arXiv:2301.02111v1 [cs.CL] 5 Jan 2023

神经编解码器语言模型是零样本文本转语音合成器

Chengyi Wang* **Sanyuan Chen*** **Yu Wu*** **Ziqiang Zhang Long Zhou**
ShujieLiuZhuo Chen Yanqing Liu Huaming Wang JinyuLi Lei He Sheng Zhao Furu
Wei微软https://github.com/microsoft/unilm

摘要

我们提出了一种用于文本转语音合成（TTS）的语言建模方法。具体而言，我们使用现成的神经音频编解码模型生成的离散码，训练一个神经编解码器语言模型（称为 VALL-E），并将文本转语音视为条件语言建模任务，而非以往工作中的连续信号回归。在预训练阶段，我们将文本转语音训练数据扩增至 6 万小时的英语语音，规模是现有系统的数百倍。VALL-E 展现出上下文学习能力，仅凭未见说话人的 3秒注册录音作为声学提示，便可合成高质量的个性化语音。实验结果表明，在语音自然度和说话人相似度方面，VALL-E 显著优于当前最先进的零样本文本转语音系统。此外，我们发现，VALL-E 在合成时能够保留声学提示中的说话人情绪和声学环境。我们的演示请见
https://aka.ms/valle。

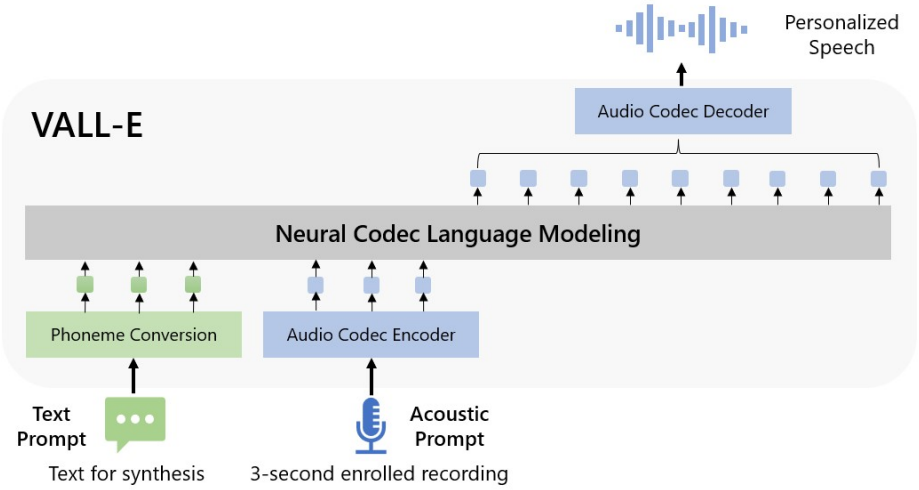


图 1：VALL-E 概览。与之前的流程（例如，音素 → 梅尔频谱图→ 波形）不同，VALL-E 的流程是音素 → 离散编码 → 波形。VALL-E 根据音素和声学编码提示生成离散音频编解码器代码，这些代码对应于目标内容和说话人的声音。VALL-E 直接支持各种语音合成应用，例如零样本文本转语音、语音编辑，以及与 GPT- 等其他生成式人工智能模型相结合的内容创作3 [Brown 等人， 2020]。

*这些作者对本研究的贡献相同。通信联系：{yuwu1,shujliu,fuwei}@microsoft.com

1 引言

过去十年间，随着神经网络和端到端建模的发展，语音合成取得了重大突破。目前，级联式文本转语音（TTS）系统 [Shen 等人，2018年，Ren 等人，2019年，Li 等人， 2019] 通常采用一套处理流程，其中包含声学模型和声码器，并以梅尔频谱图作为中间表示。尽管先进的文本转语音系统能够合成单个或多个说话人的高质量语音 [Liu 等人，2022年，Kim 等人，2021], 但仍需要来自录音棚的高质量干净数据。通过互联网爬取的大规模数据无法满足这一要求，且总会导致性能下降。由于训练数据相对较少，当前文本转语音系统仍存在泛化能力差的问题。在零样本场景中，面对未见说话人时，说话人相似度和语音自然度都会大幅下降。为解决零样本文本转语音问题，现有工作采用说话人自适应 [Chen 等人，2019年，Wang 等人， 2020] 和说话人编码 [Arik 等人，2018年，Casanova 等人，2022年] 等方法，但这些方法需要额外的微调、复杂的预先设计特征或繁重的结构工程。

与其针对这一问题设计复杂且专门的网络，不如尽可能使用大量且多样化的数据来训练模型；文本合成领域的成功为这一思路提供了启发 [Brown 等，2020，Chowdhery 等， 2022]。近年来，随着文本语言模型所用数据量的增加，其性能显著提升：从 16GB 未压缩文本 [Devlin 等， 2019], 到 160GB [Liu 等， 2019], 再到 570GB [Brown 等， 2020], ，最终达到约 1TB [Chowdhery 等， 2022]。为将这一成功经验应用于语音合成领域，我们提出了 VALL-E，这是首个利用大规模、多样化多说话人语音数据的基于语言模型的文本转语音框架。如图 1 所示，为了合成个性化语音（例如零样本文本转语音），VALL-E 根据 3秒注册录音的声学词元和音素提示，生成相应的声学词元；两者分别约束说话人和内容信息。最后，使用相应的神经编解码器解码器，将生成的声学词元合成为最终波形 [Défossez 等人， 2022]。从音频编解码模型中提取的离散声学词元，使我们能够将文本转语音视为条件编解码器语言建模；此外，还可将先进的基于提示的大模型技术（如 GPT 中的技术 [Brown 等， 2020]）应用于文本转语音任务。在推理过程中采用不同的采样策略，声学词元还使我们能够生成多样化的文本转语音合成结果。

我们使用 LibriLight 训练 VALL-E [Kahn 等人， 2020], 这是一个包含 6 万小时英语语音、涵盖 7000 多名独立说话人的语料库。原始数据仅包含音频，因此我们采用语音识别模型生成转录文本。与以往的文本转语音训练数据集（如 LibriTTS [Zen 等人） 相比， 2019], 我们的数据包含更多含噪语音和不准确的转录文本，但提供了多样化的说话人和韵律。我们认为，所提出的方法能够抵抗噪声，并通过利用大规模数据实现良好的泛化能力。值得注意的是，现有文本转语音系统通常使用数十小时的单说话人数据或数百小时的多说话人数据进行训练，其规模比 VALL-E 小数百倍以上。表 1 总结了 VALL-E 的创新之处：这是一种用于文本转语音的语言模型方法，以音频编解码器代码作为中间表示，利用大规模且多样化的数据，从而具备强大的上下文学习能力。

表1：VALL-E 与当前级联文本转语音系统的比较。

	当前系统	VALL-E
中间表示	梅尔频谱图	音频编解码器代码
目标函数	连续信号回归	语言模型
训练数据	$\leq$ 600小时	6万小时
上下文学习	✗	✓

我们在 LibriSpeech [Panayotov 等人， 2015] 以及 VCTK [Veaux 等人， 2016]数据集上评估 VALL-E，这些数据集中的所有测试说话人均未出现在训练语料库中。就语音自然度和说话人相似度而言，VALL-E 显著优于当前最先进的零样本文本转语音系统 [Casanova 等人，2022年]，在 LibriSpeech 上的比较平均意见分（CMOS）提高了 +00.12，相似度平均意见分（SMOS）提高了 +00.93。VALL-E 在 VCTK 上也优于基线，SMOS 提高了+00.11， +0CMOS 提高了 0.23。面对真实标注，它甚至取得了 +00.04 的 CMOS 分数，表明在 VCTK 上，未见说话人的合成语音与人类录音一样自然。此外，定性分析表明，VALL-E 能够针对相同文本和目标说话人合成多样化的输出，这

