arXiv:2601.10770v1 [cs.SD] 15 Jan 2026

基于自回归 Transformer 统一语音识别、语音合成与语音转换

技术报告

Runyuan Cai Yu Lin Yiming Wang Chunlin Fu Xiaodong Zeng

AutoArk-AI

{runyuan.cai, yu.lin, yiming.wang, chunlin.fu, xiaodong.zeng}@autoark.ai
https://github.com/AutoArk/GPA

摘要

传统语音系统通常依赖针对文本转语音（TTS）、自动语音识别（ASR）和语音转换（VC）分别构建的专用模型，导致流程彼此割裂，限制了可扩展性、效率和跨任务泛化能力。

本文提出**通用音频（GPA）**[*]，这是一种统一的音频基础模型，在单一大语言模型（LLM）架构中整合了多项核心语音任务。GPA 基于共享的离散音频令牌空间运行，并支持指令驱动的任务诱导，使单个自回归模型无需架构修改即可灵活执行文本转语音、自动语音识别和语音转换。

这种统一设计结合了基于离散语音令牌的完全自回归形式、跨语音领域的联合多任务训练，以及能够实现高并发能力和高吞吐量的可扩展推理流水线。由此形成的模型系列支持高效的多尺度部署，包括针对边缘设备和资源受限环境优化的轻量级3亿参数版本。这些设计选择共同表明，统一的自回归架构能够在多样化语音任务上实现有竞争力的性能，同时仍适用于低延迟的实际部署。

关键词 文本转语音 ·自动语音识别 · 语音转换 · 基础模型

1 引言

生成式人工智能的迅猛发展从根本上重塑了语音处理的格局。受大语言模型（LLM）在自然语言处理领域取得成功的推动，语音领域已从专门的特定任务流水线转向大规模、数据驱动的基础模型。在文本转语音（TTS）、自动语音识别（ASR）和语音转换（VC）等核心任务中，神经网络方法已取得重大突破，并在特定场景下展现出媲美人类的表现[49, 38, 31]。

尽管各项任务都取得了成功，当前的生态系统在架构上仍然支离破碎。虽然文本转语音、自动语音识别和语音转换具有共同的语言学和声学基础，但它们通常被视为彼此孤立的问题，各自依赖不同的建模范式和架构假设：

- **文本转语音：** 现代零样本文本转语音系统已发展出多种类型，从以自回归方式生成声学词元的离散基于词元的语言模型 [45, 22, 42, 48], ，到通过隐式学习文本与语音之间对齐关系的扩散概率映射 [25, 3], ，还包括

[*]正如学业平均绩点反映一个人在不同学科上的能力一样，GPA旨在让模型在所有音频任务中都取得不错的表现。

将语义建模与非自回归声学渲染相结合的由粗到细混合架构 [19, 13, 15]。尽管混合模型目前在合成质量和可控性之间实现了较好的平衡，但其多阶段设计带来了显著的系统复杂性，妨碍了统一建模和高效部署。

- **自动语音识别：** 自动语音识别（ASR）已迅速转向采用大规模预训练编码器和编码器—解码器模型，例如 Whisper [38], Qwen-Audio [11, 10], SenseVoice [1], 和 Seed-ASR [6],，这些模型展现出强大的多语言和多任务能力。然而，这些模型主要针对判别式识别目标进行优化，且在架构上专用于自动语音识别，因此难以在不进行大量架构修改或引入辅助组件的情况下，将其用于生成式或面向转换的语音任务 [51, 40, 27, 47, 52]。
- **语音转换：** 语音转换（VC）系统通常依赖基于解耦的框架，将语音分解为语言内容、音色和说话风格，从而实现对属性的可控操控 [36, 34]。现有方法包括基于平行数据的转换 [60, 20], 、解耦表示学习 [64, 14], 和大规模上下文学习 [46, 26, 54]。尽管取得了这些进展，语音转换方法仍常受限于可扩展性不足以及属性解耦不完善等问题；为实现稳健的零样本泛化，这些方法往往需要平行语料库、显式风格标注或额外的微调阶段 [21, 61, 35, 59]。

这种碎片化导致建模工作重复、跨任务知识迁移受限，并显著增加部署复杂度。迄今为止，仍缺乏一种**完全统一的生成框架**，能够在**单一的自回归建模与推理流水线中**原生支持语音理解、生成和语音转换，而无需依赖任务专用的指令头或路由机制。

在本文中，我们提出通用音频（General-Purpose Audio，GPA），一种将TTS、ASR和VC整合到单一自回归LLM框架中的统一语音基础模型。不同于多阶段或混合系统，GPA将所有语音任务都建模为共享离散音频令牌空间上的序列建模问题。通过以任务专用指令作为条件控制，单一模型无需进行架构修改，即可在语音识别、音频合成和语音转换之间无缝切换。

本文的主要贡献总结如下：

- **统一的自回归音频框架：** 我们提出GPA，一种用于TTS、ASR和VC的统一自回归LLM框架。GPA采用双分词器方案，同时利用GLM分词器[55] 和BiCodec分词器[48] 进行语音离散化。通过将所有任务置于共享的离散音频令牌空间中，GPA无需任何架构修改即可实现指令驱动的任务切换。
- **协同多任务学习：** 我们表明，使用 Emilia [18]数据集和补充的专有数据进行多任务联合训练，能够提升整体模型性能。这种统一方法使自动语音识别、文本转语音和语音转换能够受益于离散潜在空间中的共享表征，其效果优于分别训练各项任务。
- **高吞吐量与流式处理效率：** 我们证明，GPA 采用纯自回归架构，能够简化部署，提供更强的并发能力和更高的整体吞吐量，优于混合方案。这种设计使 GPA 非常适合流式应用和高需求推理场景。
- **边缘优化可扩展性：** 我们推出了一系列不同规模的 GPA 模型，其中包括专为在资源受限环境和边缘环境中部署而设计的超紧凑型 3亿参数版本。这使得单个模型实例能够在设备上执行多任务推理，同时保持极低的内存占用。

2 通用音频

语音理解、生成和编辑任务，例如自动语音识别、文本转语音和语音转换，通常被视为彼此独立的问题，并通过特定任务流水线加以处理。尽管这些设计行之有效，但它们使建模选择变得碎片化，并且往往依赖异构组件。

尽管大语言模型的最新进展推动了语音任务的发展，但依赖下游合成模块的混合方法[15] [9][32] 仍受延迟和系统效率的限制。相比之下，BiCodec系统[48] 提供了经验证据，表明纯自回归架构能够在文本转语音任务中取得出色表现。

这一成功表明，当语音和文本以共享的离散词元空间表示时，各类音频任务都可以统一表述为序列到序列生成问题，区别仅在于输入和输出规范。

基于这一发现，我们得出了一个统一原则：**所有语音任务都可以在完全离散的词元空间中采用单一的自回归形式**。

图1展示了我们提出的*GPA* 框架的架构。GPA 基于 Qwen3[53] 大语言模型骨干网络，采用 BiCodec 框架，并集成额外的 GLM[55] 分词器以增强语义抽象。通过将文本、语义和声学信息统一到共享的离散词元空间中，自动语音识别、文本转语音和语音转换可以自然地成为单一序列生成任务的不同实例。由此构建出一个完全基于自回归生成的通用音频框架，其任务行为完全由指令控制，无需任务特定架构或解码流程。

