AUDIOGEN: TEXTUALLY GUIDED AUDIO GENERATION

Felix Kreuk[1], **Gabriel Synnaeve**[1], **Adam Polyak**[1], **Uriel Singer**[1], **Alexandre Défossez**[1], **Jade Copet**[1], **Devi Parikh**[1], **Yaniv Taigman**[1], **Yossi Adi**[1,2]

[1]FAIR Team, Meta AI
[2]The Hebrew University of Jerusalem
felixkreuk@meta.com

ABSTRACT

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。

项目Github地址: https://github.com/binary-husky/gpt_academic/。

项目在线体验地址: https://chatpaper.org。

当前大语言模型: gpt-3.5-turbo，当前语言模型温度设定: 1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

我们致力于解决根据描述性文本生成音频样本的问题。在本研究中，我们提出了AUDIOGEN，一种基于自回归的生成模型，可以根据文本输入生成音频样本。AUDIOGEN 在学习的离散音频表示上操作。文本到音频的生成任务面临着多重挑战。由于声音经过介质传播的方式，区分“对象”可以是一项困难的任务（例如，同时分离多人说话）。这在现实世界的录音条件下进一步复杂化（例如，背景噪声，混响等）。稀缺的文本注释施加了另一个限制，限制了模型扩展的能力。最后，对高保真音频的建模需要在高采样率下对音频进行编码，导致序列变得极长。为了缓解上述挑战，我们提出了一种混合不同音频样本的增强技术，驱使模型内部学习将多个声源分离。我们筛选了包含不同类型音频和文本注释的10个数据集，以处理文本-音频数据点的稀缺性。为了更快的推断，我们探索了多流建模的使用，允许使用较短的序列同时保持类似的比特率和感知质量。我们应用了无分类器的指导来提高对文本的遵循程度。与评估的基准方法相比，AUDIOGEN 在客观和主观指标上均取得了更好的表现。最后，我们探索了所提方法在有条件和无条件下生成音频延续的能力。样本：https://felixkreuk.github.io/audiogen。

1 INTRODUCTION

神经生成模型挑战了我们创造数字内容的方式。从生成高质量图像 (Karras et al., 2019; Park et al., 2019)和语音 (Ren et al., 2021; Oord et al., 2016)，到生成长文本片段 (Brown et al., 2020; Zhang et al., 2022)，再到最近提出的文本引导图像生成 (Ramesh et al., 2022; Rombach et al., 2022)，这些模型展现了令人印象深刻的结果。这引出了一个问题文本引导生成模型的音频等效物会是什么？从生成音景到音乐或语音，解决这个问题的一个解决方案是高保真度、可控性和多样性的输出，这将是现代电影、视频游戏和任何虚拟环境创作者的有用补充。

虽然图像生成和音频生成有很多共同之处，但也有一些关键区别。音频从本质上讲是一个一维信号，因此在区分重叠的“对象”时具有较少的自由度 (Capon, 1969; Frost, 1972)。真实世界的音频固有地具有混响效果，这使得从周围环境中区分对象的任务更加困难。此外，心理声学和心理视觉特性也有所不同，例如听觉“分辨率”（等响度）在频率上呈U形，4kHz处有一个低谷和8kHz处有一个凸起 (Suzuki et al., 2003)。最后但并非最不重要的是，具有文本描述的音频数据的可用性比文本-图像配对数据低几个数量级。这使得生成未见过的音频作品成为一项困难的任务（例如生成“太空中骑马的宇航员”的音频等效物）。

在这项工作中，我们解决了在文本描述条件下生成音频样本的问题。我们另外将所提出的方法扩展到条件和非条件音频延续。在上面的提示中，模型必须生成三个类别的声学内容，其中包括在时间轴上具有不同背景/前景、持续时间和相对位置的内容组合，这在训练集中极不可能出现。因此，生成这样的音频是一项在泛化、声学保真度、制作和制作方面具有挑战性的任务。

我们提出了AUDIOGEN，一种自回归的文本引导音频生成模型。AUDIOGEN 由两个主要阶段组成。第一个阶段使用神经音频压缩模型（例如 Zeghidour et al. (2021)）将原始音频编码为离散的令牌序列。该模型以端到端的方式训练，从压缩表示中重构输入音频，并添加一组鉴别器的感知损失。这样的音频表示被设计为生成高保真度的音频样本，同时还保持紧凑。第二阶段利用一个自回归的Transformer解码器语言模型，该模型作用于从第一阶段得到的离散音频令牌上，并同时对文本输入进行条件处理。我们使用一个在大量文本语料库上预训练的独立文本编码器模型（即T5）表示文本 (Raffel et al., 2020)。预训练的文本编码器使得能够概括当前文本-音频数据集中缺失的文本概念。当使用在多样性和描述性方面有限的文本注释时，这一点尤为重要。

与现有的文本到音频工作 (Yang et al., 2022)相比，AUDIOGEN 生成的样本具有更好的客观和主观指标。特别是，AUDIOGEN 创建了听起来更自然的未见过的音频作品。最后，我们通过使用剩余矢量量化（对声学单元）和多流Transformer来实现了声音保真度和采样时间之间的权衡，从而从实证上展示了所提出的方法如何扩展到考虑条件和非条件生成的音频延续。

我们的贡献： (i) 我们提出了一个基于文本描述或音频提示进行条件生成的最先进的自回归音频生成模型，该模型经过客观和主观（人工听众）评估得到。具体而言，我们提出了两种模型变体，一个有2.85亿参数，另一个有10亿参数； (ii) 我们通过在音频语言模型之上应用无分类器的指导来提高文本的黏合度。我们通过进行即时文本和音频混合来提高组合性；(iii) 我们展示了所提出的方法如何扩展到依赖文本和无条件的音频延续；(iv) 我们通过使用剩余矢量量化（用于声学单元）和多流Transformer来探讨音质和采样时间之间的权衡。

