MusicLM：从文本中生成音乐

Andrea Agostinelli [*1] **Timo I. Denk** [*1]
Zala´n Borsos [1] **Jesse Engel** [1] **Mauro Verzetti** [1] **Antoine Caillon** [2] **Qingqing Huang** [1] **Aren Jansen** [1]
Adam Roberts [1] **Marco Tagliasacchi** [1] **Matt Sharifi** [1] **Neil Zeghidour** [1] **Christian Frank** [1]

摘要

我们介绍了MusicLM，这是一个从文本描述中生成高保真音乐的模型，如*"平静的小提琴旋律伴着扭曲的吉他声"*。MusicLM将条件音乐的生成过程设定为一个层次化的序列到序列的建模任务，它以24kHz的频率生成音乐，并在几个minutes中保持一致。我们的实验表明，MusicLM在音频质量和对文本描述的遵守方面都优于以前的系统。此外，我们证明MusicLM可以以文本和旋律为条件，因为它可以根据文本标题中描述的风格来转换口哨和哼唱的旋律。为了支持这项研究，我们公开发布了MusicCaps，这是一个由5.5千首音乐-文本对组成的数据集，其中有人类专家提供的丰富文本描述。

google-research.github.io/seanet/musiclm/examples

1. 简介

条件性神经音频生成涵盖了广泛的应用，从文本到语音（Zen等人，2013年；van den Oord等人，2016年）到歌词条件音乐审议（Dhariwal等人，2020年）和MIDI序列的音频合成（Hawthorne等人，2022b）。调节信号和相应的音频输出之间有一定程度的时间一致性，这对此类任务是有利的。相比之下，受文本到图像生成的进展启发（Ramesh等人，2021年；2022年；Saharia等人，2022年；Yu等人，2022年），最近的工作探索了从整个序列的高级标题（Yang等人，2022年；Kreuk等人，2022年）生成音频，如*"吹着风的口哨"*。虽然从这种粗略的标题中生成音频是一种突破，但这些模型仍然局限于简单的声学场景，包括在一段时期内的少数声学事件。

秒的时间。因此，将单一的文字说明变成具有长期结构和许多干系的丰富的音频序列，如音乐片段，仍然是一个公开的挑战。

AudioLM（Borsos等人，2022年）最近被提议作为一个音频生成框架。将音频合成作为离散表示空间中的语言建模任务，并利用从粗到细的音频离散单元（或*标记*）的层次结构，AudioLM实现了几十秒的高保真和长期一致性。此外，通过对音频信号的内容不做任何假设，AudioLM学会了从纯音频语料中生成真实的音频，无论是语音还是钢琴音乐，都不需要任何注释。对不同信号进行建模的能力表明，如果在适当的数据上进行训练，这种系统可以产生更丰富的输出。

除了合成高质量和连贯的音频的固有困难，另一个阻碍因素是成对的音频-文本数据的稀缺。这与图像领域形成鲜明对比，在该领域，大量数据集的可用性对最近实现的显著的图像生成质量有很大贡献（Ramesh等人，2021；2022；Saharia等人，2022；Yu等人，2022）。此外，创建一般音频的文本描述要比描述图像难得多。首先，仅用几个词就能毫不含糊地捕捉到声学场景（例如，在火车站或森林中听到的声音）或音乐（例如，旋律、节奏、人声的音色和用于伴奏的许多乐器）的突出特征，这是不简单的。其次，音频是沿着时间维度结构的，这使得整个序列的标题比图像标题的注释水平要弱得多。

在这项工作中，我们介绍了MusicLM，一个从文本描述中生成高保真音乐的模型。MusicLM利用AudioLM的多阶段自回归建模作为生成部分，同时将其扩展到包含文本调节。为了解决配对数据稀缺的主要挑战，我们依靠MuLan（Huang等人，2022年），这是一个音乐-文本联合模型，它被训练成将音乐和其相应的文本描述投射到嵌入空间中彼此接近的表示。这种共享的嵌入空间消除了在训练时对字幕的需要。