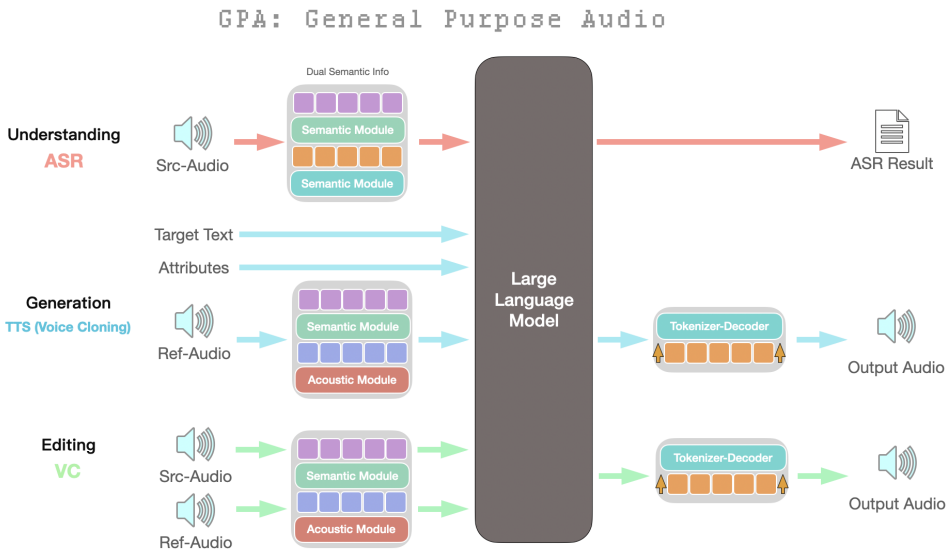


图1：**所提出的 GPA 框架架构**。该模型采用共享的大语言模型骨干网络，统一三项核心音频任务：理解、生成和编辑。根据任务不同，模型通过声学模块处理不同组合的输入，生成相应的文本或音频输出。

因此，GPA 消除了混合方法固有的效率瓶颈，实现了高并发能力和高吞吐量。BiCodec表示与GLM分词器的搭配，进一步促进了跨任务协同，提升了模型性能。我们还证明，在这一统一框架内对多个任务进行联合训练能够带来互惠收益，且始终优于单任务基线。

2.1 完全自回归架构的动机

我们采用完全自回归架构作为模型的基础，通过统一的生成式形式涵盖所有音频任务。

语音任务在其底层信息空间中天然相互关联，具有大量表征重叠，而任务专用设计或模块化设计往往未能充分利用这些重叠。要充分利用这种共享结构，必须采用能够进行端到端联合训练和跨任务推理的统一框架。

这一需求与近期研究发现相符：当单模态骨干网络进行端到端训练时，纯自回归模型即使参数更少，也能胜过交叉注意力架构[24]。这类优势通常归因于自回归模型对称且同质的优化方式；相比之下，交叉注意力架构会在联合训练期间引入架构不对称性和梯度瓶颈，从而限制统一多模态学习的可扩展性。

因此，完全自回归形式为异构音频任务的联合训练提供了更具可扩展性和延展性的基础，同时避免了基于交叉注意力的设计所带来的结构性限制。

2.2 分词与编解码器设计

BiCodec[48] 使用互补的声学和语义 token 流表示语音，其中语义 token 捕捉语音信号所固有的高层信息。尽管这种语义表示[5] 行之有效，但并未与人类语言中的离散语言单位明确对齐。

近期研究[28]表明，将来自不同架构和预训练范式模型的预训练表示相结合，能够实现更丰富、更稳健的语义建模。受这一原则启发，我们引入了 GLM [55]语音分词器——该分词器源自一个在文本监督训练[37] 下训练的大型自动语音识别模型——通过转录对齐的语义丰富统一词元空间。GLM 分词器并非取代 BiCodec 的内在语义流，而是通过提供与词汇和句法结构更为紧密对应的语言单位来对其进行补充；仅凭无监督语音表示，这些结构往往具有固有的歧义，或根本不存在。
鉴于现代大语言模型骨干网络具有强大的建模能力，系统性能主要受词元表示的语义保真度和语言基础所限制。因此，尽管这些源自自动语音识别的语义词元增加了表示复杂度，我们仍能够为骨干网络提供更加符合语言规律的信号，从而在不增加模型规模的情况下，实现表现力更强、可控性更高的音频建模。

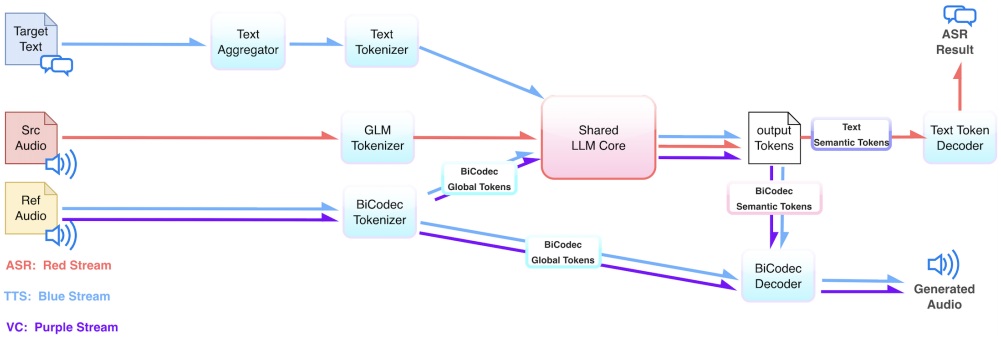


图 2：**GPA 中的分词与任务流程**。不同音频任务通过改变输入和输出的令牌构成来实现，所有任务均由共享的自回归骨干网络处理。

在这种设计下，语音理解和生成任务——包括自动语音识别、文本转语音和语音转换——可以利用同一个离散语义空间。图 2详细展示了各项任务在共享自回归骨干网络的同时所采用的分词与反分词路径。

2.3 联合多任务训练

我们采用联合多任务方式训练 GPA，使用 Emilia[18] 数据集和经过筛选的内部语料库混合而成的数据，覆盖自动语音识别、文本转语音和语音转换，并在统一的训练计划下进行训练。
与特定任务训练相比，联合优化能持续提升所有任务的性能。这一提升源于不同语音任务在共享的离散词元空间中所提供的互补归纳信号；先前对语音生成目标进行联合建模时也观察到了这一现象[57]。自动语音识别促使声学观测与具有语言依据的语义词元之间形成紧密对齐，使模型的内部表示锚定于稳定的词汇和句法结构。相比之下，文本转语音和语音转换强调忠实的声学实现以及说话人相关变化，促使模型在语义输入的条件下保留细粒度的韵律和音色信息。因此，联合训练能够形成一种平衡表示，兼具语言依据和声学表现力。

