FoundationTTS：基于生成式语言模型的自动语音识别定制文本转语音

Ruiqing Xue、*Yanqing Liu*、*Lei He*、*Xu Tan*、*Linquan Liu*、*Edward Lin*、*Sheng Zha* o

微软

arXiv:2303.02939v3 [eess.AS] 8 Mar 2023

摘要

神经文本转语音（文本转语音）通常采用级联架构，分别优化声学模型和声码器；或采用端到端架构，以连续梅尔频谱或自行提取的语音帧作为连接声学模型和声码器的中间表示。这两种架构都存在两项局限：1）仅凭音素难以预测连续声学帧，而且还需要时长或基频等声学信息来解决一对多问题；但这一做法难以扩展到大规模数据集和噪声数据集。2）为了基于连续语音特征生成多样化的语音输出，通常需要复杂的变分自编码器或基于流的模型。本文提出 FoundationTTS，这是一种新型语音合成系统，采用神经音频编解码器提取离散语音令牌并重建波形，同时使用大语言模型根据语言学（音素）词元生成离散词元。具体而言，1）我们提出一种基于矢量量化自编码器并采用对抗训练的分层编解码器网络（VQ-GAN）：首先通过细粒度编解码器提取连续的帧级语音表示，再通过粗粒度编解码器从每个连续语音帧中提取一个离散令牌；2）我们将语音令牌、语言学令牌和说话人令牌与大语言模型联合优化，并自回归地预测离散语音令牌。实验表明，与基线系统相比，FoundationTTS 的平均意见分提高了 +0.14。在自动语音识别定制任务中，与两个强大的定制自动语音识别基线相比，我们的方法的词错误率相对降低分别达到 7.09% 和 10.35%。

索引词：文本转语音、大语言模型、VQ-GAN、音频编解码器

1. 引言

文本转语音（TTS）可以通过生成合成语音，并利用文本与语音配对数据来更新自动语音识别（ASR）模型，或使用基于 TTS 数据解码出的假设文本训练拼写纠正模型 [1, 2, 3, 4, 5, 6, 7, 8, 9]，从而帮助降低自动语音识别定制中的词错误率（WER），尤其是在低资源或词汇表外（OOV）场景下。然而，与自动语音识别训练集中的真实语音相比，合成语音的适应性能较差，因为它缺少复杂的录音条件、多样且真实的说话人差异，以及适用于不同语言语境的动态韵律语音。这揭示了神经文本转语音的不足：与通常拥有多出 1000 倍数据的自动语音识别应用相比，其训练所用的数据集规模相对较小且较为干净。因此，合成语音越能贴近自动语音识别训练集中的真实语音，定制效果就越好。拥有大规模数据集和多样化语音输出的文本转语音模型，是借助合成语音实现更佳自动语音识别定制的关键因素。

近年来，神经文本到语音合成借助级联或端到端的 Transformer 或 LSTM 架构，在自然度和音频保真度方面已达到接近人类的质量。级联模型分别优化声学模型和声码器，并采用具有自回归 [12] 或非自回归 [16] Transformer层的结构；而端到端（E2E）TTS 则通过联合训练声学模型和声码器，利用传统的梅尔频谱图[15] 或自动学习的语音表示 [18]来提升语音质量，从而简化模型构建。预测得到的帧级特征通过采用对抗训练 [23, 24, 25, 26]的神经声码器转换为原始波形。DelightfulTTS 2 [18] 提出了一种基于采用对抗训练的矢量量化自编码器的新型编解码器网络，用于自动学习帧级语音表示。它不使用梅尔频谱图或其他预先设计的特征，而是利用编解码器编码器提取语音表示，使用残差向量量化器对其进行量化，然后通过编解码器解码器在对抗训练下重建波形。尽管上述工作在小型且干净数据集上实现了较高质量，但由于声学模型与声码器之间存在复杂且连续的声学特征依赖关系，因此难以扩展到更大规模或更嘈杂的数据集，同时实现多样性控制。

采用连续语音特征的 TTS 很难解决一对多问题，因为难以在帧级语音特征上实现输出随机性，从而在韵律和说话人方面实现可观察的多样性；而梅尔频谱图、线性频谱图或自行提取的语音特征等特征在时间轴和频率轴上高度相关，这会导致声学模型在预测时出现过拟合或过度平滑效应。人类语音天然具有类似源滤波器声码器中白噪声的随机采样效应 [27,28, 29, 30, 31] ，仅凭语言或声学条件不足以模拟这种效应。由于多样化控制比语言模型或图像模型更容易实现，一些研究利用离散语音令牌探索不同的架构，以替代基于连续特征的模型。 [32] 提出了一种基于 VQ-VAE [33]和神经机器翻译（NMT）的端到端文本到语音（E2E TTS）框架，其中 VQ-VAE 模型学习将语音波形映射为离散词元序列，而基于 Transformer 的 NMT 模型则训练用于根据给定输入估计离散词元序列。由于使用了离散语音令牌，该模型可以像 NMT 和自动语音识别一样在每一步进行采样，从而避免过度平滑问题。VQTTS [34] 由一个称为 txt2vec 的声学模型（AM）和一个称为 vec2wav 的声码器组成，它使用自监督向量量化（VQ）声学特征，而不是梅尔频谱图；其中，t xt2vec 是采用交叉熵损失的分类模型，而非回归模型，vec2wav 则使用额外的特征编码器来平滑这种

*本文作者对本研究贡献相同。通讯作者：RuiqingXue, ruiqingxue@microsoft.com

不连续量化特征。txt2vec 不必预测复杂的梅尔频谱图，只需考虑沿时间轴的相关性，从而缩小真实语音帧与预测语音帧之间的差距。尽管可以在推断时通过采样实现多样性，但这些离散模型通常使用非常小且相对干净的数据集，例如广受欢迎的 JUST 或 LJ Speech [35, 36], ；然而，对于大规模且噪声更多的数据集，尤其是在自动语音识别定制中的多样化语音输出方面，尚未进行系统性探索。

