arXiv:2302.04215v1 [eess.AS] 8 Feb 2023

面向真实世界自发语音文本转语音合成的向量量化方法

Li-Wei Chen、Shinji Watanabe、Alexander Rudnicky

语言技术研究所，卡内基梅隆大学 liweiche@cs.cmu.edu、
swatanab@cs.cmu.edu、air@cs.cmu.edu

摘要

近期，基于朗读语料或表演语料库训练的文本转语音（TTS）系统已达到接近人类水平的自然度。然而，人类语音的多样性往往超出这些语料库的覆盖范围。我们认为，处理这种多样性的能力对于人工智能系统实现人类水平的交流至关重要。我们的工作探索利用更丰富的真实世界数据来构建语音合成器。我们使用来自 YouTube 和播客的真实场景语音训练文本转语音系统。我们发现，基于梅尔频谱图的自回归模型在训练与推理对齐方面存在不匹配，导致合成语音难以理解；并证明，在多个码组中学习离散编码能够有效解决这一问题。我们介绍了 MQTTS 系统，其架构专为生成多个码组和实现单调对齐而设计，同时采用干净的静音提示来提升合成质量。我们进行了消融分析，以评估所提方法的有效性。结果表明，MQTTS 在多项客观和主观指标上均优于现有文本转语音系统。

1 引言

人工智能（AI），尤其是对话代理的一个关键组成部分，是合成人类水平语音的能力。随着深度学习的最新进展，基于神经网络的文本转语音（TTS）系统（Li 等人，2019；Kim 等人，2020；Ren 等人，2019）显著提升了合成语音的质量。然而，用于训练文本转语音系统的标准语料库（Ito 和 Johnson（2017）；Yamagishi、Veaux 和 MacDonald，2019）大多包含在受控环境中录制的朗读语音或表演式语音。另一方面，人类会自发地产生韵律丰富多样的语音，传递包括细微情绪在内的副语言信息。这种能力源于接触数千小时的真实世界语音。这表明，基于真实世界语音训练的文本转语音系统能够实现人类水平的人工智能。

在这项工作中，我们探索将从 YouTube 和播客收集的真实场景语音用于文本转语音。虽然最终目标是使用自动语音识别系统转录真实场景语音，但在此我们通过使用现成的转录语料库简化了这一设定，并聚焦于文本转语音。成功在真实场景语音上训练的系统能够利用野外数量无限的话语。因此，我们认为有可能复制类似于 GPT-3（Brown 等人，2020年）等大语言模型（LLM）的成功。这些系统只需少量资源，就能针对特定的说话人特征或录音环境进行微调。在这项工作中，我们探讨了在真实场景语音上训练文本转语音系统时面临的新挑战：与在受控环境中录制的朗读语音相比，韵律变化更大，背景噪声也更多。

针对真实场景语音，我们首先证明，基于梅尔频谱图的自回归系统在推理过程中无法生成恰当的文本-音频对齐，因而产生难以理解的语音。我们进一步表明，训练过程中仍然能够学习到清晰的对齐，因此，推理时对齐失败可以合理地归因于解码过程中的错误累积。我们发现，用学习得到的离散码本取代梅尔频谱图能够有效解决这一问题。我们认为，这是因为离散表示对输入噪声具有更强的鲁棒性。然而，我们的结果表明，即使增大码本规模，单一码本也会导致真实场景语音的重建失真。我们推测，单一码本不足以涵盖自发语音中丰富多样的韵律模式。因此，我们采用多个码本，并为多码本采样和单调对齐设计了专门的架构。最后，在推理过程中，我们使用干净的静音音频提示，促使模型即使在嘈杂语料上训练，也能生成干净的语音。我们将该系统命名为 MQTTS（多码本矢量量化文本转语音），并将在第 3 节介绍。

我们在第 5 节进行消融分析，并将基于梅尔频谱图的系统进行比较，以确定真实场景语音合成所需的特性。我们还进一步将 MQTTS 与非自回归方法进行比较。结果表明，我们的自回归 MQTTS 在可懂度和说话人迁移能力方面表现更好。MQTTS 的自然度略高，韵律多样性则高出许多。另一方面，非自回归模型在鲁棒性和计算速度方面表现出色。此外，使用干净的静音提示时，MQTTS 能够实现低得多的信噪比（SNR）。我们公开了代码[1]。

版权所有 © 2023，人工智能促进协会（www.aaai.org）。保留所有权利。

[1]https://github.com/b04901014/MQTTS

2 相关工作

自回归文本转语音。 典型的自回归文本转语音系统采用梅尔频谱图–声码器流水线（Shen 等人，2018；Li 等人，2019）。训练自回归模型以按顺序合成梅尔频谱图，然后单独训练神经声码器来重建波形。这类系统的一个问题是，它们对训练语料库中的差异容忍度较低。例如，Tacotron 2 的对齐学习（Shen 等人，2018）甚至会受到前导和尾随静音的影响。这些特性使得使用噪声极大的真实场景语音进行训练变得不可行。我们通过实验证明，用量化离散表示替代梅尔频谱图即可解决这一问题。另一个常见问题是推理过程中出现错误对齐，进而导致重复和删除错误。研究表明，强制采用单调对齐可以缓解这一问题（He、Deng 和 He，2019；Zeyer、Schl¨uter 和 Ney，2021）。在此，我们采用单调分块注意力（Chiu 和 Raffel，2018）中的类似概念；该方法已广泛用于语音识别（Liu 等人，2020）。不过，我们直接将注意力权重用作转移判据，而不对训练过程作任何修改。我们在第 3.3 节详细介绍我们的方法。无注意力 Tacotron（Shen 等人，2020）通过用外部时长预测器替换自回归模型的注意力机制来解决对齐问题。我们将证明，这种技术实际上会降低 MQTTS 的性能。