从优化角度来看，完全自回归形式通过共享的下一词元预测目标传导来自异质任务的梯度，并借助多样化的条件生成模式对骨干网络进行隐式正则化，这与在统一自监督目标下训练的大型自回归语言模型的观察结果一致 [39]。不同任务并非相互争夺容量，而是作为共享模型的相互约束，减轻模型对特定任务偏差的过拟合，并提升泛化能力。

以往研究也观察到，联合训练语音识别与语音合成目标可以让跨任务监督发挥隐式正则化作用[44]。在异构数据源上训练时，这种效果尤为显著[4], : 联合目标促使模型学习任务不变的语义抽象，同时增强其对说话人身份、录音条件和标注质量差异的鲁棒性。

重要的是，统一的分词方案无需进行架构修改或添加辅助任务头，即可实现任务间的无缝混合。所有任务都被转化为不同输入–输出组合下的下一词元预测，使模型能够受益于更丰富的数据多样性和更高的训练信号密度，同时不增加模型规模或推理复杂度。因此，联合多任务训练不仅能提升各个任务的性能，还能构建出更加连贯且可迁移的音频基础模型。

3 实证评估

为了评估 GPA 框架，我们针对其训练动态以及在多种语音任务上的性能开展了实证研究。我们首先介绍训练方案，包括数据组成、样本构建策略和微调流程。随后，我们对模型性能进行定量和定性分析，涵盖推理特性、延迟、吞吐量以及任务特定评估指标，从而全面了解 GPA 在实际场景中的有效性和效率。

3.1 训练细节

本小节介绍用于训练 GPA 的训练方案、数据组成和样本构建策略，重点关注可扩展性、可复现性以及与统一自回归学习的兼容性。为便于复现和进一步研究，完整的训练代码和配置文件已发布在我们的公开 GitHub 仓库中。[2]

从头训练。 所有 GPA 模型均从头训练，不依赖预训练语音或文本生成权重。这一设计选择确保学习过程不受语言模型、声学编码器或任务特定架构所引入的外部归纳偏置影响。

相反，所有模型组件都在单一的下一词元预测目标下进行联合优化。因此，不同任务（如转录、语音合成或口语意图理解）之间的行为差异，完全源于输入–输出词元序列的变化，而非专用模块或独立预测头。

预训练数据。 我们使用约**1M** 小时的语音数据进行大规模预训练，这些数据来自多个公开语料库，包括 Emilia [18], , 以及经过筛选的补充数据集。该补充数据集旨在扩大说话人多样性、录音条件和语言变化的覆盖范围。此阶段并不针对任何单一的下游任务，而是让模型接触广泛的声学–语义对应关系，从而为统一的语音理解与生成奠定坚实基础。

监督微调。 预训练之后，我们使用约**200K** 小时的数据进行监督微调，涵盖自动语音识别、文本转语音和语音转换。在此阶段，通过在相同的底层词元空间中构造不同任务专属的输入–输出词元构成，来创建训练样本。这样，模型便能适应精确的任务要求，同时保留在大规模预训练中学到的共享表征。

数据构建与表征。 为支持统一的语音建模，所有音频–文本对均经过专门设计的数据整理流程处理，该流程优先考虑语义连贯性、说话人覆盖度和声学鲁棒性。

- **音频归一化与降噪：** 沿用先前工作 [15], , 我们采用动态范围压缩和基于峰值的音量归一化。具体而言，将波形按其最大绝对振幅重新缩放：

$$\tilde{x} = 0.6 \cdot \frac{x}{\max |x|},$$

  其中，x 表示原始波形。该归一化可确保不同话语间的振幅缩放保持一致，减轻录音级音量变化导致的分布偏移，并提高训练稳定性。
  此外，我们还使用了内部开发的噪声抑制模型，该模型基于多种真实环境干扰（如交通噪声、混响和重叠语音）进行训练，以提升信号清晰度。
- **说话人感知分段：** 集成的语音活动检测和说话人分离系统会联合检测语音片段，并将其分配给身份一致的说话人。虽然我们的实现依赖于内部开发的模块，但我们鼓励使用已有研究成果（例如， [7]）来实现类似功能。

[2] https://github.com/AutoArk/GPA

这样便能提取干净且说话人一致的话语，同时丢弃受串音或非语音事件干扰的片段。在此过程中，系统会自动过滤掉那些边界处的能量特征表明可能存在词中截断的话语。

- **高置信度转录：** 文本标签采用混合自动语音识别策略生成，该策略利用多个公开可用模型之间的一致性，包括 Faster-Whisper Large-V3[37], Canary-1B&Parakeet-TDT-0.6B-v3[41], Omnilingual ASR[43], 和 SeamlessM4T-v2-Large[12]。每个模型都会独立生成一个转录假设，随后由内部的置信度感知模块进行合并与评分。

  仅保留呈现出较强跨系统一致性的片段。实际上，一致性通过所有自动语音识别假设之间的平均成对词错误率（pWER）进行量化：

$$\mathrm{pWER} = \frac{1}{N(N-1)} \sum_{i \neq j} \mathrm{WER}(y_i, y_j),$$

  其中，$\{y_1, \ldots, y_N\}$ 表示 N 个自动语音识别模型的转录输出。按照先前工作的设置[15], , 保留 pWER 为 < 15 % 的片段作为最终文本标签。
- **韵律对齐的标点：** 利用 Montreal Forced Aligner 进行强制对齐后， [29], 词间停顿时长可用于优化标点：在分句之间存在 $\geq$300 ms 静音时插入逗号，而在流畅语段中（停顿时长为 <50 ms）则删除多余标点。由此得到的文本流在时间上有依据、在语言上连贯，可用于共同离散化到共享的语音文本令牌空间中。

3.2 模型性能

统一语音模型在实际部署中的效果，不仅取决于其准确率，也取决于推理流水线的效率和可扩展性。在本小节中，我们从架构和系统层面分析 GPA 的性能特征。

我们首先讨论 GPA 纯自回归设计的推理行为，重点介绍其适合流式处理的架构如何实现简单的批处理、高并发能力以及对计算资源的高效利用。随后，我们通过具有代表性的 TTS 和 ASR 流式工作负载下的延迟和吞吐量测量，定量验证这些特性，重点关注面向部署、针对低延迟和高内存效率进行优化的 **GPA-0.3B** 配置。

这些分析共同表明，GPA 非常适合大规模服务器部署，也适用于对延迟敏感且资源受限的执行场景。

3.2.1 推理

GPA 采用纯自回归推理范式，其中文本和语音均表示为离散令牌，并由单一共享的 Transformer 主干网络按序生成。所有语音任务——包括 ASR、TTS 和语音转换——均在同一解码框架内处理，无需切换模型或在推理时调用特定任务组件。任务行为则通过令牌顺序和输入提示隐式指定，从而形成统一且轻量的推理流水线。

这种设计天然适用于流式场景。随着解码的进行，词元会逐步输出，从而实现低延迟输出，无需显式的前瞻机制或多阶段缓冲。与依赖独立编码器、解码器或后处理阶段的传统流水线相比，GPA 始终保持单一、连续的生成过程，既降低了系统复杂性，也减少了协调开销。