基于大语言模型（LLM）的声学模型在大规模语音数据集训练方面取得了显著进展。语音录音通常包含声学或语义抽象特征，例如细粒度的音素级或词级韵律、句法、语法属性，或说话人识别、风格类别、口音类型等粗粒度属性。神经编解码器模型借助自回归波形建模 [37, 38, 39] 和对抗训练 [40, 16, 41, 42]等方法，已实现接近录音质量的效果。语言模型也展现出对不同内容类型的高层次长期序列进行建模的能力，近期在文本 [43, 44, 45, 46, 47] 和图像 [48, 49, 50, 51, 52, 53]领域取得的进展便证明了这一点。音频生成 [54, 55, 56,57, 58, 59] 正朝着自然的音频合成方向发展，使音频在整个过程中保持自然且一致。[54]表明，基于离散化语音单元训练的 Transformer [60] 解码器，无需依赖文本音素即可生成连贯语音：它可以通过语音提示或音乐提示序列持续生成，并呈现动态的声学多样性、语言一致性和高音频保真度。然而，提示内容仅限于语音单元，尚未探索说话人或文本提示。

在这项工作中，我们提出了 FoundationTTS，这是一个能够在大规模语音数据集上进行长期语言与声学依赖建模，从而实现高质量、多样化文本转语音生成的框架。正如我们的实验所示，由于语言模型采用了离散采样机制，FoundationTTS 的语音输出比传统模型更加多样。我们通过结合对抗式神经音频编解码器、端到端神经文本转语音 [18], 和大语言建模 [43]领域的最新进展，实现了这一目标。具体而言，我们从原始音频波形出发，首先使用细粒度音频编解码器重建波形；其中，编解码器编码器通过一组残差量化器从波形中提取帧级语音表示，这些表示能够捕捉音频波形的声学细节，并支持编解码器解码器实现高质量重建。随后，我们构建另一种粗粒度语音编解码器，通过少量量化器重建细粒度语音表示，以获得声学／语音词元。最后，我们训练了一个仅含解码器的语言模型，该模型自回归地预测来自粗粒度语音编解码器的声学词元；这种编解码器能够捕捉局部依赖（例如语音中的语音学特征）和全局长期结构（例如语音中的语义内容）。对语言词元和声学词元进行建模，既能实现高音频质量，也能保持长期一致性。总而言之，我们的贡献如下：

- 我们提出了 FoundationTTS，这是一种文本转语音生成框架，它将离散语音标记以分层方式与大型语言模型相结合，从而在大规模、复杂的语音数据集上实现长期一致性和多样化输出。
- 我们提出了一种分层编解码器网络，其中细粒度编解码器利用帧级语音表示重建波形，而粗粒度编解码器使用更少的量化器重建语音表示，将连续表示转换为离散语音标记，以进行语言建模。
- 我们展示了 FoundationTTS 的能力：它能够生成语言一致、韵律多样且高质量的语音，适用于自动语音识别领域定制和通用神经文本转语音。

2. FoundationTTS

FoundationTTS 是一种两阶段文本转语音模型。在第一阶段，我们引入一种能够编码和重建波形的分层音频编码；更具体地说，它由两个部分组成：1）用于连续语音帧提取的细粒度音频编解码器；2）用于离散语音令牌提取的粗粒度音频编解码器。在第二阶段，我们基于第一阶段音频编解码器生成的离散语音令牌序列训练前缀语言模型。我们采用前缀语言模型而非编码器-解码器范式，因为它能够简化模型结构，并且已在类似 GPT 的模型中取得巨大成功。

2.1. 采用分层音频编解码器的语音分词器

神经音频编解码器使用更高的比特率，通过帧级离散分词器重建高质量波形，使其接近真实语音。这会增加每帧的离散语音令牌数量，需要使用多个残差量化器；但若直接使用较小的量化器，则难以重建波形，因为真实音频样本包含噪声，十分复杂。此外，语言模型在过长范围的上下文建模方面表现不佳，而基于字符或词片段的分词器，其处理单位又远短于语音帧。因此，细粒度编解码器的帧与粗粒度语言模型的建模单位之间，需要在音频编解码器重建质量和语言模型性能之间取得平衡。为解决这一矛盾，如图2所示，我们设计了一种用作语音分词器的分层音频编解码器：首先，细粒度音频编解码器通过卷积编码器将波形编码为连续特征，并通过残差量化器从瓶颈层提取特征，再由镜像解码器重建波形；然后，粗粒度音频编解码器使用更少的量化器，进一步压缩细粒度编解码器生成的连续表示，从而缩短语音令牌序列。

2.1.1. 使用 *VQGAN* 的细粒度编解码器

细粒度编解码器类似于 DelightfulTTS 2 [18], 中的帧级特征提取网络，由一个波形下采样编码器、一个与之镜像对称的上采样解码器，以及一个充当特征瓶颈的残差向量量化器组成。编解码器编码器以 16kHZ 波形为输入，通过卷积层将其下采样为帧级表示，并通过下采样层实现 600 倍压缩。随后，使用包含多个向量量化器的残差向量量化器（RVQ）对帧级语音表示进行量化；因此，我们能够将 16kHz 连续波形采样压缩为 26.7 Hz 的离散语音标记，码率为 3.4 kbps，帧长为 37.5ms。这样一来，每个语音帧可以被标记为多个离散词元，但像 [54],那样在每一步展平多个语音词元并从粗量化器层自回归地进行预测，会导致一次预测多个离散词元的速度较慢。

