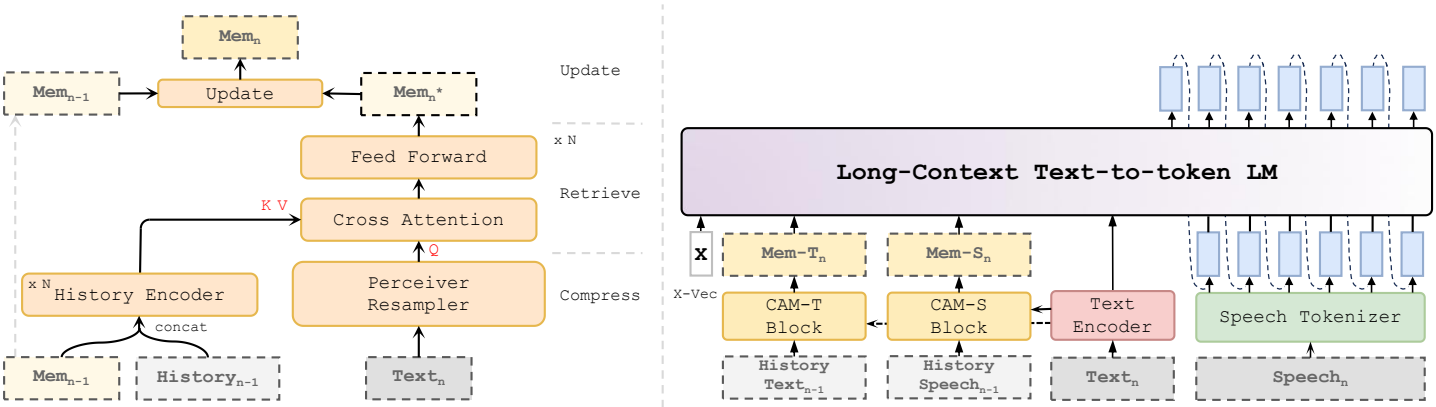

图 2: *(a)* 展示了 *CAM* 模块的概述，包含压缩、检索和更新三个阶段。*(b)* 给出了长上下文语言模型的说明。

2.2. 长上下文语言模型

为了增强韵律表达和连贯性，我们使用上述模块生成的上下文记忆 Mem-T_n 和 Mem-S_n 作为 LLM 的输入。主要架构遵循 CosyVoice[17]，我们使用 Cam++[27] 的 X-vectors，并在训练和推理阶段均采用段落级别的 X-vectors 而非语句级别。这使得 LLM 具有更大的学习空间，从而增强生成语音的自然度和连贯性。LLM 的输入如下：

$$[X_{vec}, Mem\text{-}T_n, Mem\text{-}S_n, TE(Text_n), ST(Speech_n)]$$

TE 和 ST 分别是文本编码器和语音标记化器。预训练的 Flow Matching 和 Vocoder 被应用于将生成的标记转换为波形。

在长上下文语言模型中，vec，Mem-T_n，Mem-S_n 被视为预填充信息。因此，在训练过程中，仅考虑生成的语音标记的交叉熵损失。

2.3. 前缀掩码对于大语言模型

如今，大多数文本到语义的 TTS 大型语言模型属于仅解码器架构。这些模型通常使用因果掩码，其中每个标记只能关注前面的标记和自身。然而，在整个序列训练过程中严格应用因果掩码可能会限制模型的表现 [28]，特别是对于具有上下文学习能力的长上下文 LM。因此，我们为长上下文 LM 引入了前缀掩码（图 3），它对前缀标记如 X-vec、Mem 和 Text 应用双向注意，允许沿时间维度进行前缀序列的双向编码。保持生成标记上的单向注意力以确保生成的一致性。