从部署的角度来看，自回归解码的简洁性使 GPA 能够无缝集成到标准大语言模型推理框架中。实际上，这使其在并发工作负载下具有良好的吞吐量和延迟表现，尤其是在流式场景中。相同的推理流程可一致地应用于云端和边缘环境，因此 GPA 非常适用于可扩展的实时语音应用。

为促进可复现性和实际应用，我们随公开发布一并提供完整的推理与部署代码库。

3.2.2 延迟与吞吐量

我们报告了在不同请求并发度下，使用**GPA-0.3B** 模型进行文本转语音和自动语音识别时的流式延迟与吞吐量。对于文本转语音，我们测量首块时间（TTFC）和实时率（RTF），数值越低分别表示

并发能力	平均首块时间（毫秒）	P50 首块时间（毫秒）	P99 首块时间（毫秒）	平均实时因子	P50 实时因子	P99 实时因子	音频时长（秒）
1	258.8	258.8	258.8	0.197	0.197	0.197	6.44
5	385.0	394.7	396.2	0.218	0.217	0.248	6.76
10	544.6	564.2	566.7	0.282	0.301	0.313	6.49
20	977.8	977.9	982.9	0.470	0.490	0.538	7.19
40	1797.0	1736.4	2564.5	0.421	0.400	0.587	6.33
80	3786.4	4054.4	5415.8	0.763	0.763	1.096	6.32
160	9847.9	10239.9	14350.3	1.718	1.740	2.577	6.44

表 1：不同并发能力下的 TTS 流式基准测试。TTFC 表示首块时间（毫秒），RTF 表示实时率，音频时长表示平均生成音频时长（秒）。

并发能力	平均首个令牌时间（毫秒）	P50 首个令牌时间（毫秒）	P99 首个令牌时间（毫秒）	总平均值（毫秒）
1	157.5	157.5	157.5	190.9
5	394.1	393.7	395.9	400.0
10	589.6	721.3	723.3	598.1
20	1316.3	1495.6	1500.4	1317.8
40	2690.9	2678.3	2861.4	2693.7
80	3833.4	3961.3	4027.0	3845.1
160	5037.0	5689.3	6676.0	5044.0

表 2：自动语音识别流式 负载下的延迟与并发能力的关系。TTFT 表示首个令牌时间。所有数值均以 d 毫秒为单位 .

首段音频交付越快、语音合成速度越高。对于自动语音识别，我们报告首个令牌时间（TTFT）和端到端延迟。

所有结果均在流式设置下测量，展示了随着并发负载增加，系统响应能力和总体吞吐量的变化情况。按照官方部署脚本操作即可复现[3],，这反映了真实服务场景中的端到端性能，而非离线推理性能。

3.3 模型评测

我们在具有代表性的文本转语音和自动语音识别基准测试上评估 GPA，以考察其实际效果以及在不同模型规模下的可扩展性。实验采用了两个设计目标互补的模型变体。语音转换（VC）未在单独的表格中评估，因为它与文本转语音采用相同的自动评测指标，即可懂度相关错误率和说话人相似度。由于 GPA 中的语音转换与文本转语音的主要区别在于条件控制和令牌构成，而非底层的自回归生成机制，因此单独进行数值评估会造成冗余。

精简版 GPA-0.3B 模型针对内存效率和端侧部署进行了优化，面向资源受限的场景，在这些场景中，低资源占用和流式处理能力是首要考量。虽然它的目标并非媲美最强大的大规模系统，但在其参数规模范围内实现了有竞争力的性能。相比之下，GPA-3B 模型是面向能力的配置，展示了所提出的统一自回归框架在扩大模型容量时的有效性。

对于文本转语音评估（表 3），我们在中文测试集上报告字符错误率（CER）和说话人相似度（Sim），并按照既定评估协议，在英语基准测试上报告词错误率（WER）和说话人相似度 [3]。错误率越低表示可懂度越好，而相似度分数越高则表示说话人一致性越好。对于自动语音识别，我们在 Librispeech [30],和 AISHELL-1 [8],上报告标准的 WER 或 CER 指标，涵盖英语和普通话语音识别场景。

这些基准测试涵盖了不同任务和语言中的声学保真度与语言能力。尤其是自动语音识别的性能，取决于声学建模以及捕捉长程语言结构的能力。因此，不同模型规模下的识别准确率各不相同（表 4），反映了表征能力与部署效率之间的不同权衡。综合来看，这些结果刻画了 GPA 在不同任务、语言和模型配置下的表现。

[3]https://github.com/AutoArk/GPA

模型	开源	模型规模	Seed-zh[3]		Seed-en[3]	
			字符错误率（%）↓	相似度（%）↑	词错误率（%）↓	相似度（%）↑
多阶段或非自回归方法						
人类	-	-	1.26	75.5	2.14	73.4
Seed-TTS [3]	No	-	1.12	**79.6**	2.25	**76.2**
MiniMax-Speech [56]	No	-	0.83	78.3	**1.65**	69.2
F5-TTS [9]	Yes	0.3B	1.52	74.1	2.00	64.7
Cosyvoice2 [16]	Yes	0.5B	1.45	75.7	2.57	65.9
FireRedTTS2 [17]	Yes	1.5B	1.14	73.2	1.95	66.5
Index-TTS2 [62]	Yes	1.5B	1.03	76.5	2.23	70.6
VibeVoice-1.5B [33]	Yes	1.5B	1.16	74.4	3.04	68.9
VibeVoice-Realtime [33]	Yes	0.5B	-	-	2.05	63.3
HiggsAudio-v2	Yes	3B	1.50	74.0	2.44	67.7
VoxCPM [63]	Yes	0.5B	0.93	77.2	1.85	72.9
GLM-TTS [55]	Yes	1.5B	1.03	76.1	-	-
GLM-TTS RL [55]	Yes	1.5B	0.89	76.4	-	-
Fun-CosyVoice3-0.5B-2512 [15]	Yes	0.5B	1.21	78.0	2.24	71.8
Fun-CosyVoice3-0.5B-2512_RL [15]	Yes	0.5B	**0.81**	77.4	1.68	69.5
单阶段自回归方法						
Spark TTS [48]	Yes	0.5B	1.20	66.0	1.98	57.3
GPA-0.3B	Yes	0.3B	**0.95**	65.9	**1.51**	56.5
GPA-3B	Yes	3B	**0.84**	73.0	**1.31**	71.7

表 3：TTS评估表。结果分为多阶段（非自回归）和单阶段自回归方法。根据 [15]中的评估协议，我们报告中文（Seed-zh）的字符错误率（CER）和说话人相似度（相似度），以及英语（Seed-en）的词错误率（WER）和说话人相似度（相似度）。↓表示越低越好，↑表示越高越好。