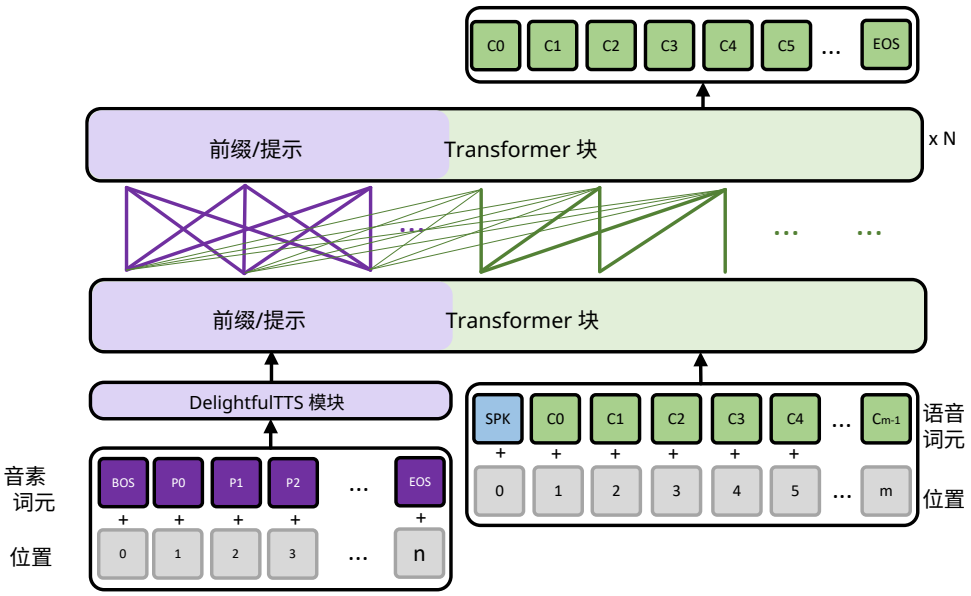


图1： *FoundationTTS* 的前缀语言模型。*DelightfulTTS* 模块充当音素编码器，以音素令牌和位置嵌入作为输入。前缀中还添加了一个说话人令牌，记作 *SPK*。一组 *Transformer* 解码器利用完整的音素令牌和说话人令牌，对语音令牌进行自回归解码。

2.1.2. 用于语言模型令牌提取的粗粒度音频编解码器

粗粒度编解码器旨在使用较少的量化器，从细粒度编解码器重建帧级特征。细粒度编解码器使用更多量化器能够更好地重建波形，但也会增加每个语音帧包含的语音令牌数量，使基于离散词元进行语言建模变得困难。粗粒度编解码器基于预训练的细粒度编解码器，在令牌数量与波形重建质量之间取得平衡。细粒度编解码器输出的帧级声学特征，作为粗粒度编解码器的重建目标。首先，将细粒度的帧级语音声学特征输入改进后的 Conformer 编码器 [16], , 然后使用单层向量量化器对输出进行量化。最后，另一个 Conformer 解码器会重建细粒度音频编解码器的声学特征。重建细粒度语音帧而非波形，旨在简化预测目标，从而只需较少的量化器就能近似细粒度编解码器表示的语音。为进一步提升重建能力，我们不直接重建特征，而是让解码器预测细粒度残差向量量化（RVQ）的权重分布。实验发现，这种方法能带来更好的词元化效果，并加快训练收敛。

对于细粒度音频编解码器和粗粒度音频编解码器，我们都使用了 DelightfulTTS2 [18], 中描述的相同判别器，这些判别器结合了多尺度和多周期判别器。

2.2. TTS 语言建模

传统神经文本转语音中的声学模型旨在预测连续语音特征，例如 TransformerTTS、FastSpeech 和 DelightfulTTS [12,13, 16] 所使用的梅尔频谱图，或 NaturalSpeech 和 DelightfulTTS 2 [22, 18] 直接从线性频谱或波形中学习的声学特征。尽管这些模型能够合成高质量的自然语音，但仍存在两个明显的缺点：1）难以利用复杂的现实世界语音或大规模语音数据集进行声学建模，例如自动语音识别语料库中噪声较多的数据，或有声读物，而不是相对干净的录音语音。2）从连续的帧级特征合成多样化的语音输出并不直接，这不利于缓解文本转语音中的一对多映射问题。通常需要采用一些可选方法来缓解多样性问题。第一类方法是在音素序列输入之外，为声学模型提供额外的声学或语言条件，例如基频或音素嵌入；[61, 62, 63], 中还将不同粒度的韵律特征加入了文本转语音。[64, 65]中也探索了矢量量化变分自编码器的有效性。另一种解决该问题的方法是，除了用于声学模型的常见 L1 或 L2 损失准则之外，利用更好的训练准则。然而，真实的声学分布要复杂得多。因此，一些研究使用归一化流 [6, 67] 来解决这一问题。归一化流将数据分布转换为已知的简单分布，并通过最大对数似然进行优化。然而，变分自编码器和流模型都需要经过精心设计，以确保可逆性和稳定性，这限制了此类模型的可扩展性。

如图1所示，为克服这些问题，我们使用离散语音标记作为建模单元，并设计了一个前缀语言模型来预测这些离散标记。具体而言，语音语言模型由一个 DelightfulTTS 编码器模块和一个仅解码器 Transformer 模块组成。DelightfulTTS 编码器从音素或文本序列中提取语言学表示，而仅解码器变换器则以语言学表示作为前缀，自回归地生成离散声学词元。语言学前缀和离散语音标记被输入同一个 Transformer 解码器，且不使用交叉注意力层。我们使用不同的位置嵌入和词元来区分二者。在语言学前缀上使用双向自注意力，在离散语音标记上使用单调自注意力。说话人