可能有利于语音识别任务中的伪数据创建。我们还发现，VALL-E 能够保留声学提示中的声学环境（例如混响）和情绪（例如愤怒）。

总而言之，我们作出以下贡献。

- 我们提出了 VALL-E，这是首个具备与 GPT-3 同等强大上下文学习能力的文本转语音框架。该框架将文本转语音视为语言模型任务，并以音频编解码器代码作为中间表示，取代传统的梅尔频谱图。它具备上下文学习能力，并支持基于提示的零样本文本转语音方法；与以往工作不同，该方法不需要额外的结构工程、预先设计的声学特征或微调。
- 我们利用海量半监督数据，在说话人维度上构建了一个通用的文本转语音系统，这表明，简单地扩展半监督数据规模对于文本转语音的重要性一直被低估。
- 对于相同的输入文本，VALL-E 能够提供多样化的输出，并保留声学提示中的声学环境和说话人情绪。
- 我们验证了 VALL-E 在零样本场景下通过提示能够合成具有高说话人相似度的自然语音。评测结果表明，在 LibriSpeech 和 VCTK 上，VALL-E 的表现显著优于当前最先进的零样本文本转语音系统。

我们鼓励读者在演示页面上收听我们的样例https://aka.ms/vall e.

2 相关工作

零样本文本转语音：当前的文本转语音方法可分为级联式和端到端方法。级联文本转语音系统 [Shen 等人，2018年，Ren 等人，2019年，Li 等人， 2019] 通常采用由声学模型和声码器组成的流水线，并以梅尔频谱图作为中间表示。为了解决声码器的不足，研究者提出了端到端文本转语音模型 [Kim 等人，2021年，Liu 等人， 2022] ，以联合优化声学模型和声码器。在实际场景中，人们非常希望能够利用少量注册录音，将文本转语音系统定制为任意声音。因此，零样本多说话人文本转语音技术日益受到关注，其中大多数研究基于级联文本转语音系统开展。作为先行者，Arik 等人 [2018] 提出了说话人自适应和说话人编码方法。在说话人自适应方向，后续研究 [Chen 等人，2019年，Wang 等人，2020年，Chen 等人， 2021] 致力于利用更少的目标说话人数据和说话人特定参数来提高自适应效率。Huang 等人 [2022] 将元学习应用于说话人自适应，只需 5-shot 即可构建性能良好的系统。与此同时，基于说话人编码的方法近年来取得了长足进展。基于说话人编码的系统由说话人编码器和文本转语音组件构成，其中说话人编码器可以在说话人验证任务上进行预训练 [Jia 等人， 2018]。Jia 等人 [2018] 和 Arik 等人 [2018], 的实验表明，对于域内说话人，模型只需 3 秒注册录音即可生成高质量输出。为提升未见说话人的合成质量，可以采用先进的说话人嵌入模型 [Cai 等人， 2018] ，但根据 Tan 等人 [2021]的研究，这仍不理想。另一种方法是设计先进但复杂的说话人编码器 [Wu 等人， 2022]。基于扩散模型的文本转语音方法 [Popov 等人，2021 年，Kim 等人， 2022] 也被扩展到零样本文本转语音领域[Kang 等人， 2022] ，并取得了良好效果。与以往工作相比 [Ren 等人，2019年，Du 等人，2022], 我们的工作沿用了级联式文本转语音路线，但首次采用音频编解码器代码作为中间表示。这是首个具备强大上下文学习能力的模型，其能力可媲美 GPT-3，且无需微调、预先设计的特征或复杂的说话人编码器。

语音生成式预训练模型：自监督学习在语音理解领域得到了广泛研究 [Baevski 等人，2020b, Hsu 等人，2021，Chen 等人， 2022] 以及语音到语音生成领域 [Lakhotia 等人，2021，Borsos 等人， 2022]。在语音到语音生成领域，一个热门课题是如何在无文本场景下合成语音。GSLM [Lakhotia 等人， 2021]提出基于 HuBERT 编码 [Hsu 等人， 2021], 合成语音，而 Polyak 等人[2021] 则通过结合 HuBERT 编码、VQVAE 编码和说话人编码器来提升性能。AudioLM [Borsos 等人， 2022] 采用类似的方法，但使用音频编解码器 [Zeghidour 等人，2022] 并结合语义编码来合成语音。值得注意的是，AudioLM 能够基于音频编解码器合成语音，而无需训练 HifiGAN 等额外的声码器[Kong 等人， 2020]。AudioLM 是语音到语音模型，而 VALL-E 是文本转语音模型，因此

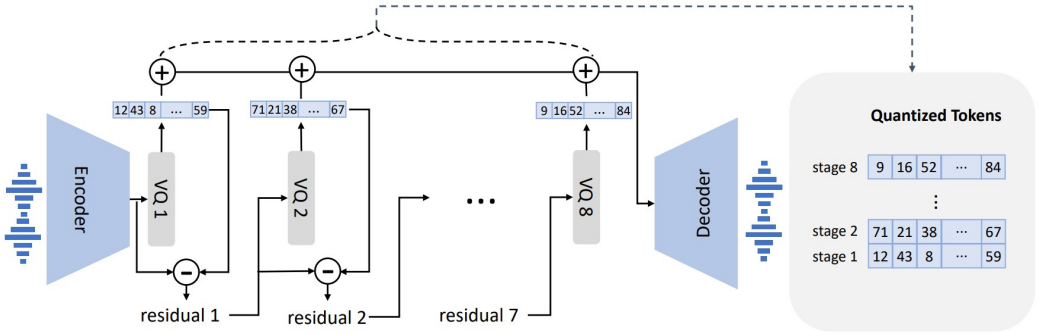


图2：重新审视神经音频编解码模型。由于采用了残差向量量化，第一量化器在重建中发挥着最重要的作用，其他量化器的影响则逐渐减小。

我们可以明确控制语音合成中的内容。另一个方向是将预训练应用于神经文本转语音。Chung 等人 [2018] 通过自回归梅尔频谱预测对文本转语音中的语音解码器进行预训练。在Ao 等人 [2022], 提出了统一模态编码器-解码器框架 SpeechT5，该框架能够利用无标签语音和文本数据，对文本转语音模型的所有组件进行预训练。Tjandra 等人 [2019] 使用 VQVAE 模型将无标签语音量化为离散令牌 [van den Oord 等人， 2017], 并使用令牌到语音序列训练模型。他们证明，预训练模型只需少量真实数据即可进行微调。Bai 等人 [2022] 提出了在梅尔频谱图上进行掩码与重建的方法，并展示了该方法在语音编辑和合成方面的更优性能。此前的文本转语音预训练工作使用的数据不足 1K 小时，而 VALL-E 则使用 60K 小时数据进行预训练。此外，VALL-E 首次使用音频编解码器代码作为中间表示，并在零样本文本转语音中展现出上下文学习能力。

3 背景：语音量化

由于音频通常以 16 位整数值序列的形式存储，因此生成模型需要在每个时间步输出 $2^{16} = 65$, 536 个概率，才能合成原始音频。此外，音频采样率超过一万，导致序列长度极长，使原始音频合成更难实现。为此，需要对语音进行量化，以压缩整数值和序列长度。 μ-law 变换可将每个时间步量化为 256 个值，并重建高质量的原始音频。它广泛应用于语音生成模型，例如 WaveNet[van den Oord 等人， 2016], 但由于序列长度没有缩短，推理速度仍然较慢。近年来，向量量化广泛应用于自监督语音模型的特征提取，例如 vq-wav2vec [Baevski 等人，2020a]和 HuBERT [Hsu 等人， 2021]。后续研究 [Lakhotia 等人，2021，Du 等人， 2022] 表明，自监督模型生成的编码也能重建内容，且推理速度快于 WaveNet。然而，说话人身份信息已被舍弃，重建质量也较低 [Borsos 等人， 2022]。AudioLM [Borsos 等人，2022] 在自监督模型生成的 k 均值词元和神经编解码器模型生成的声学词元上训练语音到语音语言模型，从而实现高质量的语音到语音生成。