模型	参数量	Librispeech test-clean（词错误率↓）	AISHELL-1（字符错误率↓）
参数量为 $<$ 0.5B的模型			
Whisper-S [37]	0.24B	3.13	–
GPA-0.3B	0.3B	8.88	4.50
参数量为 $\geq$ 0.5B 的模型			
Fun-ASR-nano[2]	0.8B	1.76	1.80
FireRed-ASR[52]	1.1B	1.84	0.54
GLM-ASR-nano[55]	1.5B	2.00	1.81
GLM-ASR-nano*[55]	1.5B	2.17	2.17
Whisper-L[37]	1.55B	1.82	4.72
Kimi-Audio[23]	–	1.32	0.71
Step-Audio2[50]	–	1.17	0.63
Seed-ASR[6]	–	1.58	0.68
Seed-ASR*[6]	–	2.80	1.63
Fun-ASR[2]	7.7B	1.51	1.22
GPA-3B	3.0B	2.52	1.93

表 4：在 Librispeech 和 AISHELL-1 上进行自动语音识别评估。Librispeech 报告词错误率（WER，%），AISHELL-1 报告字符错误率（CER，%）。鉴于自动语音识别中语言建模具有显著的规模扩展效应，结果按模型规模分组比较。基线结果取自已发表的报告[55, 58]，并遵循相同的评估协议。参数量未公开的模型列于表格底部。Seed-ASR* 的结果通过官方火山引擎 API 获取，而 GLM-ASR-nano* 的结果则使用开源检查点进行评估。

4 结论

本文介绍了 GPA，这是一种用于通用语音建模的统一自回归框架，可在单一基于大语言模型的架构中整合多种语音任务。通过将语音、文本和控制信号表示为离散令牌，并在共享的下一词元预测目标下对其进行建模，GPA 以统一的表述支持自动语音识别、文本转语音合成和语音编辑。这种统一设计减少了架构碎片化，并支持跨任务的高效训练与推理。

全面评估表明，GPA 在一系列语音基准测试中实现了具有竞争力的准确率，同时保持了出色的效率和可扩展性。这些结果验证了完全自回归、基于词元的方法在统一语音建模中的有效性，并凸显了其用于构建灵活且可部署的音频基础模型的潜力。

5 局限性与未来工作

尽管 GPA 展示了统一自回归框架用于多种语音任务的可行性，但仍存在一些局限性。首先，尽管共享的基于词元的表述简化了系统设计，但也引入了单一的建模瓶颈；与专用架构相比，这可能限制其在高度专业化任务上的峰值性能。其次，与其他大型自回归模型一样，推理成本和延迟会随序列长度增加，在未经进一步优化的情况下，这给超长文本或低延迟应用带来了挑战。第三，GPA-0.3B 模型的自动语音识别性能相对较弱，这主要是由容量限制造成的，而非统一自回归设计的根本局限性，表明扩展规模和进一步优化能够带来明显收益。

展望未来，GPA 统一的自回归形式天然支持在离散音频词元上进行强化学习。与整合独立声学模型、语言模型和特定任务模块的异构流水线不同，后者的各个组件之间往往难以明确分配信用；GPA 的单一下一词元预测目标则提供了清晰、一致的优化目标。这一特性为指标驱动的后训练打开了大门，使我们能够在统一框架内，直接针对指令遵循、感知质量和特定任务目标进行优化。

参考文献

[1] Keyu An、Qian Chen、Chong Deng、Zhihao Du、Changfeng Gao、Zhifu Gao、Yue Gu、Ting He、Hangrui Hu、Kai Hu、Shengpeng Ji、Yabin Li、Zerui Li、Heng Lu、Haoneng Luo、Xiang Lv、Bin Ma、Ziyang Ma、Chongjia Ni、Changhe Song、Jiaqi Shi、Xian Shi、Hao Wang、Wen Wang、Yuxuan Wang、Zhangyu Xiao、Zhijie Yan、Yexin Yang、Bin Zhang、Qinglin Zhang、Shiliang Zhang、Nan Zhao 和 Siqi Zheng。《Funaudiollm：面向人类与大语言模型自然交互的语音理解与生成基础模型》，2024年。[2] Keyu An、Yanni Chen、Zhigao Chen、Chong Deng、Zhihao Du、Changfeng Gao、Zhifu Gao、Bo Gong、Xiangang Li、Yabin Li、Ying Liu、Xiang Lv、Yunjie Ji、Yiheng Jiang、Bin Ma、Haoneng Luo、Chongjia Ni、Zexu Pan、Yiping Peng、Zhendong Peng、Peiyao Wang、Hao Wang、Haoxu Wang、Wen Wang、Wupeng Wang、Yuzhong Wu、Biao Tian、Zhentao Tan、Nan Yang、Bin Yuan、Jieping Ye、Jixing Yu、Qinglin Zhang、Kun Zou、Han Zhao、Shengkui Zhao、Jingren Zhou 和 Yanqiao Zhu。《Fun-ASR 技术报告》，2025年。[3] Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao、Mingqing Gong、Peisong Huang、Qingqing Huang、Zhiying Huang、Yuanyuan Huo、Dongya Jia、Chumin Li、Feiya Li、Hui Li、Jiaxin Li、Xiaoyang Li、Xingxing Li、Lin Liu、Shouda Liu、Sichao Liu、Xudong Liu、Yuchen Liu、Zhengxi Liu、Lu Lu、Junjie Pan、Xin Wang、Yuping Wang、Yuxuan Wang、Zhen Wei、Jian Wu、Chao Yao、Yifeng Yang、Yuanhao Yi、Junteng Zhang、Qidi Zhang、Shuo Zhang、Wenjie Zhang、Yang Zhang、Zilin Zhao、Dejian Zhong 和 Xiaobin Zhuang。《Seed-TTS：高质量、多用途语音生成模型系列》，2024年。[4] Junyi Ao、Rui Wang、Long Zhou、Chengyi Wang、Shuo Ren、Yu Wu、Shujie Liu、Tom Ko、Qing Li、Yu Zhang、Zhihua Wei、Yao Qian、Jinyu Li 和 Furu Wei。《Speecht5：面向口语语言处理的统一模态编码器-解码器预训练》，2021年。[5] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。《wav2vec 2.0：语音表征自监督学习框架》，2020年。[6] Ye Bai、Jingping Chen、Jitong Chen、Wei Chen、Zhuo Chen、Chuang Ding、Linhao Dong、Qianqian Dong、Yujiao Du、Kepan Gao、Lu Gao、Yi Guo、Minglun Han、Ting Han、Wenchao Hu、Xinying Hu、Yuxiang Hu、Deyu Hua、Lu Huang、Mingkun Huang、Youjia Huang、Jishuo Jin、Fanliu Kong、Zongwei Lan、Tianyu Li、Xiaoyang Li、Zeyang Li、Zehua Lin、Rui Liu、Shouda Liu、Lu Lu、Yizhou Lu、Jingting Ma、Shengtao Ma、Yulin Pei、Chen Shen、Tian Tan、Xiaogang Tian、Ming Tu、Bo Wang、Hao Wang、Yuping Wang、Yuxuan Wang、Hanzhang Xia、Rui Xia、Shuangyi Xie、Hongmin Xu、Meng Yang、Bihong Zhang、Jun Zhang、Wanyi Zhang、Yang Zhang、Yawei Zhang、Yijie Zheng 和 Ming Zou。《Seed-ASR：借助基于大语言模型的语音识别理解多样化语音及其语境》，2024年。[7] Hervé Bredin、Ruiqing Yin、Juan Manuel Coria、Gregory Gelly、Pavel Korshunov、Marvin Lavechin、Diego Fustes、Hadrien Titeux、Wassim Bouaziz 和 Marie-Philippe Gill。《pyannote.audio：用于说话人分离的神经网络构件》，2019年。[8] Hui Bu、Jiayu Du、Xingyu Na、Bengu Wu 和 Hao Zheng。《AISHELL-1：开源普通话语音语料库与语音识别基线》，2017年。[9] Yushen Chen、Zhikang Niu、Ziyang Ma、Keqi Deng、Chunhui Wang、Jian Zhao、Kai Yu 和 Xie Chen。《F5-TTS：利用流匹配生成流畅、忠实语音的"故事讲述者"》，2025年。[10] Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin、Chang Zhou 和 Jingren Zhou。《Qwen2-audio 技术报告》，2024年。[11] Yunfei Chu、Jin Xu、Xiaohuan Zhou、Qian Yang、Shiliang Zhang、Zhijie Yan、Chang Zhou 和 Jingren Zhou。《Qwen-Audio：通过统一的大规模音频语言模型推进通用音频理解》。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），2024年。[12] Seamless Communication、Loïc Barrault、Yu-An Chung、Mariano Cora Meglioli、David Dale、Ning Dong、Paul- Ambroise Duquenne、Hady Elsahar、Hongyu Gong、Kevin Heffernan、John Hoffman、Christopher Klaiber、Pengwei Li、Daniel Licht、Jean Maillard、Alice Rakotoarison、Kaushik Ram Sadagopan、Guillaume Wenzek、Ethan Ye、Bapi Akula、Peng-Jen Chen、Naji El Hachem、Brian Ellis、Gabriel Mejia Gonzalez、Justin Haaheim、Prangthip Hansanti、Russ Howes、Bernie Huang、Min-Jae Hwang、Hirofumi Inaguma、Somya Jain、Elahe Kalbassi、Amanda Kallet、Ilia Kulikov、Janice Lam、Daniel Li、Xutai Ma、Ruslan Mavlyutov、Benjamin Peloquin、Mohamed Ramadan、Abinesh Ramakrishnan、Anna Sun、Kevin Tran、Tuan Tran、Igor Tufanov、Vish Vogeti、Carleigh Wood、Yilin Yang、Bokai Yu、Pierre Andrews、Can Balioglu、Marta R. Costa-jussà、Onur Celebi、Maha Elbayad、Cynthia Gao、Francisco Guzmán、Justine Kao、Ann Lee、Alexandre Mourachko、Juan Pino、Sravya Popuri、Christophe Ropers、Safiyyah Saleem、Holger Schwenk、Paden Tomasello、Changhan Wang、Jeff Wang 和 Skyler Wang。《Seamlessm4t：大规模多语言、多模态机器翻译》，2023年。