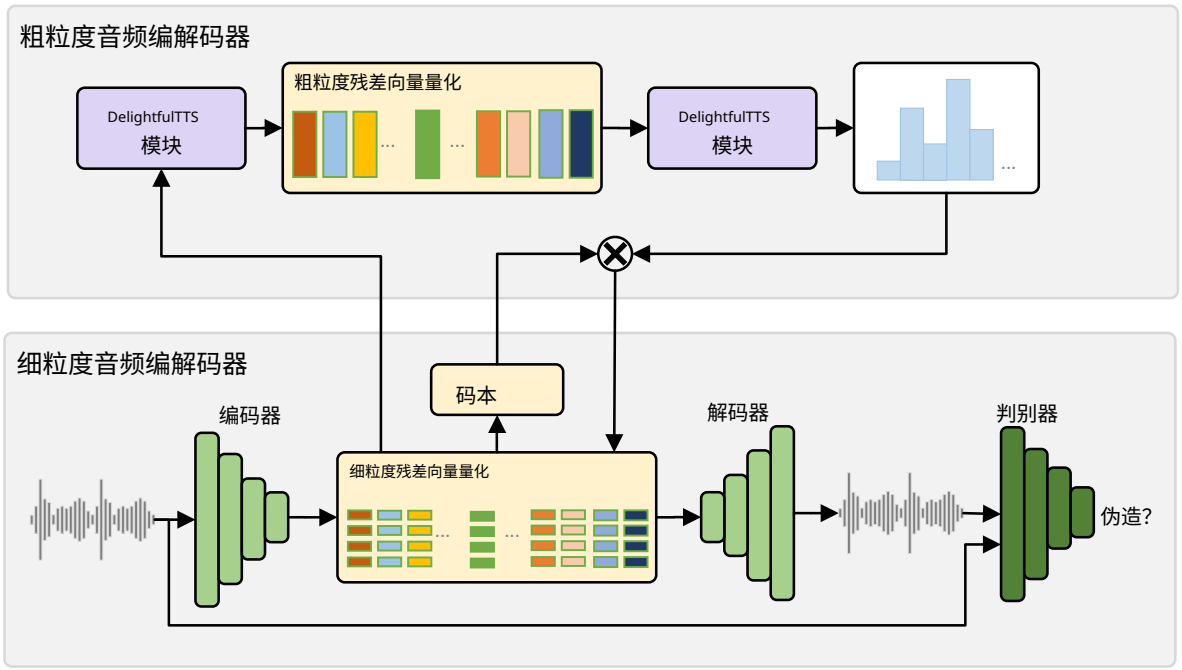


图2： 分层编解码网络由细粒度编解码器和粗粒度编解码器组成。细粒度编解码器由卷积下采样编码器、残差向量量化模块和卷积上采样解码器构成。卷积下采样编码器从音频波形中提取帧级语音表示，残差向量量化模块对帧表征进行离散化。粗粒度编解码器的结构与细粒度编解码器相似，但使用 *DelightfulTTS* 模块代替卷积网络，且残差向量量化层的数量为 *1*。

嵌入也作为说话人前缀添加到 Transformer 解码器中。据观察，就自然度或错误案例率而言，较大的语言模型比小型模型更加稳定，例如跳过音素或重复音素等问题。

2.3. 训练目标

编解码器损失 遵循 DeflightuflTTS 2，我们使用对抗目标训练分层音频编解码器。

$$L_{\text{FCODEC}} = L_{\text{adv}} + L_{\text{rvq}} + L_{\text{fm}} + L_{\text{mrs}} + L_{\text{msd}} + L_{\text{mpd}} \quad (1)$$

细粒度编解码器损失包括多周期判别器损失 L_{mpd}、多尺度判别器损失 L_{msd}、对抗损失 L_{adv}、多分辨率频谱图损失 L_{mrs}、特征匹配损失 L_{fm} 、残差向量量化损失 L_{rvq}，并使用随机分段波形。

$$L_{\text{CCODEC}} = L_{\text{adv}} + L_{\text{vq}} + L_{\text{fm}} + L_{\text{mrs}} + L_{\text{msd}} + L_{\text{mpd}} \quad (2)$$

粗粒度编解码器使用重建后的细粒度帧，通过冻结的细粒度语音编解码器，并生成重建波形；其损失与多周期判别器损失 L_{mpd}、多尺度判别器损失 L_{msd} 相同，同时还包括对抗损失 L_{adv}、多分辨率频谱图损失 L_{mrs}、特征匹配损失 L_{fm} 和向量量化损失 L_{vq}。

前缀语言模型损失 给定一个音频波形 $x = \{x_1, ..., x_{t_x}\}$，分层编解码器会将其压缩为离散语音标记，作为语言模型的目标 $c = \{c_1, ..., c_{t_c}\}$；并使用 Delight- fulTTS 模块将音素序列映射为语言学表示 $w = \{w_1, ..., w_{t_w}\}$，这样语音合成任务就可以视为一个使用交叉熵损失进行训练的语言模型问题。

$$L_{LM} = -\mathbb{E} \sum_{t=1}^{T_c} \log p_\theta \left(c_t | c_1, ..., c_{t-1}, w_1, ..., w_{t_w} \right) \quad (3)$$

其中， T_c 是帧数， t_w 是音素序列的长度， θ 是语言模型的可训练参数。

2.4. 推断

推断时，如图3所示，首先将输入音素序列送入前缀语言模型中的 DelightfulTTS 模块，以提取语言学表示；随后，仅解码器变换器根据这些语言学特征和说话人嵌入，自回归地生成语音令牌。粗粒度编解码器解码器中的另一个 DelightfulTTS 模块会将离散语音标记转换为细粒度码本中 RVQ标记的帧级权重分布，然后根据该分布生成相应的细粒度语音令牌。最后，细粒度编解码器解码器根据这些细粒度语音令牌重建波形。我们采用 Top-p采样生成多样化语音。

3. 实验

本节介绍所提出的 FoundationTTS 的实验数据集、训练和评估设置，并采用文本转语音和自动语音识别指标进行评估。

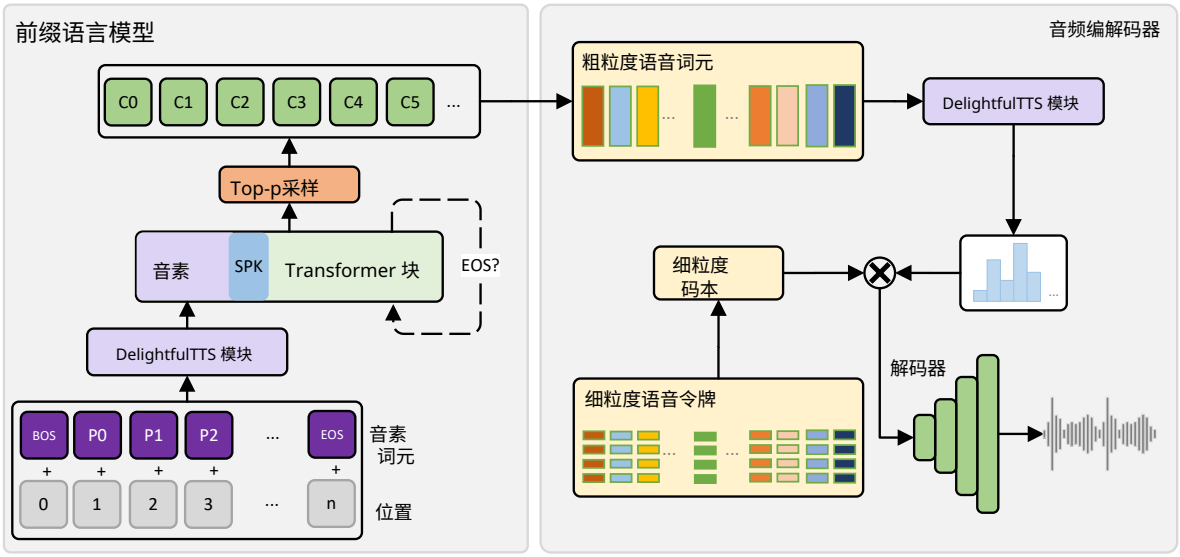


图3： 推断时，首先将音素令牌输入 *DelightfulTTS* 编码器，以提取语言学表示；仅解码器 *Transformer* 根据这些语言学特征和说话人令牌，以自回归方式生成语音令牌。我们采用 *Top-p*采样生成多样化的语音。最后，分层编解码器解码器根据预测的语音令牌重建音频波形。