本文沿用 AudioLM [Borsos 等人的做法， 2022] 利用神经编解码模型将语音表示为离散令牌。为了通过网络传输压缩音频，编解码模型能够将波形编码为离散声学码，并在训练时未见过该说话人的情况下重建高质量波形。与传统音频编解码方法相比，基于神经网络的编解码器在低比特率下表现显著更好；我们认为量化令牌包含了关于说话人和录音条件的充分信息。与其他量化方法相比，音频编解码器具有以下优势：1）它包含丰富的说话人信息和声学信息，与 HuBERT 编码相比，能够在重建时保留说话人身份 [Hsu 等人， 2021]。2）现成的编解码器解码器可以将离散令牌转换为波形，无需像基于向量量化的方法那样，额外进行基于频谱的声码器训练 [Du 等人， 2022]。3）它能够缩短时间步序列以提升效率，从而解决 μ 律变换中的问题 μ-law 变换 [van den Oord 等人， 2016]。

我们采用预训练的神经音频编解码器模型 EnCodec [Défossez 等人，2022], 作为分词器。EnCodec 是一种卷积编码器-解码器模型，其输入和输出均为 24 kHz 音频，且支持可变比特率。对于 24 kHz 的输入波形，编码器以 75 Hz 的频率生成嵌入，相当于将采样率降低了 320 倍。每个嵌入都通过残差向量量化（RVQ）进行建模，如图 2 所示，我们选用 8 个层级量化器，每个量化器包含 1024 个条目。此配置对应于 EnCodec 在 6K 比特率下对 24 kHz 音频进行重建。在此设置下，给定一个 10 秒的波形，其离散表示是一个包含 750×8 个条目的矩阵，其中 $750 = \frac{24,000 \times 10}{320}$ 表示下采样后的时间步，8 表示量化器的数量。也可以选择其他比特率设置。比特率越高，量化器数量越多，重建质量也越好。例如，如果选择 12K 比特率的 EnCodecc，则需要 16 个量化器，10 秒的波形对应一个包含 750×16 个条目的矩阵。使用所有量化器生成的离散码后，EnCodec 的卷积解码器会生成实值嵌入，并以 24 kHz 重建波形。

4 VALL-E

4.1 问题表述：将文本转语音视为条件编解码器语言建模

给定数据集 $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}$，其中 $\mathbf{y}$ 是音频样本，$\mathbf{x} = \{x_0, x_1, \ldots, x_L\}$ 是其对应的音素转写，我们使用预训练的神经编解码器模型将每个音频样本编码为离散声学码，记作 $\text{Encodec}(\mathbf{y}) = \mathbf{C}^{T \times 8}$，其中 $\mathbf{C}$ 表示二维声学码矩阵，T 表示下采样后的话语长度。每个声学码矩阵的行向量 $\mathbf{c}_{t,:}$ 表示帧 t 对应的八个声学码，而每个声学码矩阵的列向量 $\mathbf{c}_{:,j}$ 表示来自第 j 个码本的码序列，其中$j \in \{1, \ldots, 8\}$。经过量化后，神经编解码器解码器能够重建波形，记作 $\text{Decodec}(\mathbf{C}) \approx \hat{\mathbf{y}}$。

零样本文本转语音要求模型能够为未见说话人合成高质量语音。在本研究中，我们将零样本文本转语音视为条件编解码器语言建模任务。我们训练神经语言模型，根据音素序列 $\mathbf{x}$ 和声学提示矩阵 $\tilde{\mathbf{C}}^{T' \times 8}$ 生成声学码矩阵 $\mathbf{C}$，其优化目标为 $\max p(\mathbf{C}|\mathbf{x}, \tilde{\mathbf{C}})$。其中，$\tilde{\mathbf{C}}$ 由同一神经编解码器模型对注册录音进行编码得到。我们希望神经语言模型能够分别从音素序列和声学提示中提取内容和说话人信息。推理时，给定音素序列和一段未见说话人的 3秒注册录音，训练好的语言模型首先估计出包含相应内容并体现该说话人声音的声学码矩阵。随后，神经编解码器解码器合成高质量语音。

4.2 训练：条件编解码器语言建模

神经语音编解码器模型使我们能够处理离散音频表示。由于神经编解码器模型采用残差量化，令牌具有层次结构：来自前序量化器的令牌可还原说话人身份等声学属性，而后续量化器则学习精细的声学细节。每个量化器都经过训练，以建模前序量化器产生的残差。受此启发，我们以层次化的方式设计了两个条件语言模型。

对于第一量化器产生的离散令牌 $\mathbf{c}_{:,1}$，我们训练一个自回归（AR）仅解码器语言模型。该模型以音素序列$\mathbf{x}$ 和声学提示 $\tilde{\mathbf{C}}_{:,1}$为条件，形式化表示为

$$p(\mathbf{c}_{:,1}|\mathbf{x}, \tilde{\mathbf{C}}_{:,1}; \theta_{AR}) = \prod_{t=0}^{T} p(\mathbf{c}_{t,1}|\mathbf{c}_{<t,1}, \tilde{\mathbf{c}}_{:,1}, \mathbf{x}; \theta_{AR}) \tag{1}$$

由于 VALL-E 是仅解码器语言模型，$\tilde{\mathbf{c}}_{:,1}$ 与 $\mathbf{c}_{:,1}$ 的拼接构成一个完整序列，训练时我们不对二者加以区分，也不插入特定令牌。推理时，仅预测$\mathbf{c}_{:,1}$，并将前缀$\tilde{\mathbf{c}}_{:,1}$ 作为已知输入。

对于从第二个量化器到最后一个量化器所生成的离散令牌，$\mathbf{c}_{:,j\in[2,8]}$，我们训练了一个非自回归（NAR）语言模型。由于这些令牌无法以非自回归方式相互访问，为了约束说话人身份，我们使用声学提示矩阵 C 作为声学提示。因此，模型

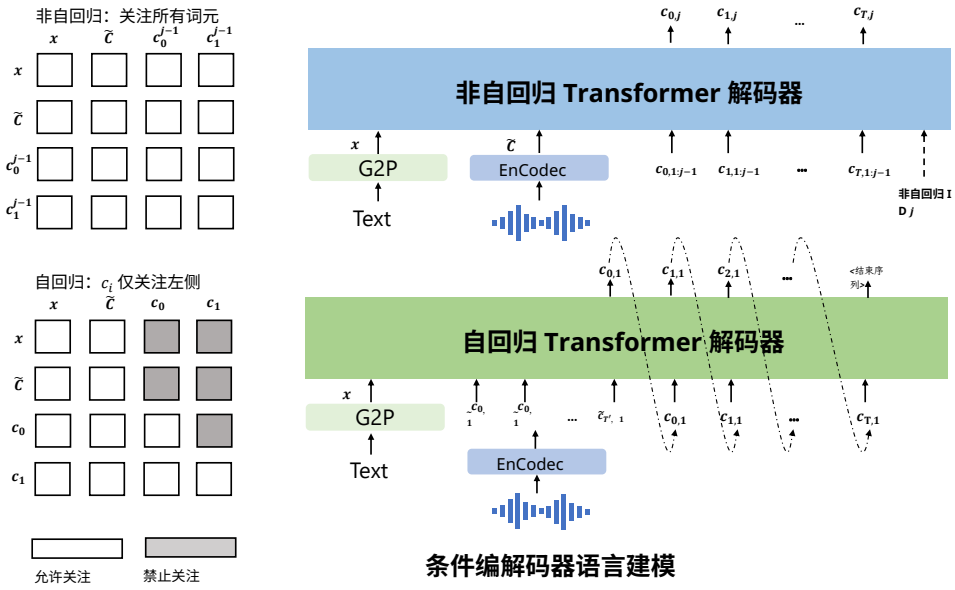


图3：条件编解码器语言建模的结构，该结构以分层方式构建。在实践中，非自回归解码器将被调用七次，以生成七个量化器中的编码。

以音素序列$\mathbf{x}$、声学提示 $\mathbf{C}$ 以及属于先前码本$\mathbf{C}_{:,<j}$的预测声学令牌为条件：

$$p(\mathbf{C}_{:,2:8}|\mathbf{x},\tilde{\mathbf{C}};\theta_{NAR}) = \prod_{j=2}^{8} p(\mathbf{c}_{:,j}|\mathbf{C}_{:,<j},\mathbf{x},\tilde{\mathbf{C}};\theta_{NAR}) \tag{2}$$