2 RELATED WORK

语音表示学习。对于无监督的语音表示学习研究，可以粗略地分为重构和自监督学习方法。自动编码是信号重构的常用方法，其中语音首先被编码为低维潜在表示，然后解码回语音。可以对编码空间施加各种约束，如时间平滑性 (Ebbers et al., 2017)、离散性 (van den Oord et al., 2017b)和层次结构 (Hsu et al., 2017)。语音的自我监督学习方法已经在自动语音识别 (Schneider et al., 2019; Baevski et al., 2020; Wang et al., 2021)、音素分割 (Kreuk et al., 2020)和音频压缩 (Zeghidour et al., 2021; Polyak et al., 2021)等方面取得了显著的成果。van den Oord et al. (2018)和Schneider et al. (2019)建议使用对比性预测编码（CPC）损失函数训练卷积神经网络，以区分真实未来样本和随机干扰样本。 Ao et al. (2022)提出了T5模型的语音版本，并展示了其在各种语音任务上的效果。与CPC类似，Baevski et al. (2020)使用了编码器和预测

器，通过对比训练来区分正样本和负样本。与 (Schneider et al., 2019)不同，它对编码器输出的片段进行了离散化和屏蔽。 Hsu et al. (2021)提出了 HuBERT 模型，该模型训练时使用了类似 BERT 的掩码预测任务，但是输入是连续的音频信号。 Chen et al. (2022)提出了一个类似的 HuBERT 版本，使用更大和增强的数据集进行训练。最近，Huang et al. (2022) 提出了一种基于掩码自动编码的方法来学习语音表示，并展示了其在几个音频分类任务中的有效性。

另一条相关的先前工作线索与建模音频离散表示有关。最近的研究表明，使用k-means量化 Self-Supervised Learning (SSL) 表示，并进行语言建模 (Lakhotia et al., 2021; Kharitonov et al., 2022a; Borsos et al., 2022)、多流处理 (Kharitonov et al., 2022b)、语音情感转换 (Kreuk et al., 2022)、口语对话 (Nguyen et al., 2022)和语音到语音翻译 (Lee et al., 2022a;b; Popuri et al., 2022)等任务。

文本到图像最近取得了很大的进展。DALL-E (Reddy et al., 2021)首先使用预训练的VQ-VAE将图像的补丁转化为离散编码。在训练过程中，表示图像补丁的编码附加到表示文本的编码中。然后，使用Transformer解码模型以自回归方式建模这些编码，而Gafni et al. (2022)提出了类似的方法，并增加了分割图以提高可控性。在Parti模型 (Yu et al.)中，作者建议使用Transformer的编码器-解码器架构将文本到图像的任务建模为序列到序列问题。

最近，扩散模型变得越来越受欢迎 (Nichol et al., 2022; Ramesh et al., 2022; Saharia et al.; Rombach et al., 2022)。DALLE-2 (Ramesh et al., 2022)使用扩散模型根据CLIP文本编码（先验）预测CLIP视觉特征，并使用另一个扩散模型根据预测的CLIP视觉特征预测图像像素（解码器）。使用一系列超分辨率模型将预测的图像上采样到更高的分辨率。Imagen (Saharia et al.)采用类似的方法，但省略了先验部分，而使用预训练的文本编码器，如T5 (Raffel et al., 2020)。**文本到音频。** 对我们的工作最相关的是Yang et al. (2022)提出的方法，他们提出了DiffSound，这是一个基于扩散过程的文本到音频模型，该模型在音频离散代码上运行。音频代码是通过在mel频谱图上训练的基于VQ-VAE的模型获得的(van den Oord et al., 2017a)。为了进一步提高模型性能，Yang et al. (2022)建议使用带有随机输入掩码的标签对扩散模型进行预训练。他们还探索了自回归Transformer解码器模型的使用，但发现该模型不及基于扩散的模型。

该提出的方法与DiffSound的不同之处在于：（i）我们的音频表示直接从原始波形中学习；（ii）我们使用数据增强创建新的音频组合，使模型能够从复杂的文本标题生成音频；（iii）我们在自回归设置下应用和研究了无分类器指导的效果；（iv）与Yang et al. (2022)的方法相反，我们凭经验证明了文本条件的自回归模型能够生成高质量的音频样本。

3 METHOD

提出的方法AUDIOGEN基于两个主要步骤：（i）使用自编码方法学习原始音频的离散表示；（ii）在音频编码器得到的学习编码的基础上，训练一个Transformer语言模型，该模型以文本特征为条件。在推理时期，我们从语言模型中采样，给定文本特征生成一组新的音频令牌。这些令牌之后可以使用步骤（i）的解码器组件解码成波形领域。该方法的视觉描述如图1所示。

3.1 AUDIO REPRESENTATION

一个持续时间为d的音频信号可以用一个序列$\boldsymbol{x} \in [-1,1]^{C_{\mathrm{a}} \times T}$表示，其中$C_{\mathrm{a}}$是音频通道的数量，$T = d \cdot f_{\mathrm{sr}}$是给定采样率$f_{\mathrm{sr}}$下的音频样本数量。在本工作中，我们设置$f_{\mathrm{sr}} = 16\mathrm{kHz}$。音频表示模型由三个组件组成：(i)编码器网络$E$，它以音频片段作为输入并输出潜在表示$\boldsymbol{z}$；