3.1. 实验设置

数据集为评估自动语音识别定制的性能，我们在内部中文上海方言数据集和英语数据集上进行了实验。内部中文上海方言数据集包含 600 小时的领域专用语料库，其转录文本通过基于规则的字素到音素转换模块转换为音素。由于转录文本中包含多音字，且基于规则的字素到音素转换模块无法准确消歧，因此将多音字保留为额外词元。英语数据集包含 900 小时的语料库，由 LibriTTS、VCTK 和内部文本转语音数据组成。 [68, 69], 文本经过标准文本转语音前端数据处理流程，包括基于规则的文本规范化和字素到音素转换。音频波形的采样率为 16kHz。

训练设置FoundationTTS由三个依次训练的组件构成。首先，使用纯音频数据训练细粒度音频编解码器。其次，使用预训练的细粒度音频编解码器训练粗粒度音频编解码器。最后，使用粗粒度音频编解码器提取的语音令牌训练前缀语言模型。训练分层编解码器和前缀语言模型时，均使用8张32GB V100 GPU。在分层编解码器训练期间，我们将波形随机切分为每段24000个采样点，以加快模型收敛。分层编解码器和前缀语言模型的训练均采用Adam优化器，并结合RAdam和Lookahead优化 [70, 71]；训练前缀语言模型时，我们将学习率设为0.001，并从4,000次迭代开始按指数衰减至0.00005。细粒度编解码器的编码器和解码器各包含四个卷积层，中间使用16个残差量化器。粗粒度编解码器的编码器和解码器各包含三个改进的Conformer模块，并在编解码器编码器之后使用一个量化器。前缀语言模型包含16个Transformer层，以及六个改进的Conformer模块作为音素编码器，参数总量约为600M。

评估设置 我们使用词错误率（WER）指标，在中文上海方言和英语测试集上，通过两个独立的自动语音识别系统 Hybrid 和 RNN-T [72, 73] 评估 FoundationTTS 的适应性能。我们还使用客观和主观指标评估 FoundationTTS 的合成音频质量以及分层编解码器的重建质量。主观指标包括 FoundationTTS 的平均意见分（MOS）和编解码器的成对比较平均意见分（CMOS）；粗粒度音频编解码器和细粒度音频编解码器的 ViSQOL [74]分数则作为客观指标进行报告。用于主客观指标评估的文本转语音和编解码器评估集从训练集中随机留出；自动语音识别适应评估集与训练集没有重叠。

3.2. 结果

3.2.1. 自动语音识别适应性能

混合式自动语音识别适配 表 1 展示了采用 FoundationTTS 领域特定适配合成数据后，混合式自动语音识别系统的词错误率降低情况。基线混合式系统是一个预训练的中文上海方言模型。需要注意的是，基线模型已经使用领域特定数据进行了适配；即便如此，我们的模型仍实现了 7.09% 的词错误率降低。评估集包含 1000 个句子。

RNN-T适配 我们在 RNN-T自动语音识别模型上开展了另一项 FoundationTTS 适配实验。作为比较，我们还列出了 FastSpeech适配的词错误率。FoundationTTS 适配和 FastSpeech适配均使用相同的领域特定数据集进行训练。表 2 显示，与 RNN-T基线相比，FoundationTTS 适配实现了 40.13% 的词错误率降低；与 FastSpeech适配相比，实现了 10.35% 的词错误率降低，这表明

	女性	Male	ALL
混合基线	9.96%	10.63%	10.29%
+FoundationTTS适配	8.88%	10.26%	9.56%
WERR	10.84%	3.48%	7.09%

表 1：上海方言测试集上的自动语音识别适应词错误率相对降低。

FoundationTTS 能够合成比传统的基于非自回归连续声学帧的文本转语音模型更加多样且逼真的音频。评估集包含 1032 个句子。在 Hybrid 和 RNN-T自动语音识别模型上的适配实验表明，合成数据具有足够的鲁棒性，能够适用于不同的自动语音识别模型，并持续实现词错误率降低。

方法	WER
RNN-T基线	15.05%
+FastSpeech适配	10.05%
+FoundationTTS适配	9.01%

表 2： *RNN-T*、*FastSpeech* 和 *FoundationTTS* 在英语适应测试集上的词错误率比较。

系统	MOS
FastSpeech	3.84 ± 0.09
FoundationTTS	3.98 ± 0.08
录音	4.45 ± 0.10

表3： *FoundationTTS* 与使用 *900* 小时多样化语料库训练的 *FastSpeech* 的平均意见分对比。

3.2.2. 文本转语音质量评估

主观指标 表3展示了通过平均意见分（MOS）对FoundationTTS与FastSpeech进行的比较。FoundationTTS和FastSpeech使用相同的训练数据进行训练，该训练数据由900小时的英语语料库和7000多名说话人组成。评估集包含80个句子，每个样本由10名评审员进行评分。结果表明，在自然度方面，FoundationTTS优于FastSpeech。

编解码器的客观指标我们在表4中报告了细粒度编解码器和粗粒度编解码器的 ViSQOL 得分。评估集包含 2939 个句子。细粒度编解码器和粗粒度编解码器的比特率分别为 3.42 和 0.27。我们可以观察到，细粒度编解码器的 ViSQOL 得分比粗粒度编解码器高出 +0.7，但粗粒度编解码器仅使用了细粒度编解码器 7.9% 的比特率；细粒度编解码器较高的比特率带来了更好的音频重构质量，而粗粒度编解码器较低的比特率则产生了更短的离散序列，更适合大语言建模。细粒度编解码器与粗粒度编解码器的结合平衡了重构质量和建模效率。需要注意的是，我们曾尝试直接在波形上训练粗粒度编解码器，但模型无法收敛。这表明，细粒度编解码器简化了音频波形的特征表示，并且使粗粒度编解码器更易于训练。