[*]平等贡献[1] Google Research[2] IRCAM - Sorbonne Universite´（在Google实习时完成的工作）。通讯员：Christian Frank <>。Christian Frank <*chfrank@google.com*>。

团结一致，并允许在大规模的纯音频语料库上进行训练。也就是说，我们在训练过程中使用从音频中计算出来的MuLan嵌入作为条件，而在推理过程中使用从文本输入中计算出来的MuLan嵌入。

当在一个大型的无标签音乐数据集上进行训练时，MusicLM学会了以24kHz的频率生成长而连贯的音乐，用于显著复杂的文本描述，如*"令人陶醉的爵士乐，有令人难忘的萨克斯风独奏和独唱"*或*"柏林90年代的电子乐，有低沉的低音和强烈的踢腿"*。为了解决这一任务缺乏评估数据的问题，我们介绍了MusicCaps，这是一个新的高质量的音乐标题数据集，有5.5万个例子，由专家音乐家准备，我们公开发布以支持未来的研究。

我们的实验表明，通过定量指标和人类评价，MusicLM在质量和对字幕的遵守方面都优于之前的系统，如Mubert（Mubert-Inc, 2022）和Riffu- sion（Forsgren & Martiros, 2022）。此外，由于用文字描述音乐的某些方面可能是困难的，甚至是不可能的，我们展示了我们的方法如何支持文本以外的信号条件。具体来说，我们对MusicLM进行了扩展，以接受音频形式的额外旋律（例如口哨声、哼唱声）作为条件，生成一个遵循所需旋律的音乐片段，以文本提示描述的风格呈现。

我们承认与音乐生成有关的风险，特别是对创造性构思的潜在盗用。根据负责任的模型开发实践，我们通过改编和扩展Carlini等人（2022）用于基于文本的大型语言模型的方法，对记忆进行了彻底研究。我们的研究结果表明，当把MuLan嵌入送入MusicLM时，生成的标记序列与训练集中的相应序列明显不同。

这项工作的主要贡献如下。

1. 我们介绍MusicLM，这是一个生成模型，它能以24kHz的频率产生高质量的音乐，在几分钟内保持一致，同时忠实于文本描述信号。

2. 我们将我们的方法扩展到其他条件信号，如根据文本提示合成的旋律。此外，我们还展示了长达5分钟的长篇连贯的音乐生成。

3. 我们发布了第一个专门为文本到音乐的生成任务收集的评估数据集。MusicCaps是一个经过手工整理的高质量的数据集，其中包括5.5万个由音乐家准备的音乐-文本对。

2. 背景和相关工作

各类生成模型的最先进技术主要由基于Transformer的自回归模型（Vaswani等人，2017）或基于U-Net的扩散模型（Ho等人，2020）所主导。在本节中，我们重新审视相关的工作，重点是在离散标记上操作的自回归生成模型，它与MusicLM有相似之处。

2.1. 量化

在自然语言处理（Brown等人，2020年；Cohen等人，2022年）和图像或视频生成（Esser等人，2021年；Ramesh等人，2021年；Yu等人，2022年；Villegas等人，2022年）中，对离散标记的序列进行自回归建模已被证明是一种强有力的方法。量化是连续信号（包括图像、视频和音频）的自回归模型成功的一个关键组成部分。量化的目的是提供一个紧凑的、离散的表示，同时允许高保真的重建。VQ-VAEs（Van Den Oord等人，2017）在不同领域的低比特率下展示了令人印象深刻的重建质量，并作为许多方法的基础量化器。