非自回归文本转语音。 由于鲁棒性更好、推理速度更快且训练难度更低，文本转语音领域的研究重点逐渐转向非自回归模型（Kim、Kong 和 Son，2021；Kim等人，2020；Ren等人，2019）。近期研究表明，在基准语料库上，这类模型在自然度方面已超越自回归模型。通常，训练时会使用时长预测器预测文本-音频对齐，并训练基于流的网络，根据对齐、音素序列和采样得到的高斯噪声合成语音。VITS（Kim、Kong 和 Son，2021）是近期提出的一种端到端模型，它采用随机时长预测器来提升多样性，在 VCTK（Yamagishi、Veaux 和 MacDonald，2019）数据集上实现了接近人类水平的自然度。尽管非自回归文本转语音系统种类繁多，我们仍选择 VITS 作为这一研究方向的代表，并分析其在真实世界数据上的表现，同时与自回归模型进行比较。

基于量化的生成模型。 一项新兴研究方向（Ramesh 等人，2021；Esser、Rombach 和 Ommer，2021）通过采用两阶段方法，在图像合成方面取得了成功。研究人员训练量化器，将图像编码为离散词元，然后使用自回归 Transformer 对图像词元的序列分布进行建模。DiscreTalk（Hayashi 和 Watanabe，2020）在文本转语音中采用了类似框架，使用 VQVAE 将语音离散化，并利用 Transformer 进行自回归生成。基于单元的语言模型（Lakhotia 等人，2021）（uLM）在由自监督学习（SSL）模型的表示派生出的离散词元上训练 Transformer，并单独训练单元到语音（u2s）声码器，将词元映射为语音。VQTTS（Du 等人，2022）进一步利用 u2s 声码器进行文本转语音，通过训练一个非自回归模型，将文本映射为离散词元。MQTTS 与这些工作的不同之处在于使用多码本，以及专为多码生成和单调对齐设计的架构。MQTTS 还学习专用于语音合成的码本，而非使用可能无法提供足够语音重建信息的自监督学习表示。

真实场景语音上的文本转语音。一种将真实场景语音用于文本转语音的方法是加入去噪。近期研究（Zhang 等，2021年）利用外部噪声数据联合训练去噪模块和文本转语音系统。我们的工作侧重于文本转语音系统本身，而非显式噪声建模，并通过简单的音频提示实现降噪，无需额外训练。另有研究（Yan 等，2021年）致力于建模自发语音的特征，例如填充停顿（即 嗯、呃）。我们没有为此设计专门的架构，而是让模型隐式地学习这些特征。

3 MQTTS

下面介绍 MQTTS 的模型架构，以及训练和推理流程。我们的模型分两个阶段进行训练。第一阶段如图 1 所示，训练一个量化器，将原始波形 $\mathbf{x}$ 映射为离散编码 $\mathbf{c} = \{c_t\}$，$t \in [1, \cdots, T]$，同时最小化量化损失 $\mathcal{L}_{VQ}$ 以及输入 $\mathbf{x}$ 与输出 $\mathbf{y}$之间的重建损失。判别器通过额外的对抗损失引导重建过程。我们将量化器中的编码器和解码器分别记为 Q_E、Q_D，将判别器记为 D，将可学习的码本嵌入记为 G_i，其中 $i \in [1, \cdots, N]$对应 N个码本。第二阶段固定量化器，并训练 Transformer，使其在说话人嵌入 s 和音素序列 $\mathbf{h}$的条件下，对 $\mathbf{c}$进行自回归生成，以支持多说话人文本转语音。这里，我们用粗体数学符号表示随时间变化的序列表示。

3.1 原始语音的量化

量化器架构。我们采用 HiFi-GAN（Kong、Kim 和 Bae，2020年）作为量化器解码器 Q_D 和判别器 D的骨干架构，因为已有研究表明，该架构能够根据梅尔频谱图生成高质量语音。我们用离散编码的学习嵌入替换输入梅尔频谱图。至于 Q_E，我们通过将反卷积层替换为卷积层，反转 HiFi-GAN 生成器的架构。我们发现，直接采用 HiFi-GAN 架构会导致量化损失 $\mathcal{L}_{VQ}$ 激增并造成训练发散。我们认为这是由于 HiFi-GAN 中残差块的输出未经过归一化。因此，我们在聚合前，对所有残差块的输出应用组归一化（Wu 和 He，2020年），每组包含16个通道。此外，我们将说话人嵌入 s 以广播相加的方式加到解码器 Q_D的输入上。

多码本学习。在第5.1节中，我们将通过实验证明，单一码本会导致真实场景语音的输入信号重建效果不佳。因此，我们借鉴 vq-wav2vec（Baevski、Schneider 和 Auli，2020年）中的概念，实现使用多个

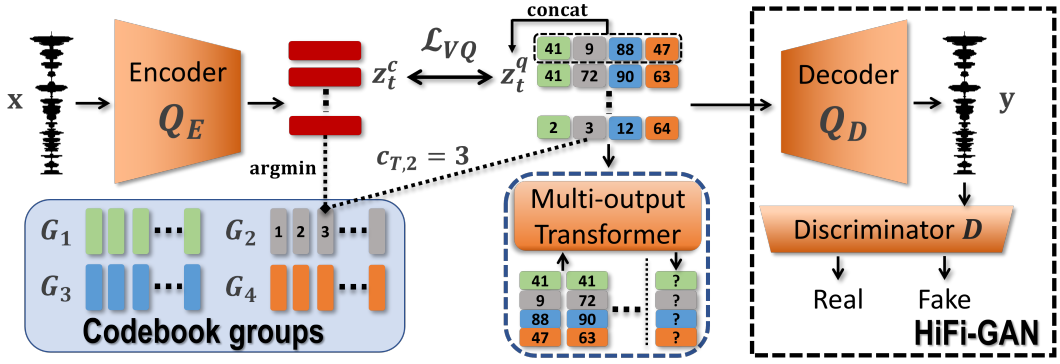


图 1：概览　MQTTS。我们用不同颜色表示来自不同码本的 4 个编码 G_i。

码本进行量化。如图1所示，现在每个 c_t 由 $N = 4$ 个编码表示，最终的嵌入是 N 个编码对应嵌入的串接。我们将与码本 G_i 对应的单个编码记为 $c_{t,\ i}$。为了从 z_t^c中获取离散编码 $c_{t,\ i}$，其中 z_t^c是 Q_E的输出，我们将 z_t^c的嵌入维度切分为 N 个等份，并将每一部分记为 $z_{t,\ i}^c$:

$$c_{t,i} = \arg\min_{z \in G_i} ||z_{t,i}^c - z||_2, \tag{1}$$

然后 z_t^q，通过直通梯度估计器获得解码器的输入:

$$z_{t,i}^q = z_{t,i}^c + \text{sg}[G_i(c_{t,i}) - z_{t,i}^c],\ z_t^q = \text{concat}[z_{t,i}^q] \tag{2}$$