编解码器的主观指标 我们还报告了细粒度编解码器重建的音频与录音之间的成对比较平均意见分（CMOS），如表5所示。总体平均分为-0.06，表明细粒度编解码器能够重新合成高保真的语音波形。

编解码器	比特率（kbps）	ViSQOL
粗粒度编解码器	0.27	3.07
细粒度编解码器	3.42	3.77

表4： 分层音频编解码器的 *ViSQOL* 对比。

系统	CMOS
编解码器重建与录音对比	-0.06

表5： 细粒度编解码器重建的成对比较平均意见分。

3.2.3. 消融分析

参数数量会影响语言模型的稳定性和音频质量。我们训练了一系列不同规模的语音词元语言模型，参数量从 20M 到 600M 不等。结果表明，参数数量会显著影响端到端质量；最小的 20M 模型完全无法收敛。更大、更多样化的训练数据需要更大的模型容量，这一点也得到了 GPT系列模型在语言领域的验证。

本研究使用的粗粒度编解码器采用一个量化器。我们对粗粒度编解码器使用不同数量的量化器进行了实验，结果表明，量化器越多，音频重建质量越好；但每个语音帧也会对应更多词元 ID，这不适用于语言建模。尤其是对于噪声较大的数据集，量化器较少的粗粒度编解码器难以更好地重建具有高信噪比或韵律变化丰富的语料库中的音频。

4. 结论

本文介绍了 FoundationTTS，这是一种神经文本转语音系统，将分层编解码网络与基于连续语音帧的对抗式矢量量化自编码器，以及基于离散语音标记的语言模型相结合。该系统无需预先设计梅尔频谱图或线性刻度频谱图等声学特征，便能通过对大规模、复杂的真实语音数据集进行建模，实现多样化的语音输出。FoundationTTS 在两项自动语音识别适应任务上的质量优于基线系统。未来，我们计划探索更大的数据集，并在更具挑战性的数据集（如多说话人或多语言数据集）上使用语言模型，以探究逼真语音合成场景中的挑战和潜在应用，以及自动语音识别适应。

FoundationTTS 能够合成高质量语音，其自然度和说话人相似度与人类语音十分接近，韵律丰富流畅。然而，这种先进能力也带来了潜在风险，例如特定个人的声音克隆可能被滥用。为应对这些风险，我们致力于在后续开发中贯彻微软人工智能原则，

这些模型适用于各种场景。[1]。

5. 参考文献

[1] J. Li、R. Zhao、Z. Meng、Y. Liu、W. Wei、S. Parthasarathy、V. Mazalov、Z. Wang、L. He、S. Zhao 等人，“开发具备定制能力、性能超越高性能混合模型的 RNN-T 模型”，*arXiv* 预印本 *arXiv:2007.15188*，2020年。

[2] K. C. Sim、F. Beaufays、A. Benard、D. Guliani、A. Kabel、N. Khare、T. Lucassen、P. Zadrazil、H. Zhang、L. Johnson 等人，“面向命名实体的移动设备端到端语音识别个性化”，载于*2019 IEEE* 自动语音识别与理解研讨会（*ASRU*）。IEEE，2019，第 23–30 页。[3] X. Wang、Y. Liu、S. Zhao 和 J. Li，“用于定制基于变换器的语音识别系统的轻量级上下文拼写纠正模型”，*arXiv* 预印本 *arXiv:2108 07493*，2021。。[4] X. Wang、Y. Liu、J. Li、V. Miljanic、S. Zhao 和 H. Khalil，“面向端到端语音识别系统定制的上下文拼写纠正”，*arXiv* 预印本 *arXiv:2203.00888*，2022。

[5] X. Wang, Y. Zhao, “通过外部声学注意力和语义感知的数据增强改进上下文拼写纠正”，*arXiv* 预印本 *arXiv:2302 11192*，2023。。

[6] G. Papamakarios、E. T. Nalisnick、D. J. Rezende、S. Mohamed 和 B. Lakshminarayanan，“用于概率建模和推断的归一化流。”机器学习研究杂志，第 22 卷，第 57 期，第 1–64 页，2021 年。[7] S. Ueno、M. Mimura、S. Sakai 和 T. Kawahara，“用于声学到词语音识别数据增强的多说话人序列到序列语音合成”，载于 *ICASSP 2019-2019 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2019，第 6161–6165 页。[8] A. Tjandra、S. Sakti 和 S. Nakamura，“具有单次说话人自适应的机器语音链”，*arXiv* 预印本 *arXiv:1803.10525*，2018年。

[9] M. Zhao, S. 三村原, “利用序列到序列的语音合成来增强声学到词语音识别”，载于*2018*年*IEEE*口语语言技术研讨会（*SLT*）。IEEE，2018年，第477–484页。

[10] Y. Wang、R. Skerry-Ryan、D. Stanton、Y. Wu、R. J. Weiss、N. Jaitly、Z. Yang、Y. Xiao、Z. Chen、S. Bengio 等，“Tacotron：迈向端到端语音合成”，*arXiv*预印本 *arXiv:1703.10135*，2017年。[11] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan 等，“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*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第4779–4783页。[12] N. Li、S. Liu、Y. Liu、S. Zhao和M. Liu，“使用Transformer网络的神经语音合成”，载于*AAAI*人工智能会议论文集，第33卷，第01期，2019，第6706–6713页。

[13] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao 和 T.-Y. Liu，“FastSpeech：快速、稳健且可控的文本转语音”，*arXiv* 预印本 *arXiv:1905.09263*，2019。

[14] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao 和 M. Liu，“Robutrans：一种基于Transformer的鲁棒文本转语音模型”，发表于*AAAI*人工智能会议论文集，第34卷，第05期，2020年，第8228–8235页。