自回归模型与非自回归模型相结合，在语音质量和推理速度之间实现了良好的平衡。一方面，生成语音的语速应与注册录音保持一致；由于不同说话人的语速可能差异很大，因此很难针对不同说话人训练长度预测器。在这种情况下，自回归模型凭借其在声学序列长度预测方面的灵活性，是更自然的选择。另一方面，对于后续阶段，由于输出槽位的数量取决于第一阶段的序列长度，非自回归模型可以将时间复杂度从 $\mathcal{O}(T)$ 降低至 $\mathcal{O}(1)$。总体而言，$\mathbf{C}$ 的预测可以建模为：

$$p(\mathbf{C}|\mathbf{x},\tilde{\mathbf{C}};\theta) = p(\mathbf{c}_{:,1}|\tilde{\mathbf{C}}_{:,1},\mathbf{X};\theta_{AR}) \prod_{j=2}^{8} p(\mathbf{c}_{:,\mathbf{j}}|\mathbf{c}_{:,<j},\mathbf{x},\tilde{\mathbf{C}};\theta_{NAR}) \tag{3}$$

4.2.1 自回归Codec语言建模

自回归语言模型生成来自第一量化器的词元。它由音素嵌入 W_x、声学嵌入 W_a、Transformer解码器和预测层组成。为了生成具有特定内容的语音，我们将音素序列用作语言模型的音素提示。因此，模型输入是以下内容的拼接： $\mathbf{x}$ 和 $\mathbf{c}_{:,1}$，并在每个序列后追加两个特殊的 <EOS> 词元。我们分别为提示词元和输入词元计算正弦位置嵌入。对于因果Transformer模型，每个词元 $c_{t,1}$ 都可以关注 $(\mathbf{x},\mathbf{c}_{\leq t,1})$ ，如图 3左侧所示。模型经过优化，以最大化第一码本中下一个词元的概率。我们将输出投影层的参数与声学嵌入的参数共享 W_a。

在自回归模型中，我们不会在训练时显式提取一段音频作为提示。训练过程就是纯粹的因果语言模型训练。这样一来，任意前缀序列 $\mathbf{c}_{<t,\ 1}$ 都可视为序列后续部分 $\mathbf{c}_{\geq t,\ 1}$的提示。在推理时，给定一段注册录音，我们应将该录音的音素序列与用于合成的音素序列拼接在一起。同时，注册录音的声学令牌序列用作

自回归解码中的前缀，具体如公式 1所示。我们将在实验中研究这种设置的优势。

4.2.2 非自回归编解码器语言建模

通过自回归模型获得第一量化器码后，我们使用非自回归（NAR）模型生成其余七个量化器的码。非自回归模型的架构与自回归模型相似，但包含八个独立的声学嵌入层。在每个训练步骤中，我们随机采样一个训练阶段 $i \in [2, 8]$。模型经过训练，以最大化第 i 个量化器码本中的声学词元。将阶段 1 至阶段 $i-1$ 的声学词元嵌入后求和，作为模型输入：

$$e_{c_{t,j}} = W_a^j \odot c_{t,j} \tag{4}$$

$$\mathbf{e_{c_t}} = \sum_{j=1}^{i-1} e_{c_t,j} \tag{5}$$

其中，$\odot$ 表示索引选择。音素序列也被视为语言模型的提示。此外，为了克隆给定说话人的独特音色，我们还使用注册语音中的声学词元作为声学提示。具体而言，我们首先使用神经编解码器模型将注册语音转换为词元，得到 $\tilde{\mathbf{C}}^{T\times 8}$。将全部八个码本的嵌入表示相加，得到声学提示 $\mathbf{e_{\tilde{c}_t}} = \sum_{j=1}^{8} e_{\tilde{c}_t,j}$。为了预测第 i 个码本中的声学词元，Transformer 的输入为 $(\mathbf{e_x}, \mathbf{e_{\tilde{c}}}, \mathbf{e_{c_{:,<i}}})$ 的拼接。提示和声学序列的位置嵌入也分别计算。当前阶段 i 通过自适应层归一化（Adaptive Layer Normalization， [Xu et al., 2019]）运算注入网络，即 $\text{AdaLN}(h, i) = a_i \text{LayerNorm}(h) + b_i$，其中 h 为中间激活， a_i 和 b_i 由阶段嵌入经过线性投影得到。与自回归模型不同，非自回归模型允许每个词元在自注意力层中关注所有输入词元。我们还共享声学嵌入层和输出预测层的参数，这意味着第 j 个预测层的权重与第（$j+1$）个声学嵌入层的权重相同。

4.3 推理：通过提示进行上下文学习

上下文学习是基于文本的语言模型一项令人惊讶的能力，它能够在不额外更新参数的情况下，为未见过的输入预测标签。对于文本转语音，如果模型无需微调就能为未见说话人合成高质量语音，我们便认为该模型具备上下文学习能力。然而，现有文本转语音系统的上下文学习能力并不强，因为它们要么需要额外微调，要么在面对未见说话人时性能会大幅下降。

对于语言模型而言，在零样本场景中，提示是实现上下文学习的必要条件。我们按如下方式设计提示和推理流程。首先，将文本转换为音素序列，并将注册录音编码为声学矩阵，从而形成音素提示和声学提示。自回归模型和非自回归模型都会使用这两种提示。对于自回归模型，我们采用以提示为条件的采样解码，因为我们观察到，束搜索可能会使语言模型陷入无限循环。此外，基于采样的方法能够显著提高输出的多样性。对于非自回归模型，我们采用贪心解码来选择概率最高的词元。最后，我们使用神经编解码器解码器，以八个编码序列为条件生成波形。声学提示在语义上可能与待合成的语音相关，也可能无关，因此分为两种情况：

VALL-E：我们的主要目标是为未见说话人生成给定内容。模型输入一个文本句子、一段注册语音及其对应的转录文本。我们将注册语音的转录音素前置于给定句子的音素序列，作为音素提示，并将注册语音的第一层声学令牌 $\tilde{c}_{:,1}$ 作为声学前缀。借助音素提示和声学前缀，VALL-E 可以生成给定文本对应的声学词元，同时克隆该说话人的声音。

VALL-E-continual：在此设置中，我们分别使用完整的转录文本和话语开头 3 秒的内容作为音素和声学提示，并要求模型生成续接内容。推理过程与 VALL-E 设置相同，不同之处在于注册语音与生成的语音在语义上是连续的。

5 实验

5.1 实验设置

数据集：我们使用 LibriLight [Kahn 等人， 2020]作为训练数据，其中包含 6 万小时的英语有声读物未标注语音。LibriLight 中不同说话人的数量约为 7000。我们按照 Kaldi 配方 [Povey 等人， 2011], 使用 960 小时带标注的 LibriSpeech 数据训练混合 DNN-HMM 自动语音识别模型。混合模型训练完成后，对未标注语音数据进行解码，并转换为最佳音素级对齐路径，帧移为 30 毫秒。我们使用 EnCodec 模型 [Défossez 等人， 2022] 为 6 万小时的数据生成声学码矩阵。