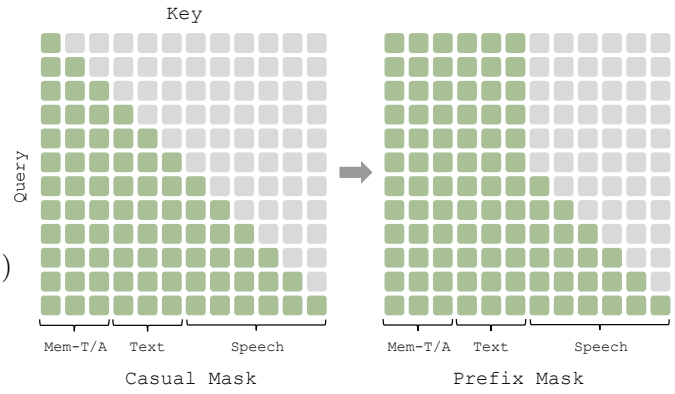

图 3: 因果掩码和前缀掩码在大语言模型中的示例说明

3. 实验

3.1. 数据集

为了训练我们提出的长上下文语言模型，我们从互联网收集了大约 15,000+小时的中文普通话有声读物，包括约 75000+完整的章节。数据主要由单播客语音组成。我们利用 Demucs[29] 从原始语音数据中提取干净的人类歌声。我们将长片段分割成较小的部分，每段不超过 30 秒，平均时

长为 16.2 秒。Paraformer[30] 用于转录这些数据。我们随机选择了 100 个额外的完整章节进行验证。

3.2. 实验设置

训练。长上下文语言模型是从 CosyVoice-LM 修改而来的，语音标记器是 50Hz 版本。对于长上下文语言模型，我们从头开始训练，并使用一个恒定的学习率 10^{-4}。在 CAM 块中，Perceiver 重采样器产生固定数量的 32 个嵌入。历史编码器由两个堆叠的 Transformer 块组成，并采用两层检索阶段。所有模型都经过了 1 亿步的训练，每批动态批量大小为 10,000 个标记，以确保完全收敛。
推理。对所有语言模型采用了随机采样解码策略。

3.3. 模型评估

3.3.1. 比较方法

为了评估我们方法的性能，我们将它与最先进的长上下文 TTS 系统进行了比较。
• **均方误差-Qformer** 薛等人。[22] 提出了一种多模态上下文增强的 Qformer，利用压缩的长上下文信息来提高 TTS 性能。
• **掌声-RAG** 薛等人。[23] 提出了一种基于提示的 TTS 框架，该框架使用了上下文感知对比语言-音频预训练模型来增强 RAG（检索增强生成）。并且它利用整个提示来引导生成过程。
• **提出的** 我们提出的具有上下文感知内存块和前缀掩码的长上下文语言模型。

我们复现了在 Cosyvoice-LM 骨干网络上的 MMCE-Qformer 和 CLAP-RAG 模型，按照原始论文的实现方法，上下文长度分别设置为 5 和 1。

3.3.2. 消融研究

我们进行消融研究以评估长上下文语言模型中关键模块的有效性。
• **基线** 使用 Cosyvoice-LM[17] 作为基础从零开始训练的 LM。
• **无内存-T** 长上下文语言模型无 CAM -T 块。
• **无内存-S** 长上下文语言模型无 CAM -S 块。
• **无前缀掩码** 带有标准随意掩码的长上下文语言模型。

我们使用客观和主观指标来评估上述模型。

3.3.3. 评估指标

主观评价 我们随机选择了 20 个句子级别的（大约 15 秒）语音样本和 10 个长段落形式的（大约 60 秒）语音样本进行评估。长段落语音是通过结合多个句子级别语音构建而成。我们进行了段落 MOS（平均意见评分）来评估真实录音和句子级合成语音的表现力和自然度。段落 CoMOS（一致性平均意见评分）用于评估整个长段落语音在风格和音色上的一致性。在每次 MOS 测试中，10 名母语为普通话的听评人对 MOS 和 CMOS 进行打分，分数范围从 1 到 5，间隔为 0.5 分。最终得分报告了 95%的置信区间以确保统计上的可靠性。
目标评估 对于客观指标，我们评估说话人相似度（SIM）、鲁棒性（CER）和语音质量（Speech-BertScore）。具体来说，对于说话人相似度，我们在推理过程中计算语言模型中使用的说话人级别 X-向量与生成样本的 X-向量之间的余弦相似度，均值代表总体相似度，方差表示音色一致性稳定性。对于鲁棒性，我们使用 Paraformer-zh 作为 ASR 模型来评估内容一致性。对于语音质量，我们采用 SpeechBERTScore[31] 进行质量估计，因为它相较于以前的方法显示出更高的与人工评分的相关性。
推理上下文成本 在句子级语音合成中，Num 表示每个长上下文方法使用的句子级上下文的数量，Prefix Len 指的是输入到语言模型中的与上下文相关的标记数量。