[15] Y. Ren、C. Hu、X. Tan、T. Qin、S. Zhao、Z. Zhao 和 T.-Y. Liu，“Fastspeech 2：快速、高质量的端到端文本转语音”，发表于国际学习表征会议，2021年。 [线上]。可访问：https://openreview. net/forum?id=piLPYqxtWuA[16] Y. Liu、Z. Xu、G. Wang、K. Chen、B. Li、X. Tan、J. Li、L. He 和 S. Zhao，“Delightfultts：微软语音合成系统参加2021年暴雪挑战赛，*12612*，2021*2110*” *arXiv*预印本，*arXiv:* 。.[17] J. Donahue、S. Dieleman、M. Bi´nkowski、E. Elsen 和 K. Simonyan，“端到端对抗性文本转语音”，*arXiv*预印本，*arXiv:2006.03575*，2020年。[18] Y. Liu、R. Xue、L. He、X. Tan 和 S. Zhao，“Delightfultts 2：采用对抗性向量量化自编码器的端到端语音合成”，*arXiv*预印本，*arXiv:2207.04646*，2022年。[19] J. Kim、J. Kong 和 J. Son，“采用对抗学习的端到端文本转语音条件变分自编码器”，*arXiv*预印本，*arXiv:2106.06103*，2021年。[20] R. J. Weiss、R. Skerry-Ryan、E. Battenberg、S. Mariooryad 和 D. P. Kingma，“Wave-tacotron：无需频谱图的端到端文本转语音合成”，发表于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年，第5679–5683页。[21] X. Tan、T. Qin、F. Soong 和 T.-Y. Liu，“神经语音合成综述”，*arXiv*预印本，*arXiv:2106.15561*，2021年。[22] X. Tan、J. Chen、H. Liu、J. Cong、C. Zhang、Y. Liu、X. Wang、Y. Leng、Y. Yi、L. He 等，“Naturalspeech：具有人类水平质量的端到端文本转语音合成”，*arXiv*预印本，*arXiv:2205.04421*，2022年。[23] K. Kumar、R. Kumar、T. de Boissiere、L. Gestin、W. Z. Teoh、J. Sotelo、A. de Br´ebisson、Y. Bengio 和 A. Courville，“Mel-gan：用于条件波形合成的生成对抗网络”，*arXiv*预印本，*arXiv:1910.06711*，2019年。[24] J. Kong、J. Kim 和 J. Bae，“Hifi-gan：用于高效、高保真语音合成的生成对抗网络”，神经信息处理系统进展，第33卷，2020年。[25] J.-H. Kim、S.-H. Lee、J.-H. Lee 和 S.-W. Lee，“Fre-gan：对抗式频率一致音频合成”，*arXiv*预印本，*arXiv:2106.02297*，2021年。[26] W. Jang、D. Lim、J. Yoon、B. Kim 和 J. Kim，“Univnet：采用多分辨率频谱图判别器的神经声码器，用于高保真波形生成”，*arXiv*预印本，*arXiv:2106.07889*，2021年。

[27] Q. Hu, K. Richmond, J. Yamagishi 和 J. Latorre，“对多种声码器类型的实验比较”，载于第八届 *ISCA* 语音合成研讨会，2013年。

[28] H. Zen、T. Nose、J. Yamagishi、S. Sako、T. Masuko、A. W. Black 和 K. Tokuda，“基于隐马尔可夫模型的语音合成系统（HTS）2.0 版。”*SSW*，第 6 卷，第 294–299 页，2007年。[29] H. Kawahara、I. Masuda-Katsuse 和 A. De Cheveigne，“使用基于基频自适应的时频平滑和基于瞬时频率的基频提取，对语音表示进行重构：声音中重复结构可能发挥的作用”，《*Speech Communication*》，第 27 卷，第 3-4 期，第 187–207 页，1999年。[30] T. Drugman 和 T. Dutoit，“残差信号的确定性加随机模型及其应用”，*IEEE* 音频、语音与语言处理汇刊，第 20 卷，第 3 期，第 968–981 页，2011年。[31] K. Tokuda、T. Kobayashi、T. Masuko 和 S. Imai，“梅尔广义倒谱分析——一种语音频谱估计的统一方法。”载于*ICSLP*，第 94 卷，1994年，第 18–22 页。[32] T. Hayashi 和 S. Watanabe，“Discretalk：将文本转语音视为机器翻译问题”，*arXiv* 预印本 *arXiv:2005 05525*，2020年。。[33] A. Razavi、A. Van den Oord 和 O. Vinyals，“使用 VQ-VAE-2 生成多样化的高保真图像”，神经信息处理系统进展，第32卷，2019。

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

[34] C. Du、Y. Guo、X. Chen 和 K. Yu，《VQTTS：基于自监督 VQ 声学特征的高保真文本转语音合成》，*arXiv* 预印本 *arXiv:2204.00768*，2022年。[35] R. Sonobe、S. Takamichi 和 H. Saruwatari，《JSUT 语料库：用于端到端语音合成的免费大型日语语音语料库》，*arXiv* 预印本 *arXiv:1711.00354*，2017年。

[36] K. Ito 和 L. Johnson，《LJ 语音数据集》，https://keithito.com/LJ-Speech-Dataset/，2017年。

[37] A. v. d. Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior 和 K. Kavukcuoglu，《WaveNet：一种用于原始音频的生成模型》，*arXiv* 预印本 *arXiv:1609.03499*，2016年。[38] J.-M. Valin 和 J. Skoglund，《LPCNet：通过线性预测改进神经语音合成》，载于 *ICASSP 2019-2019 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2019年，第5891–5895页。

[39] N. Kalchbrenner、E. Elsen、K. Simonyan、S. Noury、N. Casagrande、E. Lockhart、F. Stimberg、A. Oord、S. Dieleman 和 K. Kavukcuoglu，《高效的神经音频合成》，载于国际机器学习会议。PMLR，2018年，第2410–2419页。

[40] N. Zeghidour、A. Luebs、A. Omran、J. Skoglund 和 M. Tagliasacchi，“Soundstream：一种端到端神经音频编解码器”，*IEEE/ACM* 音频、语音与语言处理汇刊，2021。

[41] M. Tagliasacchi、Y. Li、K. Misiunas 和 D. Roblek，“Seanet：一个多模态语音增强网络”，*arXiv* 预印本 *arXiv:2009.02095*，2020年。

[42] A. Défossez, J. Copet, G. Synnaeve, 和 Y. Adi，《高保真神经音频压缩》，*arXiv* 预印本 *arXiv:2210 13438*，2022。。

[43] T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell 等，《语言模型是少样本学习者》，神经信息处理系统进展，第 33 卷，第 1877–1901 页，2020年。

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever等人，“语言模型是无监督的多任务学习者”，*OpenAI* 博客，第 1 卷，第 8 期，第 9 页，2019。

[45] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P.Mishkin, C. Zhang, S. Agarwal, K. Slama, A.Ray 等人，“通过人类反馈训练语言模型遵循指令”，*arXiv* 预印本 *arXiv:2203.02155*，2022。