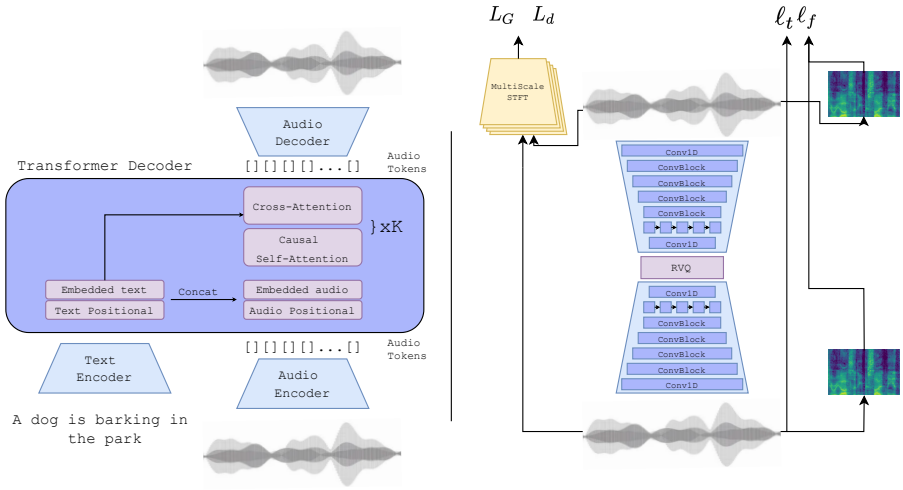


图 1: AUDIOGEN系统的总体概述如下图所示。左图：音频表示模型。右图：音频语言模型。文本和音频嵌入按时间维度进行拼接，然后与嵌入的文本一起输入具有 K 个因果自注意力和交叉注意力块的模型中。

(ii)量化层Q使用向量量化 (Vasuki & Vanathi, 2006)层生成一个压缩表示$\boldsymbol{z}_q$；(iii)解码器网络G从压缩的潜在表示$\boldsymbol{z}_q$中重构时域信号$\hat{\boldsymbol{x}}$。整个系统以端到端的方式进行训练，通过在时间域和频率域上应用重构损失以及以几个在不同时间分辨率下运行的鉴别器的形式的感知损失进行最小化。使用预训练模型，我们可以利用编码器和量化器组件作为离散特征提取器（即$Q \circ E$），并使用G将表示解码为时域信号。对于Q，我们使用一个包含2048个码字的单个码书，其中每个码字是一个128维向量。提出的方法的视觉描述可以在图 1（右侧）中看到。

架构。 我们采用类似于 Zeghidour et al. (2021); Li et al. (2021)中的自动编码器模型架构。编码器模型E由一个具有C通道的1D卷积和B个卷积块组成。每个卷积块由一个单一的残差单元组成，其后是一个由卷积核大小K为步长S的卷积层组成的下采样层。残差单元包含两个卷积和一个跳跃连接。每当进行下采样时，通道数量加倍。卷积块后面是一个用于序列建模的两层LSTM和一个最终的1D卷积层，其卷积核大小为7，输出通道数为D。我们使用$C = 32$，$B = 4$以及(2, 2, 2, 4)作为步长。我们使用ELU作为非线性激活函数 (Clevert et al., 2015)和LayerNorm (Ba et al., 2016)。解码器与编码器相反，使用转置卷积代替步进卷积，并以与编码器相反的步长顺序输出最终的音频。

训练目标。 我们优化一个基于GAN的训练目标，类似于 (Kong et al., 2020; Zeghidour et al., 2021)，共同最小化重构损失和对抗损失的组合。具体而言，我们在时间域上最小化目标音频和重构音频之间的L1距离，即$\ell_t(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1$。对于频率域损失，我们使用mel频谱图上的L1距离和L2距离的线性组合，使用几个时间尺度 (Yamamoto et al., 2020; Gritsenko et al., 2020)。形式上，

$$\ell_f(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{|\alpha| \cdot |s|} \sum_{\alpha_i \in \alpha} \sum_{i \in e} \|\mathcal{S}_i(\boldsymbol{x}) - \mathcal{S}_i(\hat{\boldsymbol{x}})\|_1 + \alpha_i \|\mathcal{S}_i(\boldsymbol{x}) - \mathcal{S}_i(\hat{\boldsymbol{x}})\|_2, \tag{1}$$

其中，$\mathcal{S}_i$是使用归一化的STFT计算的64个频谱区间的梅尔频谱图，窗口大小为2^i，跳跃长度为$2^i/4$，$e = 5, \ldots, 11$是尺度集合，α代表在L1和L2项之间平衡的标量系数集合。与Gritsenko et al. (2020)不同的是，我们设置$\alpha_i = 1$。

为了进一步提高生成样本的质量，我们额外优化了基于多尺度STFT（MS-STFT）鉴别器。多尺度鉴别器在捕捉音频信号中不同结构方面非常流行 (Kumar et al., 2019; Kong et al., 2020; You et al., 2021)。MS-STFT鉴别器基于对多尺度复值STFT进行操作的结构相同的网络，其中实部和虚部被连接。每个子网络由一个2D卷积层（使用大小为3 x 8的卷积核，32个通道）组成，随后是时间维度上具有不断增加的扩张率（1, 2和4）和在频率轴上步长为2的2D卷积。最后，通过大小为3 x 3和步长（1, 1）的2D卷积进行最终预测。我们使用了5个不同尺度的STFT窗口长度[2048，1024，512，256，128]。生成器的对抗损失构造如下，$\ell_g(\hat{\boldsymbol{x}}) = \frac{1}{K}\sum_k \max(0, 1 - D_k(\hat{\boldsymbol{x}}))$，其中$K$是鉴别器网络的数量。类似于之前在神经声码器上的工作 (Kumar et al., 2019; Kong et al., 2020; You et al., 2021)，我们还为生成器添加了特征匹配损失。形式上，

$$\ell_{feat}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} \|D_k^l(\boldsymbol{x}) - D_k^l(\hat{\boldsymbol{x}})\|_1, \tag{2}$$

其中，(D_k) 表示判别器，L 表示判别器的层数。

总体而言，判别器的训练目标是最小化以下损失函数：$L_d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{K}\sum_{k=1}^{K} \max(0, 1 - D_k(\boldsymbol{x})) + \max(0, 1 + D_k(\hat{\boldsymbol{x}}))$，其中 K 是判别器的数量；而生成器的训练目标是最小化以下损失函数：$L_G = \lambda_t \cdot \ell_t(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_f \cdot \ell_f(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_g \cdot \ell_g(\hat{\boldsymbol{x}}) + \lambda_{feat} \cdot \ell_{feat}(\boldsymbol{x}, \hat{\boldsymbol{x}})$。