模型：自回归模型和非自回归模型均采用相同的 Transformer 架构，包含 12 层、16 个注意力头、1024 的嵌入维度、4096 的前馈层维度，以及 0.1 的丢弃率。LibriLight 中波形的平均时长为 60 秒。训练期间，我们随机裁剪波形，使其时长在 10 秒至 20 秒之间。对应的音素对齐用作音素提示。我们会移除强制对齐音素序列中连续重复的部分。对于非自回归模型的声学提示词元，我们从同一段话语中随机选取一段 3 秒的波形。

模型使用 16 块 NVIDIA TESLA V100 32GB GPU 进行训练，每块 GPU 的批量大小为 6k 个声学词元，共训练 800k 步。我们使用 AdamW 优化器优化模型，将学习率在前 32k 次更新中逐渐预热至峰值 5×10^{-4}, 随后进行线性衰减。

基线：我们选择最先进的零样本文本转语音模型 YourTTS [Casanova 等人，2022年] 作为基线，该模型在由 VCTK [Veaux 等人、 2016], LibriTTS [Zen 等人、2019], 和 TTS-Portuguese [Casanova 等人，2022年]组成的混合数据集上训练。我们使用其发布的检查点*。

自动评测指标：我们采用最先进的说话人验证模型 WavLM-TDNN [Chen 等人，2022], 来评估提示（解压缩的注册语音）与合成语音之间的说话人相似度。WavLM-TDNN 在 2021年和 2022年排行榜的 VoxSRC 挑战赛中名列第一。在 Vox1-O、Vox1-E 和 Vox1-H 上，其等错误率（EER）平均值分别为 0.383、0.480 和 0.986。WavLM-TDNN 预测的相似度分数范围为 $[-1, 1]$，数值越大表示输入样本的相似度越高。

我们还评估了模型的合成鲁棒性。神经文本转语音系统存在鲁棒性问题，有时会因注意力对齐错误而出现删除、插入和替换错误。我们对生成的音频进行自动语音识别，并根据原始转写文本计算词错误率（WER）。在此实验中，我们采用在 LibriSpeech 960 小时数据集上微调的 HuBERT-Large[Hsu 等人， 2021]模型作为自动语音识别模型。该模型基于连接时序分类（CTC），且未采用语言模型融合。

人工评估：我们通过众包计算比较平均意见分（CMOS）和相似度平均意见分（SMOS），邀请 12 位和 6 位母语者分别参与 CMOS 和 SMOS 评估。SMOS 的评分范围为 1 至 5 分，间隔为 0.5 分。CMOS 的评分范围为 -3（新系统远差于基线）至 3（新系统远优于基线），间隔为 1 分。CMOS 是语音自然度的指标，而 SMOS 用于衡量语音是否与原说话人的声音相似。

5.2 LibriSpeech 评估

我们首先使用 LibriSpeech [Panayotov 等人的工作 2015]进行零样本文本转语音评估，因为 LibriLight 训练数据与 LibriSpeech test-clean 数据之间不存在说话人重叠。按照Borsos 等人 [2022], 的方法，我们使用 LibriSpeech test-clean 中时长在 4 到 10 秒之间的样本，得到一个 2.2 小时的子集。对于每个样本的合成，VALL-E 会随机选择同一说话人的另一段话语，并截取其中 3 秒的语音片段作为注册语音。每项实验运行三次，并报告平均分数。VALL-E-continual 使用真实语音的前 3 秒作为注册语音。

表2 展示了客观评估结果。我们首先计算真实标注语音的词错误率评分和说话人相似度分数，将其作为上界。为了比较说话人相似度，我们使用测试集中来自同一说话人的语音对。与 YourTTS 基线相比，我们的

*https://github.com/Edresson/YourTTS

表2：音频生成评估结果。YourTTS 和 VALL-E 是以音素为输入的文本转语音模型，而 GSLM 和 AudioLM 是以潜在代码为输入的语音转语音模型。AudioLM 的词错误率结果由 Conformer Transducer 模型获得 [Borsos 等人，2022]。由于 AudioLM* 未开源，我们无法使用自己的工具评估其说话人评分。

模型	WER	SPK
真实值	2.2	0.754
语音转语音系统		
GSLM	12.4	0.126
AudioLM*	6.0	-
文本转语音系统		
YourTTS	7.7	0.337
VALL-E	5.9	**0.580**
VALL-E-continual	**3.8**	0.508

模型在鲁棒性和说话人相似度两方面都明显更优，表明我们的生成语音高度忠实于给定文本和给定的注册语音。此外，在 VALL-E-continual 设置下，词错误率还可以进一步降低，因为前 3 秒的声学词元提取自真实标注。我们还将鲁棒性与其他基于语音到语音语言模型的生成模型 GSLM 和 AudioLM 进行了比较，这些模型以音频潜在编码作为输入。GSLM 使用 HuBERT 编码作为输入，并通过 Tacotron 模型重建波形2 [Shen 等人， 2018]以及 WaveGlow [Prenger 等人， 2019] 声码器。我们使用其发布的模型运行了开源代码，并对结果进行了评估。由于 HuBERT 编码会丢弃说话人身份，因此其说话人评分较低。对于 AudioLM，我们列出了其论文中报告的词错误率评分，该评分由 Conformer Transducer 模型得出。实验结果表明，在鲁棒性方面，VALL-E 优于其他基于语音到语音语言模型的生成系统。一个主要原因是，VALL-E 使用伪音素而非 HuBERT/w2v-BERT 编码进行训练，因此与输入文本具有更好的对齐质量。

我们从 LibriSpeech test-clean 中为人工评估随机抽取每位说话人的一个话语，共得到 40 个测试样例。表3 展示了人工评估结果。就 SMOS 而言，VALL-E 非常接近真实标注，这表明合成语音与测试中给定的未见说话人相似。它以 +0.93 的 SMOS 显著优于基线，证明了 VALL-E 在零样本场景中的有效性。在自然度方面，VALL-E 以 +0.12 的 CMOS 优于基线，这表明所提出的方法能够合成比基线更自然、更逼真的语音。

表3：在 LibriSpeech test-clean 上对 40 位说话人进行人工评估，每位说话人均使用 3秒注册录音。

	SMOS	CMOS（相较于 VALL-E）
YourTTS	$3.45_{\pm 0.09}$	-0.12
VALL-E	$4.38_{\pm 0.10}$	0.00
真实值	$4.5_{\pm 0.10}$	+0.17

消融实验： 本节将进行详细的消融实验。我们首先研究非自回归模型。我们训练了三个使用不同数量提示的非自回归模型。设置 **NAR-noprompt**在训练时不使用任何提示。设置 **NAR-phn prompt** 仅使用音素序列作为提示进行训练，而设置 **NAR-2 prompts** 则同时使用音素提示和声学标记提示作为条件。在评估时，我们将真实标注的一级声学标记作为模型输入，并计算词错误率和说话人相似度分数。结果见表 4。结果表明，即使声学输入词元是真实标注，不使用任何提示的模型在自动语音识别和说话人相似度评估中的表现都很差。加入音素提示后，词错误率从 19.6 大幅降至 3.0。这表明音素提示主要对生成内容起作用。在 **NAR-2 prompts**设置中，模型能够从声学标记提示中学习说话人信息，从而提升说话人评估质量。

表4：非自回归模型的消融实验。消融实验中，非自回归模型的输入为真实值。

	NAR-noprompt	NAR-phn prompt	NAR-2 prompts
WER	19.6	3.0	**2.8**
SPK	0.518	0.541	**0.732**

我们进一步对自回归模型进行了消融实验。在这些实验中，我们始终将 **NAR-2 prompts** 设置用作非自回归模型。在表 5中可以看到，移除声学提示（无声学提示）后，说话人相似度分数仅为0.236，这表明提示对于说话人身份至关重要。即使非自回归模型能够看到该提示，自回归模型的提示也对说话人相似度有很大贡献。