[13] Wei Deng、Siyi Zhou、Jingchen Shu、Jinchao Wang 和 Lu Wang。Indextts：面向工业级应用的可控高效零样本文本转语音系统，2025年。[14] Zhihao Du、Qian Chen、Shiliang Zhang、Kai Hu、Heng Lu、Yexin Yang、Hangrui Hu、Siqi Zheng、Yue Gu、Ziyang Ma、Zhifu Gao 和 Zhijie Yan。Cosyvoice：基于监督语义标记的可扩展多语言零样本文本转语音合成器，2024年。[15] Zhihao Du、Changfeng Gao、Yuxuan Wang、Fan Yu、Tianyu Zhao、Hao Wang、Xiang Lv、Hui Wang、Chongjia Ni、Xian Shi、Keyu An、Guanrou Yang、Yabin Li、Yanni Chen、Zhifu Gao、Qian Chen、Yue Gu、Mengzhe Chen、Yafeng Chen、Shiliang Zhang、Wen Wang 和 Jieping Ye。Cosyvoice 3：通过扩展规模和后训练实现真实场景中的语音生成，2025年。[16] Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang、Fan Yu、Huadai Liu、Zhengyan Sheng、Yue Gu、Chong Deng、Wen Wang、Shiliang Zhang、Zhijie Yan 和 Jingren Zhou。Cosyvoice 2：基于大语言模型的可扩展流式语音合成，2024年。[17] Hao-Han Guo、Yao Hu、Kun Liu、Fei-Yu Shen、Xu Tang、Yi-Chen Wu、Feng-Long Xie、Kun Xie 和 Kai-Tuo Xu。Fireredtts：面向工业级生成式语音应用的基础文本转语音框架，2025年。[18] Haorui He、Zengqiang Shang、Chaoren Wang、Xuyuan Li、Yicheng Gu、Hua Hua、Liwei Liu、Chen Yang、Jiaqi Li、Peiyang Shi、Yuancheng Wang、Kai Chen、Pengyuan Zhang 和 Zhizheng Wu。Emilia：用于大规模语音生成的广泛、多语言且多样化的语音数据集，2024年。[19] Dongya Jia、Zhuo Chen、Jiawei Chen、Chenpeng Du、Jian Wu、Jian Cong、Xiaobin Zhuang、Chumin Li、Zhen Wei、Yuping Wang 和 Yuxuan Wang。Ditar：用于语音生成的扩散变换器自回归建模，2025年。[20] Zhijun Jia、Huaying Xue、Xiulian Peng 和 Yan Lu。转换并说话：以最少监督实现零样本口音转换。发表于第*32*届*ACM*国际多媒体会议论文集（*ACM MM*），第4446–4454页，2024年。[21] Mumin Jin、Prashant Serai、Jilong Wu、Andros Tjandra、Vimal Manohar 和 Qing He。保留音色的零样本多口音转换。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第1–5页，2023年。[22] Eugene Kharitonov、Damien Vincent、Zalán Borsos、Raphaël Marinier、Sertan Girgin、Olivier Pietquin、Matthew Sharifi、Marco Tagliasacchi 和 Neil Zeghidour。说、读与提示：以最少监督实现高保真文本转语音。发表于机器学习研究汇刊（*TMLR*），2023年。[23] KimiTeam、Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang、Zhengtao Wang、Chu Wei、Yifei Xin、Xinran Xu、Jianwei Yu、Yutao Zhang、Xinyu Zhou、Y. Charles、Jun Chen、Yanru Chen、Yulun Du、Weiran He、Zhenxing Hu、Guokun Lai、Qingcheng Li、Yangyang Liu、Weidong Sun、Jianzhou Wang、Yuzhi Wang、Yuefeng Wu、Yuxin Wu、Dongchao Yang、Hao Yang、Ying Yang、Zhilin Yang、Aoxiong Yin、Ruibin Yuan、Yutong Zhang 和 Zaida Zhou。Kimi-audio 技术报告，2025年。[24] Hugo Laurençon、Léo Tronchon、Matthieu Cord 和 Victor Sanh。构建视觉语言模型时哪些因素最重要？，2024年。[25] Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 和 Wei-Ning Hsu。Voicebox：大规模文本引导的多语言通用语音生成。神经信息处理系统进展（*NeurIPS*），第36卷，2023年。[26] Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 和 Wei-Ning Hsu。Voicebox：大规模文本引导的多语言通用语音生成。发表于神经信息处理系统进展（*NeurIPS*），第36卷，2023年。[27] Yuang Li、Yu Wu、Jinyu Li 和 Shujie Liu。通过提示大语言模型实现语音识别的零样本领域自适应。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第12436–12440页，2024年。[28] Ziyi Lin、Chris Liu、Renrui Zhang、Peng Gao、Longtian Qiu、Han Xiao、Han Qiu、Chen Lin、Wenqi Shao、Keqin Chen、Jiaming Han、Siyuan Huang、Yichi Zhang、Xuming He、Hongsheng Li 和 Yu Qiao。Sphinx：多模态大语言模型中权重、任务与视觉嵌入的联合混合，2023年。[29] Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。Montreal Forced Aligner：使用 Kaldi 进行可训练的文本与语音对齐。发表于*Interspeech*，第2017卷，第498–502页，2017年。[30] Aleksandr Meister、Matvei Novikov、Nikolay Karpov、Evelina Bakhturina、Vitaly Lavrukhin 和 Boris Ginsburg。Librispeech-pc：用于评估端到端自动语音识别模型标点与大小写能力的基准，2023年。