3.2 Audio Language Modeling

回想一下，在这项工作中我们的目标是根据文本生成音频。具体来说，给定一个文本输入$\boldsymbol{c}$，Audio Language Model (ALM)组件输出一系列音频令牌$\hat{\boldsymbol{z}}_q$，可以使用G将其解码成原始音频。

考虑一个文本编码器F，它将原始文本输入映射为语义密集表示，$F(\boldsymbol{c}) = \boldsymbol{u}$。然后，一个Look-Up-Table (LUT)将音频令牌$\hat{\boldsymbol{z}}_q$嵌入到连续空间中，$\text{LUT}(\hat{\boldsymbol{z}}_q) = \boldsymbol{v}$。然后，我们将$\boldsymbol{u}$和$\boldsymbol{v}$连接起来，创建$Z = \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{T_u}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{T_v}$，其中$T_u$和$T_v$分别是文本表示和音频表示的长度。

使用以上表示，我们使用交叉熵损失函数训练一个由θ参数化的Transformer解码器语言模型：

$$L_{\text{LM}} = -\sum_{i=1}^{N}\sum_{j=1}^{T_v} \log p_\theta(\boldsymbol{v}_j^i | \boldsymbol{u}_1^1, \ldots, \boldsymbol{u}_{T_u}^i, \boldsymbol{v}_1^1, \ldots, \boldsymbol{v}_{j-1}^i). \tag{3}$$

文本表示是通过预训练的T5文本编码器获得的 (Raffel et al., 2020)。我们还尝试使用LUT学习文本嵌入。虽然它产生的结果与T5模型相当，但它限制了我们在训练过程中对未见过的单词进行泛化的能力，因此我们没有继续追求这个方向。Transformer解码语言模型采用GPT2类似的架构实现 (Radford et al., 2019)。为了实现更好的文本粘合度，我们在transformer的每个attention块之间增加了音频和文本之间的交叉注意力。在图 1（左）中可以看到一个视觉描述。整个系统也可以作为编码器-解码器模型来看，其中编码器（T5）在训练过程中是预训练的并且固定的。

分类器自由引导. Ho & Salimans (2021); Nichol et al. (2022)最近表明使用Classifier Free Guidance (CFG)方法是控制样本质量和多样性之间的权衡的有效机制。尽管CFG方法最初是针对扩散模型的分数函数估计提出的，但在这项工作中我们将其应用于自回归模型。在训练过程中，我们有条件地和无条件地优化Transformer-LM。实际上，我们在10%的训练样本中随机省略文本条件。在推理时，我们从通过条件和无条件概率线性组合获得的分布中进行采

样。形式上，我们从中采样，

$$\gamma \log p_\theta(\boldsymbol{v}_j^i|\boldsymbol{u}_1^1,\ldots,\boldsymbol{u}_{T_u}^i,\boldsymbol{v}_1^1,\ldots,\boldsymbol{v}_{j-1}^i) + (1-\gamma)\log p_\theta(\boldsymbol{v}_j^i|\boldsymbol{v}_1^1,\ldots,\boldsymbol{v}_{j-1}^i), \tag{4}$$

其中，γ是导航尺度。

多流音频输入。 为了生成高质量的音频样本，我们将原始音频进行32倍下采样，每个音频令牌对应2毫秒。这要求我们处理非常长的序列，因为每秒钟的音频由500个令牌表示。对于建模如此长的序列而言，这是一个众所周知的困难问题Rae et al. (2020); Zaheer et al. (2020); Beltagy et al. (2020)。为了缓解这个问题，我们提出了一种多流表示和建模范式。 Kharitonov et al. (2022b)的研究表明，Transformer模型能够同时建模多个流。

考虑长度为T_v的序列，我们可以使用两个大致相同比特率的并行流来学习长度为$T_v/2$的表示。这种方法可以推广到k个流，其中每个流的长度为T_v/k，每个码本的大小为$2048/k$。这样的表示可以通过将Q从单个码本的矢量量化推广为一个残差矢量量化模块来获得，就像Zeghidour et al. (2021)所做的那样。在时间t，网络接受k个离散码，并使用k个嵌入层对其进行嵌入。时间t的最终嵌入是这k个嵌入的均值。我们使网络适应使用k个语言模型预测头输出k个码。这些预测头相互独立地进行操作，我们尝试了将流i条件化到流$i-1$上，但没有观察到任何性能提升。

4 Experiments

在本节中，我们首先提供了对实验设置的详细描述。接下来，我们呈现了音频生成的主要结果，最后我们通过消融研究结束本节。

4.1 Experimental Setup

数据集。 我们使用了几个数据集：AudioSet (Gemmeke et al., 2017)、BBC音效 [1]、AudioCaps (Kim et al., 2019)、Clotho v2 (Drossos et al., 2020)、VGG-Sound (Chen et al., 2020)、FSD50K (Fonseca et al., 2021)、Free To Use Sounds [2]、Sonniss Game Effects [3]、WeSoundEffects [4]、Paramount Motion - Odeon Cinematic Sound Effects [5]。所有音频文件采样率为16kHz。

对于文本描述，我们使用两种类型的注释。第一种是多标签注释，适用于数据集：AudioSet、VGG-Sound、FSD50K、Sinniss Game Effects、WeSoundEffects、Paramount Motion - Odeon Cinematic Sound Effects。我们通过连接每个音频样本可用的标签列表来形成伪句子（例如，“狗，叫声，公园”被转换为“狗叫声公园”）。第二种类型的注释是自然语言标题，适用于数据集：AudioCaps、Clotho v2、Free To Use Sounds和BBC音效。有关使用的数据集的更详细描述可参见附录 3。我们采用了预处理步骤来更好地匹配类标注分布。具体而言，我们删除停用词和数字，最后使用NLTK（(Bird et al., 2009)中的WordNet词形还原器）对剩余单词进行词形还原（例如，“一只狗正在公园里叫”被转换为“狗叫声公园”）。由于语音是数据中占主导地位的类别，我们过滤掉所有包含单词“speech”的标签或标题的样本，以生成更平衡的数据集。总体而言，我们留下了约4k小时的训练数据。