SoundStream（Zeghidour等人，2022年）是一个通用的神经音频编解码器，能够以低比特率压缩一般的音频，同时保持高重建质量。为了实现这一点，SoundStream使用残余矢量量化（RVQ），允许扩展到更高的比特率和质量，而没有显著的计算成本。更具体地说，RVQ是一个层次化的量化方案，由一系列的矢量量化器组成，目标信号被重建为量化器输出的总和。由于量化器的组合，RVQ避免了随着目标比特率的增加而导致编码本大小的指数级膨胀。此外，每个量化器与较粗的量化器的残差相适应的事实为量化器引入了一个层次结构，其中较粗的层次对高保真度的重建更为重要。这种特性对于生成来说是可取的，因为过去的语境可以通过只关注粗大的标记来定义。最近，SoundStream被EnCodec（De´fossez等人，2022）扩展到更高比特率和立体声音频。在这项工作中，我们依靠SoundStream作为我们的音频标记器，因为它能以6kbps的高保真度重建24kHz的mu- sic。

2.2. 音频的生成模型

尽管生成具有长期一致性的高质量音频是一个挑战，但最近有一系列的方法解决了这个问题，并取得了一些成功。例如，Jukebox（Dhariwal等人，2020年）提出了不同时间分辨率的VQ-VAE的层次结构，以实现高时间性的

虽说是连贯性，但生成的音乐显示出明显的艺术性。另一方面，PerceiverAR（Hawthorne等人，2022a）提议对SoundStream标记的序列进行自回归建模，实现高质量的音频，但压缩了长期的时间一致性。

受到这些方法的启发，AudioLM（Borsos等人，2022年）通过依靠分层标记和命名方案，解决了一致性和高质量合成之间的权衡问题。具体来说，该方法区分了两种标记类型：（1）*语义*标记，允许对长期结构进行建模，这些标记是从对音频数据进行预训练的模型中提取的，目的是对语言进行建模；（2）*声学*标记，由神经音频编解码器提供，用于捕捉精细的声学细节。这使得AudioLM能够生成连贯和高质量的语音以及钢琴音乐的连续，而不需要依赖转写或符号化的音乐表示。

MusicLM建立在AudioLM的基础上，有三个重要的额外贡献：（1）我们把生成过程的条件放在描述性文本上，（2）我们表明条件可以扩展到其他信号，如旋律，和
(3)
我们对钢琴音乐以外的大量长篇音乐序列进行建模（从*爵士乐*的*鼓点*到*古典音乐*）。

2.3. 有条件的音频生成

从文本描述（如*"背景是笑声的口哨声"*）中生成音频，最近已被几项工作所解决。DiffSound（Yang等人，2022年）使用CLIP（Radford等人，2021年）作为文本编码器，并应用扩散模型来预测基于文本嵌入的目标音频的量化旋律谱特征。AudioGen（Kreuk等人，2022）使用T5（Raffel等人，2020）编码器来嵌入文本，并使用自回归变压器解码器来预测由EnCodec（De´fossez等人，2022）制作的目标音频代码。这两种方法都依赖于适量的配对训练数据，如AudioSet（Gemmeke等人，2017）和AudioCaps（Kim等人，2019）（过滤后总计不到5k小时）。

与MusicLM更接近的是，也有一些作品专注于以文本为条件的音乐生成。在Mubert（Mubert-Inc, 2022）中，文本提示被一个转化器嵌入，与编码提示相近的音乐标签被选中并用于查询歌曲生成API。根据选定的标签，Mubert生成声音的组合，而这些声音又是由音乐家和声音设计师生成的。这与Riffusion（Forsgren & Martiros, 2022）形成对比，后者在配对的音乐-文本数据集中的音乐作品的熔体频谱图上对稳定扩散模型（Rombach等人，2022a）进行微调。我们将Mubert和Riffusion作为我们工作的基线，表明我们提高了音频生成质量和对文本描述的坚持。

正如Huang等人（2019）；Hawthorne等人（2019）；Engel等人（2020）所证明的那样，音乐的符号表示（如MIDI）也可以用来驱动生成过程，作为一种强调节的形式。MusicLM能够以更自然和直观的方式提供调理信号，例如通过哼唱的旋律，也可以与文本描述相结合。