表5：自回归模型的消融实验。

	词错误率	说话人
VALL-E	**5.9**	**0.585**
无声学提示	**5.9**	0.236

5.3 VCTK评估

我们在包含108位说话人的VCTK上评估模型，这些说话人在训练期间均未出现。由于YourTTS在训练期间已接触过VCTK中的97位说话人，因此我们分别在全部107位说话人和11位未见说话人上评估YourTTS的性能。对于每位说话人，我们随机选取三段时长为3秒、5秒和10秒的话语作为提示，并将另一段话语的文本作为文本提示。

表6：在VCTK的108位说话人上进行的说话人相似度自动评估。*YourTTS在训练期间接触过97位说话人，而VALL-E没有接触过其中任何一位。

	3秒提示	5秒提示	10秒提示
108 位完整说话人			
YourTTS*	0.357	0.377	0.394
VALL-E	0.382	0.423	**0.484**
真实值	0.546	0.591	0.620
11 位未见说话人			
YourTTS	0.331	0.337	0.344
VALL-E	0.389	0.380	**0.414**
真实值	0.528	0.556	0.586

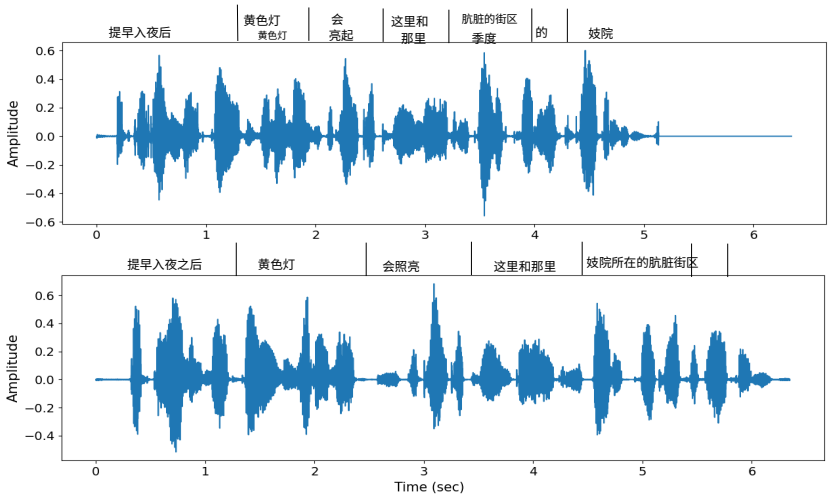
我们首先使用前文所述的说话人验证指标评估两个模型。从表6可以看出，即使基线在训练中见过 97 位说话人，VALL-E 的表现仍优于基线，这表明我们的模型能够合成说话人相似度更高的语音。在公平的设置下（11 位说话人）与基线进行比较时，性能差距会进一步扩大，尤其是在只能使用 3 秒提示的情况下。通过比较不同长度的提示可以看出，提示越长，我们的模型生成的语音就越相似，这与我们的直觉一致。

我们抽取了60位说话人进行人工评估，每位说话人各选取一段话语，其中11位是未见说话人，49位说话人在 YourTTS 的训练中出现过。VALL-E 未见过这60位说话人。在模型合成过程中，每位说话人都有一段3秒注册录音。表 7展示了我们的方法与基线及真实标注的比较。SMOS 的比较结果表明，VALL-E 的说话人相似度优于基线，即使基线在训练中见过其中一些说话人也是如此。并排比较式合成语音自然度平均意见分评估表明，VALL-E 比 YourTTS +0.23，说明其语音自然度表现显著更好。此外，VALL-E 的 CMOS 比真实标注高 +0.04，这表明在该数据集上，其与人类录音之间不存在统计学上的显著差异。与 LibriSpeech 上的评估结果相比，VALL-E 的 CMOS

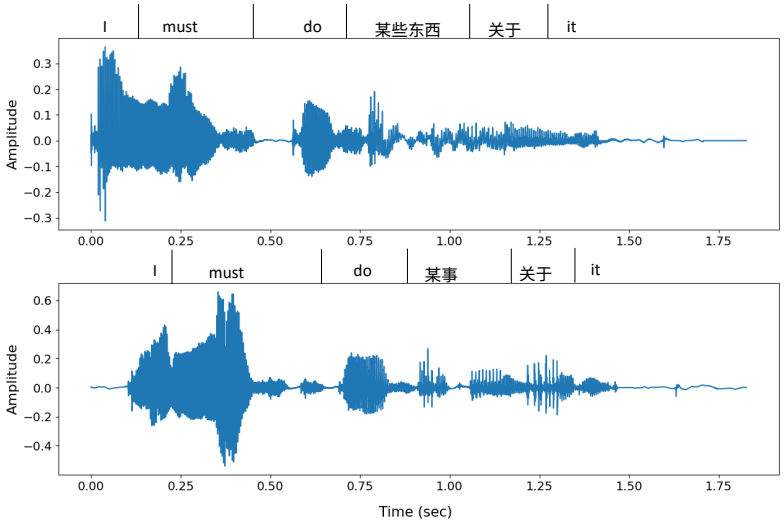
Tab表 7：使用 VCTK 上每位说话人的 3秒注册录音，对 60 位说话人进行人工评估 .

	SMOS	比较平均意见分（相较于 VALL-E）
YourTTS*	$3.70_{\pm 0.09}$	-0.23
VALL-E	$3.81_{\pm 0.09}$	0.00
真实值	$4.29_{\pm 0.09}$	-0.04

在与真实标注的比较中得分更高，这主要是因为平均句长较短，而且 VCTK 中的一些真实标注话语也存在环境噪声。在说话人相似度方面，VCTK 更具挑战性，因为其中包含口音各异的说话人，而训练数据和 LibriSpeech 测试数据中则没有带有各种口音的说话人。



(a) 一个 LibriSpeech 样本：提早入夜后，黄色的灯光便在妓院肮脏的街区各处亮起。



(b) 一个 VCTK 样本：我必须想办法处理这件事。

图4：VALL-E 的多样性分析。每条话语使用不同的随机种子合成两次。可以观察到，对于相同的输入，两次输出存在显著差异。

5.4 定性分析

多样性：以往的文本转语音系统中，输入文本与输出波形之间存在很强的一一映射关系，因为梅尔频谱的生成是基于逐步重建的，不包含随机性。由于 VALL-E 使用基于采样的方法生成离散令牌，因此推理过程中的随机性使其对同一输入文本产生的输出具有多样性。给定一个句子和一段注册录音，我们运行两次推理过程，并在图4中可视化其波形。在图4(a)中，我们观察到两次采样的长度和短语时长各不相同，其中第一次的语速更快。在图4(b)中，我们观察到两次采样的口音不同。第二次输出以更大的振幅突出“must”一词，而第一次输出则没有。我们的演示页面上提供了更多采样。

多样性对某些下游场景很重要。例如，语音识别总能从不同说话人和声学环境下的多样化输入中受益，而以往的文本转语音系统无法满足这一需求。考虑到 VALL-E 的多样性特征，它是生成语音识别伪数据的理想选择。

声学环境保持： 另一个有趣的发现是，声学提示与生成语音之间的声学环境一致性。当声学提示带有混响时，VALL-E 也能合成带有混响的语音，而基线则输出清晰语音。我们的解释是，VALL-E 在一个大规模数据集上训练，该数据集包含的声学条件比基线所用数据更多，因此 VALL-E 在训练期间能够学习声学一致性，而非仅学习清晰环境。我们在演示页面上展示了这种一致性。