数据增强。 最近提出的生成模型（(Ramesh et al., 2022; Saharia et al.; Gafni et al., 2022)）最令人印象深刻的能力之一是它们能够创建看不见的对象组合（例如，“太空中骑马的宇航

[1] https://sound-effects.bbcrewind.co.uk/
[2] https://www.freetousesounds.com/all-in-one-bundle/
[3] https://sonniss.com/gameaudiogdc
[4] https://wesoundeffects.com/we-sound-effects-bundle-2020/
[5] https://www.paramountmotion.com/odeon-sound-effects

表 1: 本文报告了DiffSound和几个AUDIOGEN的不同版本的结果。对于DiffSound数据增强，我们采用了作者建议的基于掩码的文本生成（MBTG）策略。对于主观测试，我们报告了总体质量（OVL）和文本相关性（REL.），并提供了95%的置信区间。对于客观指标，我们报告了FAD和KL。

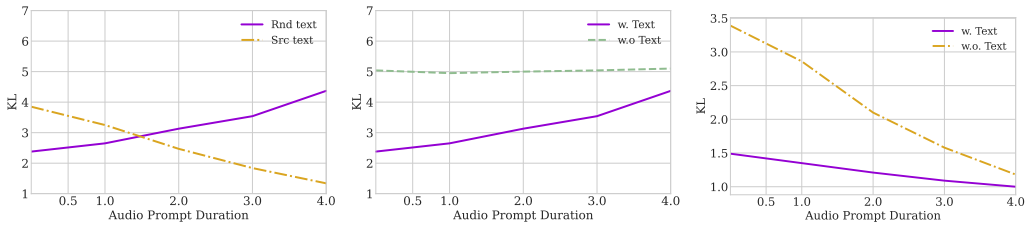
	#params	AUG.	TEXT-COND.	SUBJECTIVE OVL↑	SUBJECTIVE REL.↑	OBJECTIVE FAD↓	OBJECTIVE KL↓
Reference	-	-	-	92.08±1.16	92.97±0.85	-	-
DiffSound	400M	MBTG	CLIP	65.68±1.58	55.91±1.75	7.39	2.57
AUDIOGEN-base	285M	-	T5-base	70.85±1.06	63.23±1.65	2.84	2.14
AUDIOGEN-base	285M	Mix	T5-base	**71.68±1.89**	66.01±1.79	3.13	2.09
AUDIOGEN-large	1B	Mix	T5-large	**71.85±1.07**	**68.73±1.61**	**1.82**	**1.69**

员”）。为了实现类似的音频生成能力，我们提出了一种增强方法，在训练过程中通过融合一对音频样本及其对应的文本标题来创建新的概念组合。具体而言，给定两个音频样本 $\boldsymbol{x}_1, \boldsymbol{x}_2$ 和它们对应的文本标题 $\boldsymbol{c}_1, \boldsymbol{c}_2$，我们首先随机选择一个时间偏移来合并两个音频样本。接下来，在区间$[-5, 5]$中随机选择一个信噪比（Signal-to-Noise Ratio (SNR)），最后我们混合音频样本并连接文本标题 $\boldsymbol{c}_1, \boldsymbol{c}_2$。**评估方法.** 我们使用客观和主观指标来评估所有模型和基准模型。对于客观指标，我们计算真实样本和生成样本上的Fréchet Audio Distance (FAD) (Kilgour et al., 2019)。FAD 是从音频领域的 Fréchet Inception Distance (FID) 改编而来的一种无参考评估指标，与人类感知密切相关。类似于 Yang et al. (2022)，我们还计算了先进的音频分类模型 (Koutini et al., 2021) 在原始样本和生成音频上的 KL-散度。 FAD 已经被证明在音频质量方面与人类感知相关。另一方面，KL-散度是使用预训练分类器生成的标签分布计算得出的。因此，它更多地反映了录音中出现的更广泛的音频概念。因此，这两个指标是互补的。

对于主观评估方法，我们遵循与 (Yang et al., 2022) 类似的设置。我们要求人工评估员评估音频信号的两个主要方面：(i) 整体质量 (OVL) 和 (ii) 与文本输入的相关性。我们使用 MUSHRA 协议 (Series, 2014)，同时采用隐藏参考和低锚点。对于整体质量测试，评估员需要在1到100的范围内评价提供的样本的感知质量。对于文本相关性测试，评估员需要在1到100的范围内评价音频与文本之间的匹配程度。我们使用亚马逊 Mechanical Turk 平台招募评估员。我们评估 AudioCaps 测试集中随机采样的 100 个文件，每个样本至少由 5 个评估员评估。我们验证了大多数样本（85%）至少包含两个音频概念（例如，“一只狗在一只鸟鸣叫的时候叫”）。我们使用 CrowdMOS 软件包[6] 来过滤嘈杂的注释和离群值。我们删除了没有听完录音的注解人员，对参考录音评分低于 85 的注解人员，以及 CrowdMOS 的其余推荐策略 (Ribeiro et al., 2011)。本研究中的参与者至少获得了美国的最低工资标准。

超参数. 我们训练了两组ALM，一个有 2.85 亿参数 (base)，另一个有 10 亿参数 (large)。在较小的模型中，我们使用 768 的隐藏大小，24 层和 16 个注意力头，而对于较大的变体，我们使用 1280 的隐藏大小，36 层和 20 个注意力头。我们使用 Adam 优化器，批大小为 256，学习率为 5e-4，先进行 3k 步的预热，然后采用逆平方根衰减。小型模型在 64 个 A100 GPU 上训练了 200k 步（约 5 天），大型模型在 128 个 A100 GPU 上训练了 200k 步（约 1 周）。对于小型模型，我们使用 T5-base，对于大型模型，我们使用 T5-large。对于采样，我们使用 top-p (Holtzman et al., 2019) 采样，$p = 0.25$。对于CFG，我们使用 $\gamma = 3.0$。

[6] http://www.crowdmos.org/download/



(a) (b) (c)