其中，sg[·] 是停止梯度算子，$G_i(\cdot)$返回给定编码的嵌入，而 concat[·]算子会沿嵌入维度拼接所有编码。最终损失 $\mathcal{L}_F$ 由 HiFi-GAN 损失项 $\mathcal{L}_{GAN}$和额外的量化损失 $\mathcal{L}_{VQ}$组合而成:

$$\mathcal{L}_F = \lambda \mathcal{L}_{VQ}(Q_E, Q_D, G) + \mathcal{L}_{GAN}(Q_E, Q_D, G, D) \tag{3}$$

$$\begin{aligned}\mathcal{L}_{VQ}(Q_E, Q_D, G) =& \mathbb{E}_x \Bigg[\frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} (||\text{sg}[z_{t,i}^c] - z_{t,i}^q||_2^2 \\ &+ \gamma ||z_{t,i}^c - \text{sg}[z_{t,i}^q]||_2^2) \Bigg]\end{aligned} \tag{4}$$

其中，γ 和 λ 是承诺损失权重和量化损失权重的超参数。$\mathcal{L}_{GAN}$ 的详细定义请参见 HiFi-GAN 论文。请注意，我们的 $\mathcal{L}_{VQ}$ 不同于以往图像生成领域的工作；我们不包含重建损失，因为 $\mathcal{L}_{GAN}$ 中的梅尔频谱损失已实现这一目的。

3.2 使用 Transformer 进行条件合成

在第二阶段，我们训练一个 Transformer，根据过去的序列 $\mathbf{c}_{<t}$ 自回归地生成 c_t。如图 2 所示，Transformer 还以全局说话人嵌入 $s = E_{SPK}(\mathbf{x})$ 和音素序列 $\mathbf{h}$ 为条件。我们使用一个预训练的说话人嵌入提取器[2]，后接两个带 ReLU 激活的线性层，

[2]https://huggingface.co/pyannote/embedding

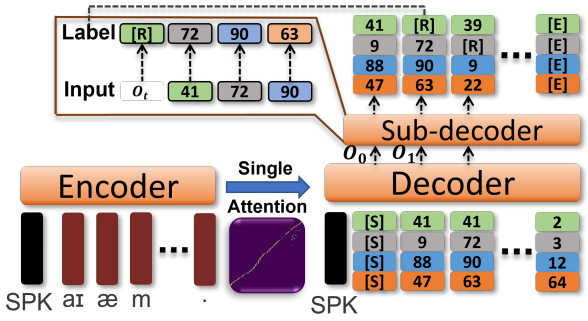


图 2：训练期间多输出 Transformer 的详细视图。 [R]表示重复标记。 [S]和 [E]分别表示起始标记和结束标记。SPK 表示处理后的说话人嵌入。

用于 E_{SPK}。训练 Transformer 以最大化对数似然:

$$\mathcal{L}_T = \mathbb{E}_{\mathbf{c}} \left[-\sum_t \log p(c_t \mid \mathbf{c}_{<t}, s, \mathbf{h}) \right] \tag{5}$$

架构设计。不同于常用的编码器-解码器架构（其所有解码器层都通过交叉注意力连接到编码器输出状态），我们采用了不同的配置。典型的编码器-解码器架构假定两个输入之间存在复杂的语义关联，需要从低层到高层表示对其进行建模；这对许多 NLP 任务而言确实如此。然而，在文本转语音中，交叉注意力只需建立一种独特的对齐映射，因此利用低层音频表示预先估计对齐模式所带来的益处较小。因此，我们仅在解码器的最后一层应用单个交叉注意力层，并使用单头交叉注意力来实现这种独特的对齐。更重要的是，唯一的对齐映射使我们能够采用单调对齐，我们将在第 3.3 节中介绍这种方法。我们使用 ALiBi（Press、Smith 和 Lewis，2022年）替代位置编码，使模型能够对编码器和解码器的长输入序列进行外推和合成。

多输出 Transformer。虽然采样技术（Holtzman 等，2020；Fan、Lewis 和 Dauphin

2018年）在自回归模型的性能中发挥着重要作用，但要将这些技术应用于我们的多编码设置并非易事。如果我们只使用一个线性层一次预测 N 个编码，就只能分别获取每个编码的分布。核采样（Holtzman 等，2020）在理论上不可行，因为它需要从 N 个编码的联合分布中选出 top-p 候选项。为解决这一问题，我们使用一个额外的 Transformer 解码器来显式建模编码的条件分布，并将其称为子解码器模块。与沿时间序列运行的主解码器不同，子解码器按顺序处理固定大小为 N 的码本组。如图 2 所示，在时间步 t，子解码器以主解码器的输出状态 O_t 作为初始条件，并在给定 $\mathbf{c}_t$ 的情况下，按顺序生成 $c_{t,\,i}$。我们为每个码本组使用独立的嵌入，不使用位置编码。这种配置使我们能够分别在时间步 t 对每个条件分布 $p(c_{t,\,i} \mid \mathbf{c}_{t,\,<i},\ O_t)$ 应用核采样。生成完整的 c_t 后，将 c_t 以自回归方式反馈给主解码器，以生成下一个时间步 O_{t+1}。

重复标记。 语音信号的表征通常具有很高的时间相似性，导致 Transformer 连续预测重复编码。我们通过引入一个额外的标记——重复标记，显式地对这类转移进行建模。连续重复的编码会被替换为该标记，并在推理时解码回原始标记。