3.4. 实验结果

3.4.1. 性能比较

我们对三种长上下文方法进行了比较。首先，我们分析了推理中的上下文成本。对于每个句子级别的语音合成，MMCE-Qformer 以五个上下文作为输入，并生成 64 个标记作为前缀令牌来指导生成；CLAP-RAG 使用 CLAP 检索所有上下文

表 1: 所提方法、*SOTA* 长上下文方法及消融研究的评估结果。

	主观		目标			上下文成	
	MOS (↑)	CoMOS (↑)	SpeechBERT (↑)	CER (↓)	SIM (↑)	Num	Prefix
Ground Truth	$4.406_{\pm 0.095}$	$4.870_{\pm 0.056}$	100	4.286%	$93.716_{(0.046)}$	–	-
MMCE-Qformer	$3.557_{\pm 0.082}$	$3.885_{\pm 0.112}$	79.031	5.075%	$85.110_{(0.021)}$	5	Fixed
CLAP-RAG	$3.489_{\pm 0.133}$	$3.717_{\pm 0.183}$	78.892	6.234%	$84.920_{(0.037)}$	1	Vari
Baseline	$3.468_{\pm 0.113}$	$3.460_{\pm 0.120}$	77.776	5.850%	$85.051_{(0.035)}$	–	-
提议	$\mathbf{3.796_{\pm 0.091}}$	$\mathbf{3.992_{\pm 0.127}}$	**80.448**	**4.140%**	$\mathbf{85.685_{(0.019)}}$	1	Fixed
无内存-T	$3.604_{\pm 0.146}$	$3.887_{\pm 0.135}$	78.358	5.036%	$85.461_{(0.031)}$	1	Fixed
无内存-S	$3.516_{\pm 0.155}$	$3.827_{\pm 0.129}$	78.063	5.496%	$85.150_{(0.039)}$	1	Fixed
无前缀掩码	$3.661_{\pm 0.158}$	$3.846_{\pm 0.125}$	80.243	4.633%	$85.135_{(0.026)}$	1	Fixed

中相关性最高的句子，并利用完整的文本和语音（~900 个标记）作为可变长度的前缀令牌。这种方法对键值缓存 (KV) 带来了巨大的计算负担。相比之下，提出的方法结合了这两种方法的优点，只需要以前的上下文历史 $_{n\text{-}1}$ 和记忆 $\text{Mem}_{n\text{-}1}$ 作为输入，同时生成 64 个记忆标记用于合成。这显著减少了检索和自回归推理的消耗。

主观评估 MOS 和 CoMOS 表明，使用检索到的完整提示来引导生成（CLAP-RAG）可以提高连贯性，但并没有在自然度方面显示出显著改善。同时，所提出的方法在整体性能上优于 MMCE-Qformer 和 CLAP-RAG。在客观测试中，所提出方法在 SIM 上表现出更好的均值和方差，表明它生成的语音与目标说话人的音色最为相似且稳定。在 SpeechBERTScore 中，所提出的方法也略微优于其他两种模型。我们将其优势归因于引入了记忆令牌，这些令牌通过平衡最新的上下文信息与长期信息来持续更新记忆，从而通过关键的上下文线索引导语音生成。此外，CLAP-RAG 在 CER 方面表现最差，我们认为这是由于在推理过程中由过长提示序列引起的幻觉效应增加所致。相比之下，所提出的方法采用了固定数量的前缀令牌，减轻了推理不稳定并增强了鲁棒性 (CER)。