[46] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le，“微调后的语言模型是零样本学习者”，*arXiv* 预印本 *arXiv:2109.01652*，2021。[47] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann 等人，“PaLM：利用路径扩展语言建模”，*arXiv* 预印本 *arXiv:2204.02311*，2022。

[48] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge 和 Y. Wu，“向量量化图像建模”，*arXiv* 预印本 *arXiv:2110.04627*，2021。

[49] A. Ramesh、P. Dhariwal、A. Nichol、C. Chu 和 M. Chen，“基于 CLIP 潜变量的分层文本条件图像生成”，*arXiv* 预印本 *arXiv:2204.06125*，2022。[50] A. Ramesh、M. Pavlov、G. Goh、S. Gray、C. Voss、A. Radford、M. Chen 和 I. Sutskever，“零样本文本到图像生成”，载于国际机器学习会议。PMLR，2021 年，第 8821–8831 页。[51] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick，“掩码自编码器是可扩展的视觉学习器”，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022 年，第 16 000–16 009 页。[52] H. Bao、L. Dong 和 F. Wei，“Beit：图像变换器的 BERT 预训练”，*arXiv* 预印本 *arXiv:2106.08254*，2021 年。

[53] D. Lee、C. Kim、S. Kim、M. Cho 和 W.-S. Han，“使用残差量化的自回归图像生成”，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022 年，第 11 523–11 532 页。

[54] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, O. Teboul, D. Grangier, M. Tagliasacchi 和 N. Zeghidour，“AudioLM：一种用于音频生成的语言建模方法”，*arXiv* 预印本 *arXiv:2209.03143*，2022。

[55] F. Kreuk，G. Synnaeve，A. Polyak，U. Singer，A. D´efossez，J. Copet，D. Parikh，Y. Taigman 和 Y. Adi，“AudioGen：文本引导的音频生成”，*arXiv* 预印本 *arXiv:2209 15352*，2022。.[56] K. Lakhotia，E. Kharitonov，W.-N. Hsu，Y. Adi，A. Polyak，B. Bolte，T.-A. Nguyen，J. Copet，A. Baevski，A. Mohamed 等，“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，第9卷，第1336–1354页，2021。[57] F. Kreuk，Y. Taigman，A. Polyak，J. Copet，G. Synnaeve，A. D´efossez 和 Y. Adi，“使用感知引导的离散表示进行音频语言建模”，*arXiv* 预印本 *arXiv:2211.01223*，2022。

[58] S. Maiti、Y. Peng、T. Saeki 和 S. Watanabe，“Speechlmscore：使用语音语言模型评估语音生成”，*arXiv*预印本*arXiv:2212.04559*，2022。

[59] G. Maimon 和 Y. Adi，“利用离散自监督单元进行说话风格转换”，*arXiv* 预印本 *arXiv:2212.09730*，2022。[60] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“Attention is all you need”，载于神经信息处理系统进展，2017年，第5998–6008页。[61] A. Ła´ncucki，“FastPitch：基于基频预测的并行文本转语音”，载于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议*(ICASSP)*。IEEE，2021年，第6588–6592页。[62] Y. Wang、D. Stanton、Y. Zhang、R.-S. Ryan、E. Battenberg、J. Shor、Y. Xiao、Y. Jia、F. Ren 和 R. A. Saurous，“风格令牌：端到端语音合成中的无监督风格建模、控制与迁移”，载于国际机器学习会议。PMLR，2018年，第5180–5189页。[63] W.-N. Hsu、Y. Zhang、R. J. Weiss、H. Zen、Y. Wu、Y. Wang、Y. Cao、Y. Jia、Z. Chen、J. Shen 等，“用于可控语音合成的分层生成建模”，*arXiv* 预印本 *arXiv:1810.07217*，2018年。[64] G. Sun、Y. Zhang、R. J. Weiss、Y. Cao、H. Zen 和 Y. Wu，“用于可解释语音合成的全分层细粒度韵律建模”，载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2020年，第6264–6268页。[65] C.-M. Chien 和 H.-y. Lee，“非自回归语音合成的分层韵律建模”，载于*2021*年 *IEEE* 口语技术研讨会 *(SLT)*。IEEE，2021年，第446–453页。[66] X. Wang、S. Takaki、J. Yamagishi、S. King 和 K. Tokuda，“用于统计参数语音合成的向量量化变分自编码器 (VQ-VAE) 自回归神经f 0 模型”，*IEEE/ACM* 音频、语音与语言处理汇刊，第28卷，第157–170页，2019年。[67] J. Kim、S. Kim、J. Kong 和 S. Yoon，“Glow-TTS：通过单调对齐搜索实现文本转语音的生成流模型”，神经信息处理系统进展，第33卷，第8067–8077页，2020年。[68] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen 和 Y. Wu，“LibriTTS：一个由 LibriSpeech 衍生的文本转语音语料库”，*arXiv* 预印本 *arXiv:1904.02882*，2019年。[69] C. Veaux、J. Yamagishi、K. MacDonald 等，“更新版 CSTR VCTK 语料库：适用于 CSTR 语音克隆工具包的英语多说话人语料库”，2016年。

[70] L. Liu、H. Jiang、P. He、W. Chen、X. Liu、J. Gao 和 J. Han，“关于自适应学习率方差及其他问题”，*arXiv* 预印本 *arXiv:1908.03265*，2019。[71] M. Zhang、J. Lucas、J. Ba 和 G. E. Hinton，“Lookahead 优化器：前进 k 步，后退 1 步”，神经信息处理系统进展，第32卷，2019年。[72] A. Graves，“利用循环神经网络进行序列转导”，*arXiv* 预印本 *arXiv:1211.3711*，2012年。[73] K. Rao、H. Sak 和 R. Prabhavalkar，“探索用于基于 RNN-T 的流式端到端语音识别的架构、数据和单元”，载于 *2017*年 *IEEE* 自动语音识别与理解研讨会 *(ASRU)*。IEEE，2017年，第193–199页。[74] A. Hines、J. Skoglund、A. C. Kokaram 和 N. Harte，“ViSQOL：一种客观语音质量模型”，*EURASIP* 音频、语音与音乐处理期刊，第2015卷，第1期，第1–18页，2015年。