3.3 推理

对于 Transformer 推理，我们采用核采样（Holtzman 等人，2020年），并将完全解码后的编码传递给 Q_D，以重建原始波形。

单调对齐。 我们将 $A(i, j)$ 表示为第 i 步的编码器状态与第 j步的解码器状态之间的交叉注意力值。由于我们的交叉注意力是单头单层的，因此 $A(i, j)$ 是唯一的。于是，在推理过程中，给定时间 t的新解码器状态，我们只计算其与 $[b_k, b_k + 1, \cdots, b_k + N_w]$ 处的 N_w 个编码器状态之间的交叉注意力，而不计算整个编码器状态序列； b_k 递归定义如下：

$$b_0 = 0, b_{k+1} = \begin{cases} b_k + 1, & \text{if } \frac{\exp\{A(b_k,k)\}}{\sum_{i=0}^{N_w}\exp\{A(b_k,k+i)\}} < \frac{1}{N_w} \\ b_k, & \text{otherwise} \end{cases} \tag{6}$$

直观而言，解码器在每个时刻只能依次关注一定范围的编码器状态上下文窗口 N_w，且只有当第一个编码器状态（位于 b_k）的 Softmax 注意力权重低于给定阈值（此处设为 ${N_w}^{-1}$）时，上下文窗口才会向前移动。一般来说，N_w 越小，单调对齐越严格，但语音上下文也越小。实际听测表明， N_w 取 3 到 6 时效果良好。此外，这种构造还使我们能够将对齐进度用作停止准则，而不必预测停止标记；后者容易因训练话语时长而过拟合。

音频提示。 在含噪语料库上训练时，似乎不可避免地会在推理过程中采样到不需要的背景噪声。在这里，我们使用一个简单的技巧来提示模型合成干净的语音。在实际合成之前，我们预先添加 3 帧由 Q_E 从一段静音片段编码而成的编码。干净的语音通常 preceded by 干净的静音，我们希望这能鼓励模型在没有背景噪声的情况下继续合成。

4 实验设置

4.1 数据集

我们在 GigaSpeech（Chen 等人，2021）上训练模型。该自动语音识别语料库包含有声读物、播客和 YouTube 的转录音频，采样率为 16 kHz。我们仅使用播客和 YouTube 语音，并采用附录 D 中详述的数据清洗流程，去除质量不佳的语音。最终数据集包含 896 小时的训练数据和 2.4 小时的验证数据，话语时长为 5 至 15 秒。为进行人工评估并确保说话人多样性，我们使用 VoxCeleb 测试集（Nagrani 等人，2019年）中的说话人参考音频，从不同说话人中选取 40 条话语；这些音频同样是带噪的自发语音。我们从 GigaSpeech 中随机选取相应文本，并从训练集和验证集中移除这些转录内容。

4.2 训练与模型配置

我们在公式 4 中使用 $\gamma = 0.25$，在公式 3 中使用 $\lambda = 10$，以平衡相对较高的 $\mathcal{L}_{GAN}$。训练量化器时，使用 Adam 优化器（Kingma 和 Ba，2015），设置为$\beta_1 = 0.5$、 $\beta_2 = 0.9$。对于 Transformer，我们使用 Adam 优化器，设置为 $\beta_1 = 0.9$、 $\beta_2 = 0.98$，每个 GPU 的批量大小为 200 个输入帧。学习率从 2×10^{-4} 线性衰减至 0。所有模型均使用 4 块 RTX A6000 GPU，以 bfloat16 精度进行训练。量化器训练了 600k 步，Transformer 训练了 800k 步。优化细节见发布的代码。推理时，我们使用核采样，设置为 $p = 0.8$。

基线模型。 对于 Transformer TTS（Li 等人，2019），编码器和解码器均采用 6 层，并采用与 MQTTS 相同的方法添加说话人嵌入。我们使用官方实现[3] 在 GigaSpeech 上训练 HiFi-GAN 声码器，并将其用作声码器。我们还使用 NVIDIA 的实现[4]，训练了 Tacotron 2（Shen 等人，2018），并按照 GST Tacotron（Wang 等人，2018）的方法，将说话人嵌入广播后加到编码器状态中。对于 VITS，我们基于官方实现进行修改，将其固定的说话人嵌入替换为自己的说话人嵌入模块，并训练了 800k 步。我们遵循原论文，将随机时长预测器的噪声缩放设为 0.8，将先验分布设为 0.667。

模型变体。 为测试 MQTTS 中各组件的有效性，我们评估了两个经过消融的模型版本：一个不使用单调对齐，另一个将子解码器模块替换为数量相同的线性层。对于不含子解码器的版本，我们对每个代码分别应用核采样

[3] https://github.com/jik876/hifi-gan [4] https://github.com/NVIDIA/DeepLearningExamples

。我们还评估了模型参数规模与性能之间的可扩展性。我们报告了三个模型版本：40M、100M 和 200M，它们使用相同的量化器，但 Transformer 的规模不同。由于对齐更清晰，40M 和 100M 版本的上下文窗口 N_w 设置为 4，200M 版本设置为 3。每个版本的详细架构见附录A。为了更好地与 MQTTS 进行比较，我们还训练了一个采用唯一对齐的单交叉注意力版本 Transformer TTS。最后，我们评估了无注意力 Tacotron（Shen 等人，2020）在提升对齐鲁棒性方面的方法。我们使用 VITS 的时长预测器生成的对齐，替换 Transformer TTS（单交叉注意力版本）和 MQTTS 的注意力对齐。

4.3 评估指标

ASR 错误率。自动语音识别（ASR）模型的错误率是评估合成语音可懂度的常用指标（Hayashi 等，2020）。首先，我们采用相对字符错误率（RCER）评估量化器模型。这里的"相对"是指使用 ASR 模型对输入语句的转录作为真实标注。我们认为，这更能反映重建对可懂度的影响。为此，我们采用 ESPNet 中基于 GigaSpeech 预训练的 ASR 模型[5]（Watanabe 等，2018）。我们还报告整个文本转语音系统的典型词错误率（WER），作为可懂度的客观指标。我们使用 Google 语音转文本 API 的视频模型作为 ASR 模型，并评估了 1472 个合成语音样本。

人工评估。我们报告两项采用 5 分制的平均意见分：用于量化器评估的 MOS-Q 和 MOS-N，以及用于比较文本转语音系统的 MOS-N。MOS-Q 要求人工评估者对语音的整体音频质量进行评分，而 MOS-N 则评估语音的自然度。我们使用了亚马逊土耳其机器人（MTurk）平台。人工实验的详细信息见附录B。我们同时展示平均意见分及其 95% 置信区间。