[31] Seyed Hamidreza Mohammadi 和 Alexander Kain。语音转换系统概述。*Speech Communi-cation*，第88卷：65–82，2017年。[32] William Peebles 和 Saining Xie。基于变换器的可扩展扩散模型，2023年。[33] Zhiliang Peng、Jianwei Yu、Wenhui Wang、Yaoyao Chang、Yutao Sun、Li Dong、Yi Zhu、Weijiang Xu、Hangbo Bao、Zehua Wang、Shaohan Huang、Yan Xia 和 Furu Wei。Vibevoice 技术报告，2025年。[34] Adam Polyak、Yossi Adi、Jade Copet、Eugene Kharitonov、Kushal Lakhotia、Wei-Ning Hsu、Abdelrahman Mohamed 和 Emmanuel Dupoux。基于离散解耦自监督表示的语音再合成。发表于*Interspeech* 论文集，第3615–3619页，2021年。[35] Tianhua Qi、Wenming Zheng、Cheng Lu、Yuan Zong 和 Hailun Lian。Pavits：探索用于端到端情感语音转换的韵律感知 VITS。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第12697–12701页，2024年。[36] Kaizhi Qian、Yang Zhang、Shiyu Chang、David Cox 和 Mark Hasegawa-Johnson。通过三重信息瓶颈实现无监督语音分解。发表于第*37*届国际机器学习会议（*ICML*）论文集，第7836–7846页，2020年。[37] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别，2022年。[38] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。国际机器学习会议（*ICML*），2023年。[39] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督的多任务学习者。*OpenAI* 博客，第1卷（8），2019年。[40] Paul K. Rubenstein、Chulayuth Asawaroengchai、Duc Dung Nguyen、Ankur Bapna、Zalán Borsos、Félix de Chau- mont Quitry、Peter Chen、Dalia El Badawy、Wei Han、Eugene Kharitonov、Hannah Muckenhirn、Dirk Padfield、James Qin、Danny Rozenberg、Tara Sainath、Johan Schalkwyk、Matt Sharifi、Michelle Tadmor Ramanovich、Marco Tagliasacchi、Alexandru Tudor、Mihajlo Velimirovi´c、Damien Vincent、Jiahui Yu、Yongqiang Wang、Vicky Zayats、Neil Zeghidour、Yu Zhang、Zhishuai Zhang、Lukas Zilka 和 Christian Frank。Audiopalm：能够说和听的大语言模型。发表于*Interspeech* 论文集，2023年。[41] Monica Sekoyan、Nithin Rao Koluguri、Nune Tadevosyan、Piotr Zelasko、Travis Bartley、Nikolay Karpov、Jagadeesh Balam 和 Boris Ginsburg。Canary-1b-v2 与 Parakeet-TDT-0.6B-v3：用于多语言自动语音识别和自动语音翻译的高效高性能模型，2025年。[42] Yakun Song、Zhuo Chen、Xiaofei Wang、Ziyang Ma 和 Xie Chen。Ella-v：基于对齐引导的序列重排实现稳定的神经编解码器语言建模。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第10281–10285页，2024年。[43] Omnilingual ASR 团队、Gil Keren、Artyom Kozhevnikov、Yen Meng、Christophe Ropers、Matthew Setzler、Skyler Wang、Ife Adebara、Michael Auli、Can Balioglu、Kevin Chan、Chierh Cheng、Joe Chuang、Caley Droof、Mark Duppenthaler、Paul-Ambroise Duquenne、Alexander Erben、Cynthia Gao、Gabriel Mejia Gonzalez、Kehan Lyu、Sagar Miglani、Vineel Pratap、Kaushik Ram Sadagopan、Safiyyah Saleem、Arina Turkatenko、Albert Ventayol-Boada、Zheng-Xin Yong、Yu-An Chung、Jean Maillard、Rashel Moritz、Alexandre Mourachko、Mary Williamson 和 Shireen Yates。Omnilingual ASR：面向 1600+ 语言的开源多语言语音识别，2025年。[44] Hawau Olamide Toyin、Hao Li 和 Hanan Aldarmaki。Sttatts：统一的语音转文本与文本转语音模型，2024年。[45] Chengyi Wang、Sanyuan Chen、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li、Lei He、Sheng Zhao 和 Furu Wei。神经编解码器语言模型是零样本文本转语音合成器。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第1–5页，2023年。[46] Chengyi Wang、Sanyuan Chen、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li、Lei He、Sheng Zhao 和 Furu Wei。神经编解码器语言模型是零样本文本转语音合成器。发表于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第1–5页，2023年。[47] Mingqiu Wang、Wei Han、Izhak Shafran、Zelin Wu、Chung-Cheng Chiu、Yuan Cao、Yongqiang Wang、Nanxin Chen、Yu Zhang、Hagen Soltau、Paul Rubenstein、Lukas Zilka、Dian Yu、Zhong Meng、Golan Pundak、Nikhil Siddhartha、Johan Schalkwyk 和 Yonghui Wu。SLM：弥合语音与文本基础模型之间的细微差距。发表于*IEEE* 自动语音识别与理解研讨会（*ASRU*），2023年。

[48] Xinsheng Wang、Mingqi Jiang、Ziyang Ma、Ziyu Zhang、Songxiang Liu、Linqin Li、Zheng Liang、Qixi Zheng、Rui Wang、Xiaoqin Feng、Weizhen Bian、Zhen Ye、Sitong Cheng、Ruibin Yuan、Zhixian Zhao、Xinfa Zhu、Jiahao Pan、Liumeng Xue、Pengcheng Zhu、Yunlin Chen、Zhifei Li、Xie Chen、LeiXie、Yike Guo 和 Wei Xue。《Spark-tts：一种采用单流解耦语音令牌的高效大语言模型文本转语音模型》，2025年。