3.4.2. 消融分析结果

我们进行了消融研究以探索 Proposed method 中每个组件的影响。

实验结果表明，缺乏文本和语音记忆模块的基线模型无法利用上下文信息进行指导，导致在表现力和连贯性方面的性能不佳。此外，通过 MOS 和 CoMOS 测试，我们发现使用前缀标记来引导生成并具备上下文学习能力的方法明显受益于前缀掩码。这表明前缀掩码提高了模型生成上下文驱动预测的能力。另外，我们将语音记忆和文本记忆的效果进行了比较，结果表明语音上下文信息表现更好。我们认为这是由于文本与语音之间存在一对一多的关系，单一的文本输入可以对应多个具有音调和情感变化的有效语音输出。与文本上下文相比，语音上下文本质上捕捉了更丰富的声学和韵律信息，使得语音记忆嵌入在增强生成语音的一致性和音色连贯性方面更加有效。此外，通过数据检查，我们发现真实值的 CER 分数受到源分离技术 Demucs 的影响，一些数据显示声音质量下降，导致更高的 CER 分数。尽管存在这一挑战，基于 LLM 的模型的强大性能帮助补偿了这些退化现象，使得生成的语音的 CER 分数低于真实值数据。

最后，需要注意的是，尽管所提出的方法表现

出有希望的性能，但在自然度和连贯性方面与真实有声读物数据（ ground truth ）之间仍存在明显的差距。在段落级语音生成方面仍有很大的改进潜力，这需要在未来进行进一步的研究和完善。

4. 结论

在这项研究中，我们提出了一种有效的长上下文 TTS 模型，该模型利用压缩的上下文感知记忆来提升句子级语音合成中的自然度和连贯性。CAM 模块整合并检索长期记忆和局部上下文细节，动态更新记忆以保持段落内的关键上下文历史。最新上下文记忆用作前缀信息指导 LM 模型中的标记生成，并使用前缀掩码增强上下文学习。在中国普通话有声读物语料库上的实验表明，所提出的方法在段落阅读中实现了更高的表现力、连贯性和更低的上下文计算成本，相较于基线模型和之前的长上下文方法。

5. 致谢

该工作得到南沙重点项目的部分支持，资助编号为 2022ZD011；广东省基础与应用基础研究基金（2025A1515011203）的部分支持；以及广东省人数字孪生重点实验室（2022B1212010004）的部分支持。

6. References

[1] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[3] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[4] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[6] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Hiftnet: A fast high-quality neural vocoder with harmonic-plus-noise filter and inverse short time fourier transform," *arXiv preprint arXiv:2309.09493*, 2023.

[7] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," in *Interspeech 2023*, 2023, pp. 4374–4378.

[8] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, J. Bian *et al.*, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *The Twelfth International Conference on Learning Representations*.

[9] D. Yang, D. Wang, H. Guo, X. Chen, X. Wu, and H. Meng, "Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models," in *Interspeech 2024*, 2024, pp. 4398–4402.

[10] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan *et al.*, "E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts," *arXiv preprint arXiv:2406.18009*, 2024.

[11] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[12] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.

[13] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[14] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[17] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[18] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao *et al.*, "Seed-tts: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[19] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, "Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.

[20] D. Xin, S. Adavanne, F. Ang, A. Kulkarni, S. Takamichi, and H. Saruwatari, "Improving speech prosody of audiobook text-to-speech synthesis with acoustic and textual contexts," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] Y. Xiao, S. Zhang, X. Wang, X. Tan, L. He, S. Zhao, F. K. Soong, and T. Lee, "Contextspeech: Expressive and efficient text-to-speech for paragraph reading," in *Interspeech 2023*, 2023, pp. 4883–4887.

[22] J. Xue, Y. Deng, Y. Han, Y. Gao, and Y. Li, "Improving audio codec-based zero-shot text-to-speech synthesis with multi-modal context and large language model," in *Interspeech 2024*, 2024, pp. 682–686.

[23] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Retrieval augmented generation in prompt-based text-to-speech synthesis with context-aware contrastive language-audio pretraining," in *Interspeech 2024*, 2024, pp. 1800–1804.

[24] T. Munkhdalai, M. Faruqui, and S. Gopal, "Leave no context behind: Efficient infinite context transformers with infini-attention," *arXiv preprint arXiv:2404.07143*, 2024.

[25] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[26] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech 2024*, 2024, pp. 4978–4982.

[27] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[28] N. Ding, T. Levinboim, J. Wu, S. Goodman, and R. Soricut, "Causallm is not optimal for in-context learning," *arXiv preprint arXiv:2308.06912*, 2023.

[29] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *Interspeech 2022*, 2022, pp. 2063–2067.

[31] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, "Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics," in *Interspeech 2024*, 2024, pp. 4943–4947.