2.4. 以文本为条件的图像生成

文字条件下的音频合成的前身是文字条件下的图像生成模型，由于架构的改进和大量高质量的配对训练数据的可用性，这些模型在质量上取得了显著的进步。著名的基于变压器的自回归方法包括Ramesh等人（2021）；Yu等人（2022），而Nichol等人（2022）；Rombach等人（2022b）；Saharia等人（2022）提出基于扩散的模型。文本到图像的方法已经被扩展到从文本提示生成视频（Wu等人，2022a；Hong等人，2022；Vil- legas等人，2022；Ho等人，2022）。

在这些作品中，与我们的方法最接近的是*DALL*-E 2（Ramesh等人，2022）。特别是，与DALL-E 2依赖CLIP的方式类似（Radford等人，2021年）。在文本编码方面，我们也使用了一个音乐-文本联合嵌入模型来达到同样的目的。与使用扩散模型作为解码器的DALL-E 2相比，我们的解码器是基于AudioLM的解码器。此外，我们还省略了将文本嵌入映射到音乐嵌入的先验模型，这样基于AudioLM的解码器就可以在仅有音频的数据集上进行训练，在推理过程中，音乐嵌入被简单地替换成文本嵌入。

2.5. 音乐和文本的联合嵌入模型

MuLan（Huang等人，2022）是一个音乐-文本联合嵌入模型，由两个嵌入塔组成，每个模态一个。这些塔使用对比学习将两种模式映射到128维的共享嵌入空间，其设置类似于（Radford等人，2021；Wu等人，2022b）。文本嵌入网络是一个BERT（Devlin等人，2019年），在一个大型纯文本数据的语料库上进行了预训练，而我们使用的是音频塔的ResNet-50变体。

MuLan是在成对的音乐片段和它们相关的文本注释上进行训练的。重要的是，MuLan对其训练数据质量的要求很低，即使在音乐-文本对只有微弱关联的情况下也能学习跨模式的对应关系。将Music与无约束的自然语言描述联系起来的能力使其适用于检索或零拍的音乐标签。在这项工作中，我们依靠Huang等人（2022）的预训练和冻结模型。

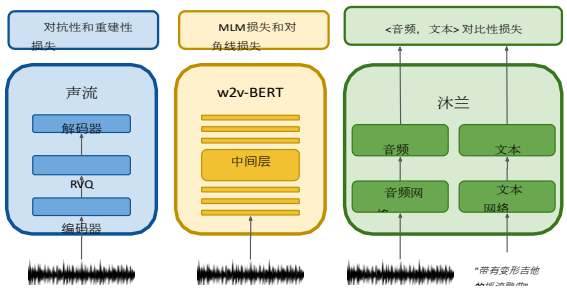


图 1. 为 MusicLM提供audio和文本表征的模型的独立预训练：SoundStream（Zeghidour等人，2022），w2v-BERT（Chung等人，2021），和MuLan（Huang等人，2022）。

3. 方法

在这一节中，我们将描述MusicLM和它的组件。第3.1节描述了提供音频表现的模型。然后，我们在第3.2节中展示了我们如何使用这些表征来生成文本条件的音乐。

3.1. 音频和文本的表示和标记化

我们使用三种模型来提取将用于条件自回归音乐生成的音频表示，如图1所示。特别是，通过遵循AudioLM的方法，我们使用SoundStream（Zeghidour等人，2022）的自我监督音频表征，作为声学标记以实现高保真合成，以及w2v-BERT（Chung等人，2021），作为语义标记以促进长期一致性生成。为了表示连贯性，我们在训练时依靠MuLan音乐嵌入，在推理时依靠MuLan文本嵌入。这三个模型都是独立预训练的，然后冻结，这样它们就为序列到序列的建模提供了离散的音频和文本表示。

SoundStream。我们使用一个SoundStream模型，用于24kHz的单声道音频，跨度系数为480，结果是50Hz的嵌入。这些嵌入的量化在训练期间由一个具有12个量化器的RVQ学习，每个量化器的词汇量为1024。这导致了6kbps的比特率，其中一秒钟的音频由600到Kens表示。我们把这些称为*声学标记*，用A表示。

w2v-BERT。与AudioLM类似，我们使用w2v-BERT模型的掩蔽语言建模（MLM）模式的中间层，有600M的参数。在对模型进行预训练和冻结后，我们从第7层提取嵌入，并使用学习过的k-means的中心点对嵌入进行量化。我们使用1024个节点和25赫兹的采样率，结果是每秒钟的音频有25个*语义标记*，用S表示。