[49] Yuxuan Wang、R. J. Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J. Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio、Quoc V. Le、Yannis Agiomyrgiannakis、Rob Clark 和 Rif A. Saurous。《Tacotron：迈向端到端语音合成》。载于*Interspeech* 论文集，第 4006–4010 页，2017年。

[50] Boyong Wu、Chao Yan、Chen Hu、Cheng Yi、Chengli Feng、Fei Tian、Feiyu Shen、Gang Yu、Haoyang Zhang、Jingbei Li、Mingrui Chen、Peng Liu、Wang You、Xiangyu Tony Zhang、Xingyuan Li、Xuerui Yang、Yayue Deng、Yechang Huang、Yuxin Li、Yuxin Zhang、Zhao You、Brian Li、Changyi Wan、Hanpeng Hu、Jiangjie Zhen、Siyu Chen、Song Yuan、Xuelin Zhang、Yimin Jiang、Yu Zhou、Yuxiang Yang、Bingxin Li、Buyun Ma、Changhe Song、Dongqing Pang、Guoqiang Hu、Haiyang Sun、Kang An、Na Wang、Shuli Gao、Wei Ji、Wen Li、Wen Sun、Xuan Wen、Yong Ren、Yuankai Ma、Yufan Lu、Bin Wang、Bo Li、Changxin Miao、Che Liu、Chen Xu、Dapeng Shi、Dingyuan Hu、Donghang Wu、Enle Liu、Guanzhe Huang、Gulin Yan、Han Zhang、Hao Nie、Haonan Jia、Hongyu Zhou、Jianjian Sun、Jiaoren Wu、Jie Wu、Jie Yang、Jin Yang、Junzhe Lin、Kaixiang Li、Lei Yang、Liying Shi、Li Zhou、Longlong Gu、Ming Li、Mingliang Li、Mingxiao Li、Nan Wu、Qi Han、Qinyuan Tan、S haoliang Pang、Shengjie Fan、Siqi Liu、Tiancheng Cao、Wanying Lu、Wenqing He、Wuxun Xie、Xu Zhao、Xueqi Li、Yanbo Yu、Yang Yang、Yi Liu、Yifan Lu、Yilei Wang、Yuanhao Ding、Yuanwei Liang、Yuanwei Lu、Yuchu Luo、Yuhe Yin、Yumeng Zhan、Yuxiang Zhang、Zidong Yang、Zixin Zhang、Binxing Jiao、Daxin Jiang、Heung-Yeung Shum、Jiansheng Chen、Jing Li、Xiangyu Zhang 和 Yibo Zhu。《Step-audio 2 技术报告》，2025年。

[51] Jian Wu、Yashesh Gaur、Zhuo Chen、Long Zhou、Yimeng Zhu、Tianrui Wang、Jinyu Li、Shujie Liu、Bo Ren、Linquan Liu 和 Yu Wu。关于用于语音转文本和大语言模型集成的仅解码器架构。载于*IEEE* 自动语音识别与理解研讨会（*ASRU*），第 1–8 页，2023年。

[52] Kai-Tuo Xu、Feng-Long Xie、Xu Tang 和 Yao Hu。Fireredasr：从编码器-解码器到大语言模型集成的开源工业级普通话语音识别模型，2025年。

[53] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。Qwen3 技术报告，2025年。

[54] Dongchao Yang、Jinchuan Tian、Xu Tan、Rongjie Huang、Songxiang Liu、Haohan Guo、Xuankai Chang、Jiatong Shi、Sheng Zhao、Jiang Bian、Zhou Zhao、Xixin Wu 和 Helen M. Meng。《Uniaudio：借助大语言模型实现通用音频生成》。载于第 *41* 届国际机器学习会议（*ICML*）论文集，2024 年。

[55] Aohan Zeng、Zhengxiao Du、Mingdao Liu、Kedong Wang、Shengmin Jiang、Lei Zhao、Yuxiao Dong 和 Jie Tang。《GLM-4-Voice：迈向智能且类人化的端到端语音聊天机器人》，2024年。[56] Bowen Zhang、Congchao Guo、Geng Yang、Hang Yu、Haozhe Zhang、Heidi Lei、Jialong Mai、Junjie Yan、Kaiyue Yang、Mingqi Yang、Peikai Huang、Ruiyang Jin、Sitan Jiang、Weihua Cheng、Yawei Li、Yichen Xiao、Yiying Zhou、Yongmao Zhang、Yuan Lu 和 Yucen He。《MiniMax-Speech：采用可学习说话人编码器的内在零样本文本转语音》，2025年。[57] Mingyang Zhang、Xin Wang、Fuming Fang、Haizhou Li 和 Junichi Yamagishi。《使用多源 Tacotron 和 WaveNet 的文本转语音与语音转换联合训练框架》，2019年。[58] Qinglin Zhang、Luyao Cheng、Chong Deng、Qian Chen、Wen Wang、Siqi Zheng、Jiaqing Liu、Hai Yu、Chaohong Tan、Zhihao Du 和 Shiliang Zhang。《OmniFlatten：用于无缝语音对话的端到端 GPT 模型》，2025年。[59] Xueyao Zhang、Xiaohui Zhang、Kainan Peng、Zhenyu Tang、Vimal Manohar、Yingru Liu、Jeff Hwang、Dangna Li、Yuhao Wang、Julian Chan、Yuan Huang、Zhizheng Wu 和 Mingbo Ma。《Vevo：通过自监督解耦实现可控的零样本语音模仿》，2025年。[60] Guanlong Zhao、Shaojin Ding 和 Ricardo Gutierrez-Osuna。《无需参考语音转换外国口音语音》。*IEEE/ACM* 音频、语音与语言处理汇刊，29：1167–1181，2021年。

[61] Kun Zhou、Berrak Sisman、Rajib Rana、Björn W. Schuller 和 Haizhou Li。《情感强度及其在情感语音转换中的控制》。*IEEE/ACM 音频、语音与语言处理汇刊*，31：1836–1848，2023年。[62] Siyi Zhou、Yiquan Zhou、Yi He、Xun Zhou、Jinchao Wang、Wei Deng 和 Jingchen Shu。《Indextts2：在情感表现力和时长控制方面取得突破的自回归零样本文本转语音》，2025年。[63] Yixuan Zhou、Guoyang Zeng、Xin Liu、Xiang Li、Renjie Yu、Ziyang Wang、Runchuan Ye、Weiyue Sun、Jiancheng Gui、Kehan Li、Zhiyong Wu 和 Zhiyuan Liu。《VoxCPM：无需分词器的文本转语音，实现上下文感知语音生成和逼真语音克隆》，2025年。[64] Mateusz Łajszczak、Guillermo Cámbara、Yang Li、Fatih Beyhan、Arent van Korlaar、Fan Yang、Arnaud Joly、Álvaro Martín-Cortinas、Ammar Abbas、Adam Michalski、Alexis Moinet、Sri Karlapati、Ewa Muszy´nska、Haohan Guo、Bartosz Putrycz、Soledad López Gambino、Kayeon Yoo、Elena Sokolova 和 Thomas Drugman。《Base TTS：构建基于十万小时数据的十亿参数文本转语音模型的经验》，2024年。