说话人情绪保持： 情感文本转语音是语音合成的一个经典子领域，旨在合成带有指定情绪的语音。传统方法 [Lei 等人， 2021] 总是在有监督的情感文本转语音数据集上训练模型，其中语音与转录文本及情绪标签相对应。我们发现，VALL-E 在零样本设置下能够保留提示中的情绪。我们从 EmoV-DB 中选取声学提示 [Adigwe 等人， 2018], 这是一个包含五种情绪语音的数据集。即使模型未在情感文本转语音数据集上进行微调，VALL-E 也能在语音合成中保留提示中的相同情绪。我们在演示页面上提供了音频样本。

6 结论、局限性与未来工作

我们提出了 VALL-E，这是一种将音频编解码器代码作为中间表示的文本转语音语言模型方法。我们使用 6 万小时的语音数据对 VALL-E 进行预训练，并展示了其在零样本场景中的上下文学习能力。我们在 LibriSpeech 和 VCTK 上取得了新的零样本文本转语音最先进结果。此外，VALL-E 能够在合成时保留声学环境和说话人情绪，并在不同的基于采样的解码过程中生成多样化的输出。

尽管取得了显著进展，VALL-E 仍存在一些问题。

合成鲁棒性： 我们发现，在语音合成中，有些词可能发音不清、遗漏或重复。这主要是因为音素到声学语言部分采用了自回归模型，其中存在注意力对齐混乱的问题，且没有针对该问题的约束。这种现象在原始 Transformer 语音合成中也有观察到，过去通过采用非自回归模型或修改建模中的注意力机制来解决。未来，我们希望借助这些技术解决这一问题。

数据覆盖范围： 即使我们使用 6 万小时的数据进行训练，也仍然无法覆盖每个人的声音，尤其是带口音的说话者。与 LibriSpeech 相比，模型在 VCTK 上的效果较差，这也表明对带口音说话者的覆盖不足。此外，说话风格的多样性也不够，因为 LibriLight 是一个有声读物数据集，其中大多数话语都是朗读风格。未来，我们将进一步扩大训练数据规模，从韵律、说话风格和说话人相似度等方面提升模型性能。我们认为，通过扩大模型和数据规模，我们的方法几乎可以解决零样本语音合成任务。

模型结构： 目前，我们使用两个模型来预测不同量化器的编码。一个很有前景的方向是使用大型通用模型来预测这些编码。另一个值得探索的方向是在该框架中使用完整的非自回归模型，以加速模型推理。

更广泛的影响： 由于 VALL-E 能够合成保留说话人身份的语音，因此模型可能存在被滥用的潜在风险，例如伪造语音身份识别或冒充

特定说话人。为降低此类风险，可以构建检测模型，以判别音频片段是否由 VALL-E 合成。在进一步开发模型时，我们还将践行微软人工智能原则[*]。

参考文献

Adaeze Adigwe、Noé Tits、Kevin El Haddad、Sarah Ostadabbas 和 Thierry Dutoit。《情感语音数据库：实现对语音生成系统中情绪维度的控制》。*arXiv* 预印本 *arXiv:1806.09514*，2018年。

Junyi Ao、Rui Wang、Long Zhou、Chengyi Wang、Shuo Ren、Yu Wu、Shujie Liu、Tom Ko、Qing Li、Yu Zhang 等。《SpeechT5：面向口语语言处理的统一模态编码器-解码器预训练》。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第5723–5738页，2022年。

Serca n Ömer Arik、Jitong Chen、Kainan Peng、Wei Ping 和 Yanqi Zhou。神经语音克隆 wi th a f 新样本。在*NeurIPS*，第 10040–10050 页，2018年。

Alexei Baevski、Steffen Schneider 和 Michael Auli。vq-wav2vec：离散语音表示的自监督学习。载于*ICLM*，2020a。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 MichaelAuli。wav2vec 2.0：一种用于语音表示自监督学习的框架。*NeurIPS*，33:12449–12460，2020b。

He Bai、Renjie Zheng、Junkun Chen、Mingbo Ma、Xintong Li 和 Liang Huang。A t[3]：面向语音合成与编辑的对齐感知声学和文本预训练。载于国际机器学习大会，*ICML 2022*，*2022*年*7*月*17*日至*23*日，巴尔的摩，马里兰州，美国，第 162 卷机器学习研究论文集，第 1399–1411 页。PMLR，2022年。

Zalán Borsos、Raphaël Marinier、Damien Vincent、Eugene Kharitonov、Olivier Pietquin、Matthew Sharifi、Olivier Teboul、David Grangier、Marco Tagliasacchi 和 Neil Zeghidour。《AudioLM：一种用于音频生成的语言建模方法》。*CoRR*，abs/2209.03143，2022年。

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。《语言模型是小样本学习者》。载于*NeurIPS*，2020年。

Weicheng Cai、Jinkun Chen 和 Ming Li。探索端到端说话人和语言识别系统中的编码层与损失函数。载于 *Odyssey 2018*：说话人和语言识别研讨会，*2018*年*6*月*26*日至*29*日，法国 *Les Sables d'Olonne*，第 74–81 页。ISCA，2018年。

Edresson Casanova、Arnaldo Cândido Júnior、Christopher Shulby、Frederico Santos de Oliveira、João Paulo Ramos Teixeira、Moacir Antonelli Ponti 和 Sandra M. Aluísio。TTS-Portuguese语料库：用于巴西葡萄牙语语音合成的语料库。语言资源评估，56(3)：1043–1055，2022年。

Edresson Casanova、Julian Weber、Christopher D Shulby、Arnaldo Candido Junior、Eren Gölge 和 Moacir A Ponti。YourTTS：面向所有人的零样本多说话人文本转语音和零样本语音转换。载于*ICML*，第 2709–2720 页。PMLR，2022年。

Mingjian Chen、Xu Tan、Bohan Li、Yanqing Liu、Tao Qin、Sheng Zhao 和 Tie-Yan Liu。AdaSpeech：面向定制语音的自适应文本转语音。发表于*ICLR*，2021年。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等。Wavlm：面向全栈语音处理的大规模自监督预训练。*IEEE*信号处理精选主题期刊，16(6)：1505–1518，2022年。

[*] https://www.microsoft.com/ai/responsible-ai

Yutian Chen、Yannis M. Assael、Brendan Shillingford、David Budden、Scott E. Reed、Heiga Zen、Quan Wang、Luis C. Cobo、Andrew Trask、Ben Laurie、Çaglar Gülçehre、Aäron van den Oord、Oriol Vinyals 和 Nando de Freitas。样本高效自适应文本转语音。发表于*ICLR*，，2019年。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vinodkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael Isard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。PaLM：借助路径扩展语言建模规模。*CoRR*，abs/2204.02311，2022年。

Yu-An Chung、Yuxuan Wang、Wei-Ning Hsu、Yu Zhang 和 R. J. Skerry-Ryan。用于提升端到端语音合成数据效率的半监督训练。载于*ICASSP*，第 6940–6944 页。IEEE，2018年。

Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。高保真神经音频压缩。*arXiv* 预印本 *arXiv:2210.13438*，2022年。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。载于*NAACL*，第 4171–4186 页，2019年。

Chenpeng Du、Yiwei Guo、Xie Chen 和 Kai Yu。VQTTS：利用自监督矢量量化声学特征实现高保真文本转语音合成。载于*Interspeech 2022*，国际语音通信协会第*23*届年会，韩国仁川，*2022*年*9*月*18*日至*22*日，第 1596–1600 页。ISCA，2022年。doi：10.21437/Interspeech.2022-489。

Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。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》，29：3451–3460，2021年。

Sung-Feng Huang、Chyi-Jiunn Lin、Da-Rong Liu、Yi-Chen Chen 和 Hung-yi Lee。Meta-TTS：面向少样本说话人自适应文本转语音的元学习。*IEEE ACM* 音频、语音与语言处理汇刊，30：1558–1571，2022年。