韵律 FID 分数（P-FID）。对于韵律多样性和自然度的评估，我们借鉴 Fr´echet Inception 距离（FID）（Heusel 等，2017），这是一种广泛用于评估图像生成模型的客观指标。FID 假设真实分布和合成分布均为高斯分布，并计算二阶 Wasserstein 距离，从而同时反映自然度和多样性。我们使用预训练分类器提取每个样本的稠密表示。在文本转语音中，由于数据集规模不足，FID 很少被使用。然而，GigaSpeech 中的 500k 条话语使 FID 所需的采样复杂度成为可能。我们使用 Wagner 等，2022 发布的情感分类器，该分类器在 MSP-Podcast（Lotfian 和 Busso，2019）上进行了预训练。我们使用最终决策层之前的输入，并将其缩放 10 倍，以计算 FID 分数。我们从训练集中随机采样 50k 条话语，并使用相同的文本集合生成 50k 条合成语音，同时打乱说话人参考音频。

说话人相似度分数（SSS）。对于多说话人文本转语音，评估合成器传递说话人特征的效果同样十分重要。本文采用余弦相似度来比较参考语音与合成语音的说话人嵌入。我们对与 P-FID 相同的 50k 个样本进行评估。

[5]https://zenodo.org/record/4630406#.YoT0Ji-B1QI

表1：VoxCeleb 测试集上量化器和声码器重构质量的比较。HF-GAN 即 HiFi-GAN。

方法	码本大小 (#组)	RCER (%)	MOS-Q (95% 置信区间)	MOS-N (95% 置信区间)
GT	*n/a*	*n/a*	3.66(.06)	3.81(.05)
HF-GAN	*n/a*	**12.8**	3.47(.06)	3.62(.06)
MQTTS 量化	1024 (1)	56.5	3.38(.07)	3.49(.06)
	65536 (1)	59.9	3.40(.06)	3.48(.06)
	160 (4)	19.7	**3.63**（。06）	3.67(.06)
	160 (8)	14.2	3.56(.06)	**3.71**(.06)

梅尔倒谱失真（梅尔倒谱失真）。我们遵循先前的研究（Hayashi 等人（2021）），使用梅尔倒谱失真来衡量在文本和说话人参考语音相同的情况下，真实标注与合成结果之间的重建误差。我们提取 23 个梅尔倒谱系数，并使用欧氏距离作为度量。采用动态时间规整（Giorgino（2009））进行对齐。我们使用验证集中的真实语音片段计算梅尔倒谱失真。

5 结果

5.1 量化器分析

表 1 展示了量化器的重建质量。从 RCER 可以清楚地看出，使用单个码本（$N = 1$）会严重扭曲输入的音素序列。其明显偏低的 MOS-Q 和 MOS-N 分数也表明，$N = 1$不足以生成高质量且自然的语音。即使将码本大小大幅增加至 65536，与 1024 个编码相比，各项指标几乎没有差异。这一结果表明，使用多个码组是必要的：单个码本甚至无法重建原始输入，更不用说将其应用于整个流程。有趣的是，我们发现对于 LJSpeech（Ito 和 Johnson（2017）），单个码本就已足够，这表明多说话人真实语音的复杂性要求使用多个码本。与基于梅尔频谱图的 HiFi-GAN（表 1 中的 HF-GAN）相比，我们的量化器在 RCER 方面略高，但语音质量（MOS-Q）明显更好。在其余实验中，我们使用码本大小为 160 的 $N = 4$，因为我们观察到，将码组数量从 $N = 4$ 增加到 8 时，可感知的语音质量和可懂度差异很小。

5.2 性能

在给定的真实世界语音语料库上训练的文本转语音系统的性能如表 2 所示。

自回归模型。 在对系统进行定量比较之前，我们先展示 Tacotron2 和 Transformer TTS 的一些定性特征。图 3 比较了自回归模型的对齐情况。可以看出，Tacotron 2 在推理时的对齐完全失败。另一方面，

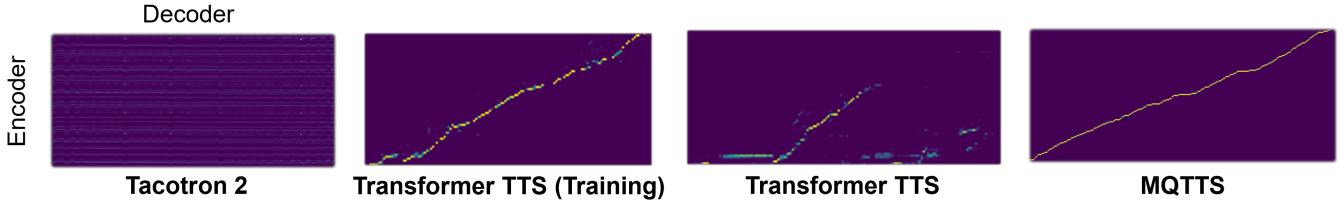


图 3：不同模型推理编码器-解码器对齐的比较。对于 Transformer TTS，我们选取了一个能够学习对齐的交叉注意力。所有注意力头和层的交叉注意力请参见附录 C。

表 2：比　　较文本转语音模型。平均意见分附有 95% 置信区间。梅尔倒谱失真附有一个标准差　　n.