图 2: 音频延续结果。在(a)中，我们将模型生成的音频与对应于SRC 和RND 的音频进行比较。在(b)中，我们将模型生成的音频与不带文本条件的音频进行比较，该音频对应于RND文本。在(c)中，我们将模型生成的音频与不带文本条件的音频进行比较，该音频对应于SRC文本。在所有设置中，我们报告KL结果。具体细节请参见第 4.2节。

4.2 RESULTS

我们首先将AUDIOGEN与DiffSound进行比较。我们在AudioCaps测试集上报告了客观指标。我们使用DiffSound作者提供的官方预训练模型 [7]。该模型在AudioSet上进行了训练，并在AudioCaps上进行了微调。结果在表 1中呈现。

AUDIOGEN-base在考虑所有指标的情况下优于DiffSound，而在参数数量方面则较小。与预期相反，AUDIOGEN-large在性能上远远优于DiffSound和AUDIOGEN-base。注意，对于AUDIOGEN-base模型，无混合训练的模型在FAD和可比的OVL方面表现更好。然而，当考虑与文本的相关性，即KL和REL.时，混合训练模型达到了更好的性能。这在REL.指标中尤为明显，因为它主要包含复杂的组合（见表 1）。我们还将AUDIOGEN-base与DiffSound使用相同的数据集设置进行了比较（即在AudioSet和AudioCaps上进行训练）。与DiffSound相比，AUDIOGEN-base的KL分数为2.46，DiffSound为2.57；FAD分数为4.39，DiffSound为7.39。这些结果表明AUDIOGEN仍然明显优于DiffSound。

接下来，我们尝试只将ALM组件预训练为只音频标记的模型（学习音频先验）。我们没有观察到使用ALM预训练带来的增益。我们假设这是由于预训练过程是在相同的标注数据上进行的。样本可以在以下链接找到：https://felixkreuk.github.io/audiogen。

音频延续。与 (Lakhotia et al., 2021)类似，我们首先将音频提示编码为音频标记序列，输入到ALM中，并采样生成音频延续。与先前的工作不同的是，我们还可以将生成的音频引导到文本标题。我们根据音频提示长度（[0.5秒，1秒，2秒，3秒，4秒]）和不同文本提示组合（i）无文本；（ii）与音频提示对应的文本（SRC）；（iii）来自随机抽样音频文件的文本（RND））来评估我们的模型。对于上述每个设置，我们测量生成的音频与源文本或目标文本之间的KL距离。

在图 2(a)中，我们将音频提示与RND文字一起输入模型。我们评估使用生成的音频的分类模型的输出与音频提示或与RND对应的音频之间的KL距离。结果表明，通过使用短音频提示，我们可以更好地将生成的音频引导到文本上。相反，使用较长的音频提示留下了较少的空间用于文本引导。文本和音频提示在约1.5秒时产生的影响大致相同。在图 2(b)中，我们将音频提示与有无RND文本作为条件一起输入模型，评估使用生成的音频的音频分类模型的输出与与RND对应的音频之间的KL距离。尽管使用较长的音频提示留下了较少的空间用于文本引导，但即使使用较长的音频提示（约4秒），我们仍观察到带有文本和不带文本的生

[7]预训练模型可在https://github.com/yangdongchao/Text-to-sound-Synthesis找到

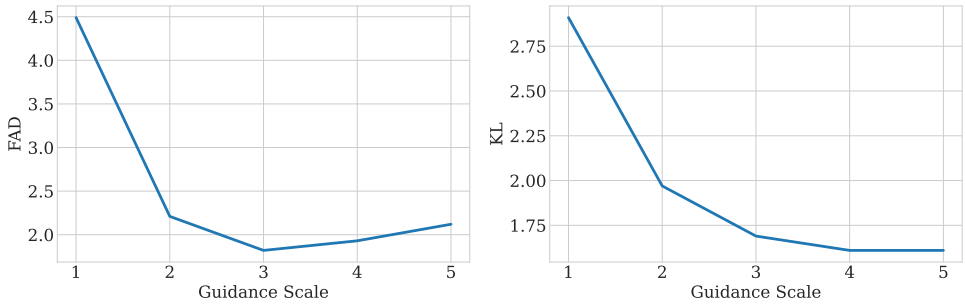


图 3: 根据引导尺度的变化，FAD（左）和KL（右）的结果。

表 2: 多流结果：我们报告了两组编码和生成的结果。对于编码度量，我们报道了比特率（kbps）、SI-SNR（dB）和ViSQOL。对于生成得分，我们报告了FAD、KL和推理加速。另外，我们还包括了流的数量和下采样因子（DSF）。请注意，由于使用了相同的音频表示模型，大模型和基础模型的编码度量是相同的。

			Encoding			Generation		
	# streams	DSF	Bitrate (kbps)	SI-SNR↑	ViSQOL↑	FAD↓	KL↓	Speed-Up
base	1	x32	5.37	5.1	3.95	3.13	2.09	x1.0
	2	x64	4.88	4.5	3.94	10.35	2.17	x2.0
	4	x128	4.39	4.2	3.91	9.68	2.36	x5.1
large	1	x32	5.37	5.1	3.95	1.82	1.69	x1.0
	2	x64	4.88	4.5	3.94	6.89	1.86	x2.3
	4	x128	4.39	4.2	3.91	10.89	2.59	x3.6

成之间的差距。这表明文本对生成过程具有显着影响。最后，在图 2(c)中，我们将模型条件设置为音频提示有无Src文本。我们评估使用生成的音频的音频分类模型的输出与输入音频之间的KL距离。结果表明，使用短音频提示时，文本对生成的输出有重大影响。然而，随着我们输入更长的序列，音频在将生成的输出引导到目标概念类上已足够，而添加文本只带来了最小的增益。

完整结果以及所有设置的FAD分数和可视化结果可在附录中的表 4中找到（见附录 7.2）。

4.3 Ablation Study.

分类器导向尺度的影响。正如 Ho & Salimans (2021)所指出的，CFG尺度的选择在样本多样性和与条件文本相关的质量之间存在着权衡。为了评估我们设置中参数γ对图 3的影响，我们报告了$\gamma \in \{1.0, 2.0, 3.0\}$的结果。需要注意的是，将$\gamma = 1$设置为单纯的采样过程（无CFG）。