Ye Jia、Yu Zhang、Ron J. Weiss、Quan Wang、Jonathan Shen、Fei Ren、Zhifeng Chen、Patrick Nguyen、Ruoming Pang、Ignacio Lopez-Moreno 和 Yonghui Wu。从说话人验证迁移学习到多说话人文本转语音合成。发表于*NeurIPS*，第4485–4495页，2018年。

Jacob Kahn、Morgane Rivière、Weiyi Zheng、Evgeny Kharitonov、Qiantong Xu、Pierre-Emmanuel Mazaré、Julien Karadayi、Vitaliy Liptchinsky、Ronan Collobert、Christian Fuegen 等。《Libri-light：一项面向有限监督或无监督自动语音识别的基准》。发表于*ICASSP*，第 7669–7673 页。IEEE，2020年。

Minki Kang、Dongchan Min 和 Sung Ju Hwang。《基于扩散模型的任意说话人自适应文本转语音合成》。《style id='1'>CoRR，abs/2211.09383，2022年。doi：10.48550/arXiv.2211.09383。

Heeseung Kim、Sungwon Kim 和 Sungroh Yoon。《Guided-TTS：一种通过分类器引导实现文本转语音的扩散模型》。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvári、Gang Niu 和 Sivan Sabato 编，国际机器学习大会，*ICML 2022*，*2022*年*7*月*17*日至*23*日，美国马里兰州巴尔的摩，第 162 卷，机器学习研究论文集，第 11119–11133 页。PMLR，2022年。

Jaehyeon Kim、Jungil Kong 和 Juhee Son。用于端到端文本转语音的对抗学习条件变分自编码器。载于 *ICML*，第 139 卷 机器学习研究论文集，第 5530–5540 页。PMLR，2021年。

Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae。HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。载于*NeurIPS*，2020年。Kushal Lakhotia、Evgeny Kharitonov、Wei-Ning Hsu、Yossi Adi、Adam Polyak、Benjamin Bolte、Tu Anh Nguyen、Jade Copet、Alexei Baevski、Adelrahman Mohamed 和 Emmanuel Dupoux。基于原始音频的生成式口语语言建模。*CoRR*，abs/2102.01192，2021年。Yi Lei、Shan Yang 和 Lei Xie。情感语音合成中的细粒度情感强度迁移、控制与预测。载于 *2021*年 *IEEE* 口语语言技术研讨会（*SLT*），第 423–430 页。IEEE，2021年。Naihan Li、Shujie Liu、Yanqing Liu、Sheng Zhao 和 Ming Liu。基于 Transformer 网络的神经语音合成。载于*AAAI*，第 6706–6713 页。AAAI，2019年。Yanqing Liu、Ruiqing Xue、Lei He、Xu Tan 和 Sheng Zhao。DelightfulTTS 2：采用对抗式向量量化自编码器的端到端语音合成。载于 *Interspeech 2022*，国际语音通信协会第*23*届年会，韩国仁川，*2022*年*9*月*18*日至*22*日，第 1581–1585 页。ISCA，2022年。doi：10.21437/Interspeech.2022-277。Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*，2019年。Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。LibriSpeech：一个基于公有领域有声读物的自动语音识别语料库。载于*ICASSP*，第 5206–5210 页。IEEE，2015年。Adam Polyak、Yossi Adi、Jade Copet、Eugene Kharitonov、Kushal Lakhotia、Wei-Ning Hsu、Ab- delrahman Mohamed 和 Emmanuel Dupoux。基于离散解耦自监督表征的语音重合成。载于*Interspeech*，第 3615–3619 页。ISCA，2021年。Vadim Popov、Ivan Vovk、Vladimir Gogoryan、Tasnima Sadekova 和 Mikhail A. Kudinov。Grad-TTS：一种用于文本转语音的扩散概率模型。载于 Marina Meila 和 Tong Zhang 编，第*38*届国际机器学习大会论文集（*ICML 2021*），*2021*年*7*月*18*日至*24*日，线上会议，第 139 卷，机器学习研究论文集，第 8599–8608 页。PMLR，2021年。网址：http://proceedings.mlr.press/v139/popov21a.html。Daniel Povey、Arnab Ghoshal、Gilles Boulianne、Lukas Burget、Ondrej Glembek、Nagendra Goel、Mirko Hannemann、Petr Motlicek、Yanmin Qian、Petr Schwarz 等。Kaldi 语音识别工具包。载于 *IEEE 2011* 自动语音识别与理解研讨会，编号 CONF。IEEE 信号处理学会，2011年。Ryan Prenger、Rafael Valle 和 Bryan Catanzaro。WaveGlow：一种用于语音合成的基于流的生成网络。载于*ICASSP*，第 3617–3621 页。IEEE，2019年。Yi Ren、Yangjun Ruan、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。FastSpeech：快速、稳健且可控的文本转语音。载于*NeurIPS*，第 3165–3174 页，2019年。Jonathan Shen、Ruoming Pang、Ron J. Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、RJ-Skerrv Ryan、Rif A. Saurous、Yannis Agiomyrgiannakis 和 Yonghui Wu。通过以梅尔频谱图预测为条件的 WaveNet 实现自然文本转语音合成。载于*ICASSP*，第 4779–4783 页。IEEE，2018年。Xu Tan、Tao Qin、Frank K. Soong 和 Tie-Yan Liu。神经语音合成综述。*CoRR*，abs/2106.15561，2021年。Andros Tjandra、Berrak Sisman、Mingyang Zhang、Sakriani Sakti、Haizhou Li 和 Satoshi Nakamura。面向 ZeroSpeech 挑战赛 2019 的 VQVAE 无监督单元发现与多尺度 code2spec 逆变器。载于 *Interspeech*，第 1118–1122 页。ISCA，2019年。Aäron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew W. Senior 和 Koray Kavukcuoglu。WaveNet：一种用于原始音频的生成模型。载于第*9*届 *ISCA* 语音合成研讨会，第 125 页。ISCA，2016年。

Aäron van den Oord、Oriol Vinyals 和 Koray Kavukcuoglu。神经离散表示学习。载于 神经信息处理系统进展 *30*：神经信息处理系统年度会议 *2017，2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩，第 6306–6315 页，2017年。Christophe Veaux、Junichi Yamagishi、Kirsten MacDonald 等。Superseded-cstr VCTK 语料库：用于 CSTR 声音克隆工具包的英语多说话人语料库。2016年。Tao Wang、Jianhua Tao、Ruibo Fu、Jiangyan Yi、Zhengqi Wen 和 Rongxiu Zhong。用于少样本说话人自适应的语音内容与声音因子分解。载于 *Interspeech*，第 796–800 页。ISCA，2020年。Yihan Wu、Xu Tan、Bohan Li、Lei He、Sheng Zhao、Ruihua Song、Tao Qin 和 Tie-Yan Liu。AdaSpeech 4：零样本场景下的自适应文本转语音。载于 *Interspeech 2022*，国际语音通信协会第*23*届年会，韩国仁川，*2022*年*9*月*18*日至*22*日，第 2568–2572 页。ISCA，2022年。doi：10.21437/Interspeech.2022-901。Jingjing Xu、Xu Sun、Zhiyuan Zhang、Guangxiang Zhao 和 Junyang Lin。理解并改进层归一化。载于 神经信息处理系统进展 *32*：神经信息处理系统年度会议 *2019，NeurIPS 2019，2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第 4383–4393 页，2019年。Neil Zeghidour、Alejandro Luebs、Ahmed Omran、Jan Skoglund 和 Marco Tagliasacchi。Sound-stream：端到端神经音频编解码器。*IEEE ACM* 音频、语音与语言处理汇刊，30：495–507，2022年。Heiga Zen、Viet Dang、Rob Clark、Yu Zhang、Ron J. Weiss、Ye Jia、Zhifeng Chen 和 Yonghui Wu。LibriTTS：源自 LibriSpeech 的文本转语音语料库。载于 *Interspeech*，第 1526–1530 页。ISCA，2019年。