方法	#参数	词错误率 (%)↓	SSS ↑	P-FID ↓	梅尔倒谱失真 ↓	MOS-N ↑
GT	*n/a*	14.9	*n/a*	0.11	*n/a*	4.01±0.06
VITS	≈ 40M	28.4	0.718	15.48	8.94±1.16	3.84±0.06
	≈ 100M	24.8	0.734	12.04	**8.73**±1.19	3.84±0.06
Transformer TTS	≈ 100M	82.6	0.434	442.29	12.49±2.77	2.34±0.09
Transformer TTS（使用单注意力）		90.2	0.331	491.49	12.79±4.35	2.23±0.10
Transformer TTS（带时长预测器）		74.3	0.451	196.09	11.70±2.12	2.75±0.09
MQTTS	≈ 40M	24.9	0.727	11.67	10.67±1.78	3.79±0.06
	≈ 100M	22.3	0.751	8.58	10.22±1.68	3.87±0.06
	≈ 200M	**22.2**	**0.762**	**6.21**	10.17±1.70	**3.89**±0.06
MQTTS（无单调对齐）	≈ 100M	34.3	0.740	10.20	10.86±2.00	3.76±0.06
MQTTS（无子解码器）		27.0	0.740	12.12	10.47±1.81	3.75±0.07
MQTTS（带时长预测器）		53.2	0.725	16.65	10.88±1.62	3.67±0.06

Transformer TTS 在训练期间能够生成清晰的对齐。然而，在推理时只形成了对齐序列的一个子序列。从感知效果来看，Transformer TTS 只合成了开头的几个词，随后便出现无法辨认的噪声。至于 Tacotron 2，则只生成了高频噪声。由于其合成结果无法理解，我们决定不将 Tacotron 2 纳入定量评估。从表 2 可以看出，Transformer TTS 在所有指标上都逊于其他系统。推理时对齐不佳是性能低下的主要原因之一，其中包括显著偏高的 P-FID，因为合成结果包含很长的非语音片段。应用时长预测器解决了对齐问题，并显著提升了性能，但与其他系统相比仍存在较大差距。我们还注意到，在 Transformer TTS 中使用单个注意力会导致性能变差，这表明尽管各对齐模式高度相似，使用多个交叉注意力仍可能提升鲁棒性。此外，参数量较大（增加了交叉注意力层）也可能是另一个影响因素。尽管如此，这也意味着无法使用外部时长预测器或单调对齐，因为部分注意力头经过训练，会对更大的上下文取平均。

模型变体。 我们可以从表 2 中看出单调对齐和子解码器的有效性。两个消融版本的 MOS-N 分数都较低。移除单调对齐后，模型的词错误率明显升高。另一方面，替换子解码器模块对可懂度的影响较小，但会导致 P-FID 分数降低。这对可懂度的影响较小，但会导致 P-FID 分数降低。这是合理的，因为单调对齐主要用于缓解重复和删除错误，对韵律的影响较小；而使用子解码器则能改善采样，从而提升韵律的自然度和多样性。然而，这些经过消融的 MQTTS 版本仍远胜于 Transformer TTS，这表明相较于连续梅尔频谱图，离散编码在真实场景语音中具有优势。用外部时长预测器替代对齐会使性能大幅下降。检查合成样本后发现，发声往往会出现不自然的过渡。我们认为，这是因为给定的对齐与 MQTTS 规划的对齐差异很大，导致发声中途出现突兀的过渡。

非自回归模型。 首先，我们注意到 VITS 和 MQTTS 的 MOS-N 分数都相当不错。然而，MQTTS 的 P-FID 明显低于 VITS，这表明 MQTTS 生成的样本更多样。我们推测，与自回归建模相比，并行生成语音信号是一项困难得多的任务。因此，由于模型容量不足，非自回归模型往往会聚焦于语料库中较少的一组常见说话风格。如果将 VITS 的规模从 4000 万参数扩大到 1 亿参数，我们会发现，在 MOS-N 不变的情况下，P-FID 有所下降。这表明，增大模型容量可以提升对多样性的建模能力，但不会改善自然度。我们没有纳入 2 亿参数版本的 VITS，因为未能找到能使训练收敛的配置。5 个样本——

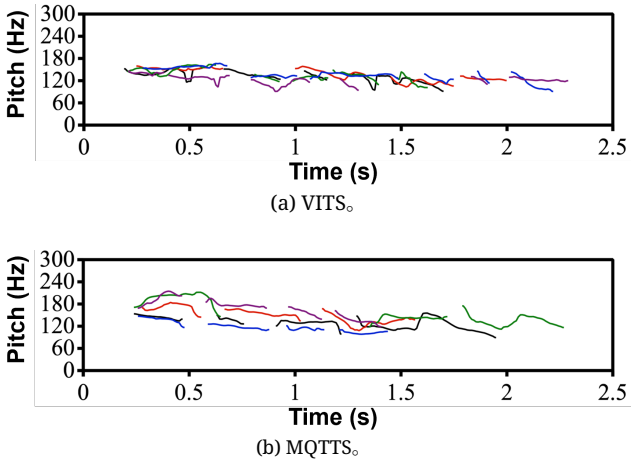
图4中两个系统针对同一文本生成的样本进一步支持了我们关于多样性对比的观点。这或许也能解释 VITS 较低的梅尔倒谱失真：其合成语音的韵律模式较为保守，因此在梅尔倒谱失真方面更具容错性。一个可能的例子是内部停顿的时长。VITS 合成的语音大多包含短暂、间歇性的停顿，而 MQTTS 则常生成较长的停顿。虽然自然语音中也并不少见，但这可能导致与真实标注出现较大的对齐偏差，从而降低梅尔倒谱失真。

MQTTS 在可懂度和说话人迁移能力方面均表现更佳。我们发现，与 VITS 相比，MQTTS 更能捕捉话语级属性（即情感、语速）。在自然度方面，随着模型容量增大，MQTTS 的 MOS-N 持续提升。这表明两者呈现出不同的缩放特性：更高的模型容量能提升 MQTTS 的自然度，却能提升 VITS 的多样性。比较 VITS 和 MQTTS 参数量相同（100 M）的模型，除梅尔倒谱失真（MCD）外，MQTTS 在所有指标上均胜出，原因已在前文解释。这表明只要资源充足，MQTTS 总体上优于 VITS。此外，我们观察到 MQTTS 的 100M 和 200M 版本都出现了过拟合，且 200M 版本的过拟合程度更严重。这解释了从 100M 到 200M 的提升幅度较小，也表明若要进一步提升性能，需要更大的训练语料库。