去除CFG会导致性能较差，相比于$\gamma > 1.0$。FAD分数在$\gamma = 3.0$时达到最小值，而KL呈单调递减，且在$\gamma = 4.0$时收敛。这意味着在评估设置中，使用$\gamma = 3.0$提供了最好的质量和多样性之间的权衡。

多流处理。为了更好地理解使用多个流而不是单个流的好处，我们使用Down Sampling Factors (DSF)为 {x32, x64, x128}，分别使用{1, 2, 4}个不同的码本来优化三种不同的音频离散

表示模型。为了进行公平比较，我们将所有模型都保持在2048个总码字（例如，对于x64模型而言，使用大小为1024的2个码本，对于x128模型而言，使用大小为512的4个码本）。

为了评估音频表示的质量，我们报告了三个编码度量标准，即Scale-Invariant Signal-to-Noise Ration (SI-SNR)，ViSQOL (Chinen et al., 2020)和比特率。这些度量只对音频表示进行表征，而忽略了ALM。值得注意的是，SI-SNR和ViSQOL都是基于参考的度量，因此它们是在从学习到的表示重构的音频与原始音频序列之间进行计算的。结果报告在表 2上。虽然单流编码器取得了更好的SI-SNR结果，但在ViSQOL所测得的感知质量方面，所有表示都是可比较的。虽然所有设置中的码字总量都相同，但多流模型的有效比特率降低，导致较低的SI-SNR值。

接下来，我们报告了ALM在学习到的表示之上的FAD、KL和推理加速度。结果表明，与单流相比，增加流的数量会降低基础模型和大型模型的性能。 AUDIOGEN-base在KL分数上改进了DiffSound (2.57)， 同时产生了更高的FAD。 AUDIOGEN-large使用两个流，在FAD和KL方面都比DiffSound有所改进（分别为7.39和2.57）。虽然与单流模型相比，多流设置显示出较差的FAD和KL结果，但提供了推理质量和速度之间的权衡。

5 LIMITATIONS

由于我们使用相对较小的下采样因子来处理音频标记，音频标记序列可能非常长。这带来了两个主要限制：（i）对长范围序列的建模；（ii）高推理时间。在这项工作中，我们提出了对第一个限制的一种可能的缓解方法，然而这种方法会以产生质量较低的音频样本为代价。当考虑高分辨率音频样本时（例如48kHz的采样率），这些问题会变得更加严重。提出方法的另一个限制与音频作曲有关。尽管混音增强大大提高了模型分离源并创建复杂作曲的能力，但它仍然不具备对场景中时间顺序的理解，例如，狗在“然后”鸟在哼唱，在“背景中”狗在叫鸟在哼唱。最后，由于我们在训练集中省略了大部分语音样本，所提出的方法经常产生难以理解的语音。可以通过使用更多的语音数据、更好的语音数据增强配方或提供额外的语音特征来缓解这个问题。使用的数据集的另一个限制是其多样性。这些数据集主要来自YouTube，其中特定的人口统计和地理位置比其他方面更具代表性。这可能会在生成的样本中产生偏见。

6 CONCLUSION

在这项工作中，我们提出了一种基于Transformer的生成模型，命名为AUDIOGEN，它在学习到的离散音频表示上运行。与以前的工作不同，我们从实证上证明了自回归模型可以有条件地或无条件地生成高质量的音频样本。我们展示了即时文本和音频混合增强可以改善模型性能，并提供了一个去除试验，分析了Classifier Free Guidance和多流处理的效果。

至于更广泛的影响，这项工作为构建更好的文本到音频模型奠定了基础。此外，所提出的研究可能开拓了未来与基准测试、语义音频编辑、从离散单元中分离音频源等相关方向的研究。

参考文献

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language

processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738, 2022.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alexei Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *ICLR*, 2020.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jack Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. Visqol v3: An open source production ready objective speech and audio metric. In *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Janek Ebbers et al. Hidden markov model variational autoencoder for acoustic unit discovery. In *INTERSPEECH 2017*, 2017.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Otis Lamont Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pp. 89–106. Springer, 2022.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Alexey Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, and Nal Kalchbrenner. A spectral energy distance for parallel speech synthesis. *Advances in Neural Information Processing Systems*, 33:13062–13072, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Po-Yao Huang, Hu Xu, Juncheng B Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=MAMOi89bOL.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Eugene Kharitonov, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Paden Tomasello, Ann Lee, Ali Elkahky, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, et al. textless-lib: a library for textless spoken language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pp. 1–9, 2022a.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, et al. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8666–8681, 2022b.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, 2019.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

Felix Kreuk, Joseph Keshet, and Yossi Adi. Self-supervised contrastive learning for unsupervised phoneme segmentation. In *INTERSPEECH 2020*, 2020.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgan Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. Textless speech emotion conversion using discrete & decomposed representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11200–11214, 2022.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3327–3339, 2022a.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 860–872, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.63. URL https://aclanthology.org/2022.naacl-main.63.

Yunpeng Li, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek. Real-time speech frequency bandwidth extension. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 691–695. IEEE, 2021.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *arXiv preprint arXiv:2203.16502*, 2022.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2337–2346, 2019.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. In *INTERSPEECH*, 2021.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation. In *Proc. Interspeech 2022*, pp. 5195–5199, 2022. doi: 10.21437/Interspeech.2022-11032.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Mr D Murahari Reddy, Mr Sk Masthan Basha, Mr M Chinnaiahgari Hari, and Mr N Penchalaiah. Dall-e: Creating images from text. *UGC Care Group I Journal*, 8(14):71–75, 2021.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2021.

F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2416–2419, 2011. doi: 10.1109/ICASSP.2011.5946971.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *INTERSPEECH*, 2019.

B Series. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2014.