误差分析。 尽管 MQTTS 在表 2 中的平均表现更好，但我们发现，与非自回归系统相比，它的采样鲁棒性较低。这是合理的，因为更高的多样性不可避免地伴随着更高的采样出较差合成结果的风险。我们观察到，在使用语速较慢风格的说话人参考语音时，部分样本中的元音会出现不自然的延长，而 VITS 中很少出现这种情况。此外，以较差录音环境开头的样本通常会导致糟糕的合成结果。删除错误的出现频率在 MQTTS 中也更高；与替换错误相比，我们认为删除错误更加不可取。删除错误占 MQTTS-100M 中 22.3% 词错误率的 8.4 个百分点，而在 VITS-100M 中仅占 24.8% 的 6.8 个百分点。我们推测，由于间歇性停顿没有被明确标注，这会促使 MQTTS 即使在关注某个音素时也生成静音。然而，如果音素能够被成功生成，其语音通常比 VITS 合成结果中的音素听起来更加自然、清晰。随着模型容量增大（从 40M 增至 200M），我们观察到这些错误逐渐得到纠正，这表明更多数据和更大的模型最终能够解决这些问题。

5.3 音频提示与信噪比

为了更好地理解音频提示对合成效果的影响，我们将音频提示设为从不同标准差的 σ 中采样得到的白噪声（随后通过Q_E进行编码）。图5展示了 σ 与信噪比（SNR）之间的关系。我们使用 WADA-SNR 算法（Kim和Stern 2008年）估计信噪比。从图5可以清楚看出，随着 σ 增加，信噪比会下降，这证实了模型受到了该提示的引导。与不使用任何音频提示相比，使用小于 10^{-4} 的 σ 可以有效提高信噪比。我们其他所有实验均使用 $\sigma = 10^{-5}$完成。我们还注意到，VITS 的信噪比更低



(a) VITS。

(b) MQTTS。

图4：同一说话人的两个模型生成的话语“How much variation is there?”的基频轮廓。

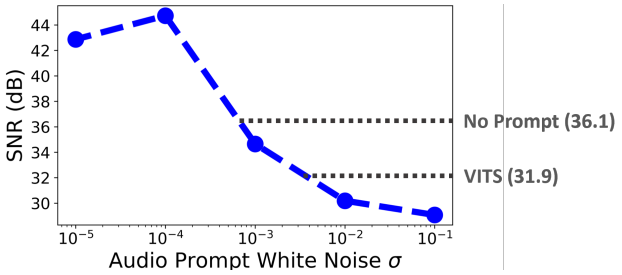


图5：音频提示中不同噪声等级下信噪比的比较。信噪比基于1472次合成计算。

。从感知上来看，我们可以在所有合成结果中听到一种轻微的模糊噪声。我们推测，VITS 还试图对环境噪声进行建模并从中采样，而这极其困难。建模失败导致它只能合成单一类型的噪声。

6 结论

在真实场景语音上，我们的实证结果表明，对于自回归合成器而言，多种离散编码优于梅尔频谱图。通过适当的建模，MQTTS 的性能也优于非自回归合成器。尽管如此，我们性能最佳的合成语音与人类语音之间仍然存在较大差距。我们认为，弥合这一差距对于发展人工智能领域的人类级交流至关重要。获取更多数据是一种直接的解决方案。在这方面，我们有兴趣结合并利用自动语音识别模型，对真实场景语音语料库进行转录。另一方面，也可以设计更好的建模方法，以缓解我们在误差分析中提到的问题。例如，可以在解码过程中使用静音检测，以防止在发声之前发生音素过渡，从而减轻删除错误。此外，我们计划进一步比较并结合自监督离散语音表征和韵律表征，以及我们学习得到的码本。

致谢

我们感谢 Amazon Alexa 对本研究的支持。感谢 Sid Dalmia 和 Soumi Maiti 参与讨论并提供反馈。

参考文献

Baevski, A.; Schneider, S.; and Auli, M. 2020年。vq-wav2vec：离散语音表征的自监督学习。载于第八届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020年。语言模型是少样本学习者。载于 Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H. 编，《神经信息处理系统进展》，第33卷，1877–1901页。Curran Associates, Inc.

Chen, G.；Chai, S.；Wang, G.；Du, J.；Zhang, W.-Q.；Weng, C.；Su, D.；Povey, D.；Trmal, J.；Zhang, J.；Jin, M.；Khu- danpur, S.；Watanabe, S.；Zhao, S.；Zou, W.；Li, X.；Yao, X.；Wang, Y.；Wang, Y.；You, Z.；以及 Yan, Z. 2021年。《GigaSpeech：一个包含10,000 小时转录音频、不断发展的多领域自动语音识别语料库》。载于*Interspeech 2021*会议论文集。

Chiu, C.；以及 Raffel, C. 2018年。《单调分块注意力机制》。载于 第*6*届国际学习表征会议，*ICLR 2018*，加拿大不列颠哥伦比亚省温哥华，*2018*年*4*月*30*日至*5*月*3*日，会议专题论文集。OpenReview.net。

Du, C.; Guo, Y.; Chen, X.; and Yu, K. 2022年. VQTTS：基于自监督矢量量化声学特征的高保真文本转语音合成. In*Interspeech 2022*论文集, 1596–1600.

Esser, P.; Rombach, R.; 和 Ommer, B. 2021年。驯服用于高分辨率图像合成的 Transformer。载于*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR)*，12873–12883。

Fan, A.; Lewis, M.; 和 Dauphin, Y. 2018年。分层神经故事生成。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），889–898。澳大利亚墨尔本：计算语言学协会。Giorgino, T. 2009。使用 R 计算和可视化动态时间规整对齐：dtw软件包。《统计软件杂志》，31(7)：1–24。

Hayashi, T.; 和 Watanabe, S. 2020年。DiscreTalk：将文本转语音视为机器翻译问题。*CoRR*，abs/2005.05525。

Hayashi, T.; Yamamoto, R.; Inoue, K.; Yoshimura, T.; Watanabe, S.; Toda, T.; Takeda, K.; Zhang, Y.; 和 Tan, X. 2020年。Espnet-TTS：统一、可复现且可集成的开源端到端文本转语音工具包。载于*2020*年*IEEE*国际声学、语音与信号处理会议，*ICASSP 2020*，西班牙巴塞罗那，*2020*年*5*月*4*日至*8*日，7654–7658。IEEE。