Yôiti Suzuki, Volker Mellert, Utz Richter, Henrik Møller, Leif Nielsen, Rhona Hellman, Kaoru Ashihara, Kenji Ozawa, and Hisashi Takeshima. Precise and full-range determination of two-dimensional equal loudness contours. *Tohoku University, Japan*, 2003.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a. URL https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017b.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

A Vasuki and PT Vanathi. A review of vector quantization techniques. *IEEE Potentials*, 25(4): 39–47, 2006.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*, pp. 10937–10947. PMLR, 2021.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203. IEEE, 2020.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022.

Jaeseong You, Dalhyun Kim, Gyuhyeon Nam, Geumbyeol Hwang, and Gyeongsu Chae. Gan vocoder: Multi-resolution discriminator is all you need. *arXiv preprint arXiv:2103.05236*, 2021.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

7 APPENDIX

7.1 DATASETS

我们使用了几个数据集：AudioSet (Gemmeke et al., 2017)，BBC音效 [8]，AudioCaps (Kim et al., 2019)，Clotho v2 (Drossos et al., 2020)，VGG-Sound (Chen et al., 2020)，FSD50K (Fonseca et al., 2021)，Free To Use Sounds [9]，Sonniss Game Effects [10]，WeSoundEffects [11]，Paramount Motion - Odeon Cinematic Sound Effects [12]。所有音频文件均以16kHz采样。

表 3: 数据集描述。持续时间以原始音频的小时为单位报告，即在预处理之前。

DATASET	TEXT CONDITIONING	DURATION (H)
AudioSet	tags	5.42k
BBC	captions	463
AudioCaps	captions	145
Clotho v2	captions	37
VGG-Sound	tags	560
FSD50K	tags & captions	108
Free To Use Sounds	tags & captions	176
Sonniss Game Effects	tags	85
WeSoundEffects	tags	12
Paramount Motion	tags	20

7.2 ADDITIONAL RESULTS

我们在图 4 中展示了条件和非条件音频延续的视觉定性描述。我们使用以下文本作为条件，即“演讲和山羊的叫声”，并使用其对应音频的前一秒作为提示。我们用虚线标记生成段的开始位置。在图 4 的左侧，我们可视化了无条件音频延续（即无文本）。在图 4 的右侧，我们可视化了基于文本条件的音频延续。由于音频提示只包含人类说话声，无条件模型生成的延续包含了说话的话语，但没有产生任何山羊的声音。相反，基于文本条件的模型成功地生成了人类说话声和山羊的声音（左侧）。

我们在表 4中报告了图 2中呈现的完整结果。

最后，我们分析了在考虑文本编码器和自动语音识别模型（ALM）时，模型大小对生成的音频的影响。在表格 5中，我们报告了四种不同组合的KL和FAD分数：{T5-base, T5-large}×{ALM-base, ALM-large}。当使用较大的T5编码器时，我们观察到在KL方面有很大的改进，并在FAD方面有较小的改进。另一方面，当使用较大的ALM时，我们观察到KL方面有类似的改进，但在FAD方面有显著的改进。使用T5-large和ALM-large的组合总体上得到最佳结果。

[8] https://sound-effects.bbcrewind.co.uk/
[9] https://www.freetousesounds.com/all-in-one-bundle/
[10] https://sonniss.com/gameaudiogdc
[11] https://wesoundeffects.com/we-sound-effects-bundle-2020/
[12] https://www.paramountmotion.com/odeon-sound-effects

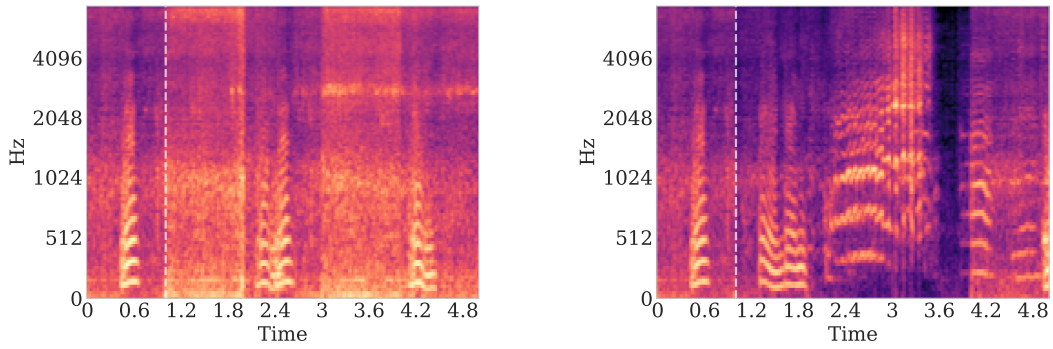


图 4: 一个文本引导的音频延续的视觉示例。我们绘制了没有文本限制的音频延续的Mel-频谱图（左）和文本引导的音频延续的Mel-频谱图（右）。输入文本是：“说话和山羊的叫声”。

TEXT	AUDIO PROMPT DURATION (SEC.)	FAD	KL W. RND	KL W. SRC
SRC	0.5	1.97	-	1.49
no text	0.5	5.66	5.04	3.39
RND	0.5	3.08	2.38	3.85
SRC	1	1.96	-	1.35
no text	1	4.70	4.95	2.86
RND	1	3.17	2.65	3.25
SRC	2	1.92	-	1.21
no text	2	3.15	5.00	2.10
RND	2	2.92	3.13	2.47
SRC	3	1.90	-	1.09
no text	3	2.46	5.04	1.58
RND	3	2.58	3.54	1.84
SRC	4	1.97	-	1.00
no text	4	2.14	5.10	1.18
RND	4	2.29	4.37	1.34

表 4: 对于所有文本条件和音频提示设置，FAD和KL指标的完整结果。

T5	ALM	KL	FAD
Base	Base	2.09	3.13
Base	Large	1.92	2.27
Large	Base	1.91	3.03
Large	Large	1.69	1.82

表 5: 消融研究。我们报告使用四种不同文本编码器和ALM设置的KL和FAD分数，考虑基础和大型模型。