Hayashi, T.; Yamamoto, R.; Yoshimura, T.; Wu, P.; Shi, J.; Saeki, T.; Ju, Y.; Yasuda, Y.; Takamichi, S.; 和 Watanabe, S. 2021年。ESPnet2-TTS：拓展文本转语音研究的边界。*CoRR*, abs/2110.07840。

He, M.; Deng, Y.; and He, L. 2019年. 用于神经文本转语音的逐步单调注意力鲁棒序列到序列声学建模. In Kubin, G.; and Kacic, Z., eds.,*Interspeech 2019*，国际语音通信协会第*20*届年度会议，奥地利格拉茨，*2019*年*9*月*15*日至*19*日,1293–1297. ISCA.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017年. 通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds.,神经信息处理系统进展, 第30卷.Curran Associates, Inc.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; 和 Choi, Y. 2020年。《神经文本退化的奇特案例》。载于国际学习表征会议。

Ito，K.；和 Johnson, L. 2017年。《LJ 语音数据集》集。

Kim, C.；和 Stern, R. M. 2008年。基于波形幅度分布分析的鲁棒信噪比估计。载于*Interspeech 2008* 论文集，2598–2601。

Kim, J.；Kim, S.；Kong, J.；和 Yoon, S. 2020年。Glow-TTS：通过单调对齐搜索实现文本转语音的生成流。载于 Larochelle, H.；Ranzato, M.；Hadsell, R.；Balcan, M.；和 Lin, H. 编，《神经信息处理系统进展》，第33卷，8067–8077页。Curran Associates, Inc.

Kim, J.; Kong, J.; and Son, J. 2021年. 用于端到端文本转语音的对抗学习条件变分自编码器. In Meila,M.; and Zhang,T., eds., 第*38*届国际机器学习会议论文集, 机器学习研究论文集第139卷, 5530–5540. PMLR.

Kingma, D. P.; and Ba, J. 2015年. Adam：一种随机优化方法. In 第*3*届国际学习表征会议, *ICLR 2015,* 圣迭戈, 加利福尼亚州, 美国, *2015*年*5*月*7*日至*9*日, *ConferenceTrack Proceedings*.

Kong, J.; Kim, J.; and Bae, J. 2020年. HiFi-GAN：用于高效、高保真语音合成的生成对抗网络. In 第*34*届神经信息处理系统国际会议论文集,NIPS'20.雷德胡克,纽约州, 美国：Curran Associates Inc. ISBN 9781713829546.

Lakhotia, K.；Kharitonov, E.；Hsu, W.-N.；Adi, Y.；Polyak, A.；Bolte, B.；Nguyen, T.-A.；Copet, J.；Baevski, A.；Mo- hamed, A.；和 Dupoux, E. 2021年。《基于原始音频的生成式口语语言建模》。计算语言学协会汇刊，9：1336–1354。

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019年. 基于Transformer网络的神经语音合成. *AAAI*人工智能会议论文集, 33(01): 6706–6713.

Liu, B.; Cao, S.; Sun, S.; Zhang, W.; and Ma, L. 2020年. 用于在线语音识别的多头单调分块注意力. *CoRR*, abs/2005.00205. Lotfian, R.; and Busso, C. 2019年. 通过从现有播客录音中检索情感语音，构建自然且情感均衡的语音语料库. *IEEE*情感计算汇刊, 10(4): 471–483. Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2019年. Voxceleb：大规模野外说话人验证. 计算机科学与语言. Press, O.; Smith, N.; and Lewis, M. 2022年. 短训练、长测试：带线性偏置的注意力支持输入长度外推. In 国际学习表征会议. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Rad- ford, A.; Chen, M.; and Sutskever, I. 2021年. 零样本文本到图像生成. In Meila, M.; and Zhang, T., eds., 第*38*届国际机器学习会议论文集, 机器学习研究论文集第139卷, 8821–8831. PMLR. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019年. FastSpeech：快速、稳健且可控的文本转语音. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alch´e-Buc, F.; Fox, E.; and Garnett, R., eds., 神经信息处理系统进展, 第32卷. Curran Associates, Inc. Shen, J.; Jia, Y.; Chrzanowski, M.; Zhang, Y.; Elias, I.; Zen, H.; and Wu, Y. 2020年. 无注意力 Tacotron：包含无监督时长建模的稳健、可控神经文本转语音合成. *CoRR*, abs/2010.04301. Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; Saurous, R. A.; Agiomvrgiannakis, Y.; and Wu, Y. 2018年. 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. In *2018*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*）, 4779–4783. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; and Schuller, B. W. 2022年. Transformer时代的语音情感识别：弥合效价差距. Wang, Y.; Stanton, D.; Zhang, Y.; Skerry-Ryan, R. J.; Bat- tenberg, E.; Shor, J.; Xiao, Y.; Ren, F.; Jia, Y.; and Saurous, R. A. 2018年. 风格标记：端到端语音合成中的无监督风格建模、控制与迁移. In*ICML*. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Enrique Yalta Soplin, N.; Heymann, J.; Wiesner, M.; Chen, N.; Renduchintala, A.; and Ochiai, T. 2018年. ESP-net：端到端语音处理工具包 . In*Interspeech* 论文集 , 2207–2211. Wu, Y.; and He, K. 2020 年 . 组归一化 . 国际计算机视觉杂志, 128(3): 742–755. Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019年. CSTR VCTK语料库：面向CSTR语音克隆工具包的英语多说话人语料库（版本0.92）。

Yan, Y.; Tan, X.; Li, B.; Zhang, G.; Qin, T.; Zhao, S.; Shen, Y.; Zhang, W.; 和 Liu, T. 2021年。AdaSpeech 3：面向自发风格的自适应文本转语音。*CoRR*，abs/2107.02530。Zeyer, A.; Schl¨uter, R.; 和 Ney, H. 2021年。潜在单调注意力变体研究。*CoRR*，abs/2103.16710。Zhang, C.; Ren, Y.; Tan, X.; Liu, J.; Zhang, K.; Qin, T.; Zhao, S.; 和 Liu, T. 2021年。Denoispeech：利用帧级噪声建模进行去噪文本转语音。载于*IEEE* 国际声学、语音与信号处理会议（*IC ASSP 2021*），多伦多，安大略省，加拿大，*2021*年*6*月*6*日至*11*日，7063–7067。IEEE。