MuLan。为了训练MusicLM，我们从MuLan的音频嵌入网络中提取目标音频序列的表示。注意，这个表示是连续的，可以直接用作基于变压器的自回归模型的调节信号。然而，我们选择将MuLan嵌入量化的方式，使音频和调节信号都有一个基于离散标记的同质表示，这有助于进一步研究调节信号的自回归模型。

由于MuLan在10秒的音频输入上操作，而我们需要处理较长的音频序列，所以我们在10秒的窗口上计算音频嵌入，跨度为1秒，并对所得嵌入进行平均。然后，我们通过应用12个矢量量化器的RVQ对所得嵌入进行分解，每个矢量量化器的词汇量为1024。这个过程产生了12个MuLan音频标记M_A，用于一个audio序列。在推理过程中，我们使用从文本提示中提取的MuLan文本嵌入作为条件，并用与音频嵌入相同的RVQ对其进行量化，以获得12个标记M_T。

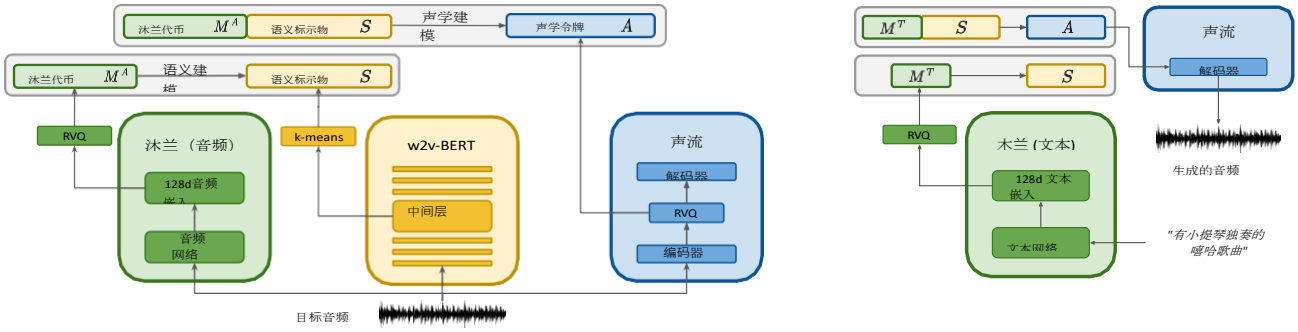
在训练期间对M_A，有两个主要的好处。首先，它允许我们轻松地扩展我们的训练数据，因为我们不受文字说明需求的限制。其次，通过利用像MuLan这样的模型，用对比性损失进行训练，我们提高了对噪声文本描述的鲁棒性。

3.2. 音频表现的分层建模

我们将上面提出的离散音频表示与AudioLM结合起来，以实现文本条件的音乐生成。为此，我们提出了一个分层的序列到序列的建模任务，其中每个阶段都由一个单独的仅有解码器的转化器进行自回归建模。拟议的方法在图2中得到说明。

第一阶段是*语义建模*阶段，它学习从MuLan音频标记到语义的映射。通过对分布$p(S_t \mid S_{<t}, M_A)$的建模，可以看出S*的*tic tokens。其中t是序列中对应于a的位置的时间步骤。第二阶段是*声学建模*阶段，声学标记A_q是根据MuLan的音频标记和语义来预测的。tokens，对分布$p(A_t \mid A_{<t}, S, M_A)$进行建模。

值得注意的是，为了避免长的标记序列，AudioLM建议将声学建模阶段进一步分成粗略和精细的建模阶段。我们依靠同样的方法，粗略阶段对SoundStream RVQ输出的前四级进行建模，精细阶段对后八级进行建模--我们参考Borsos等人（2022）的细节。



图

*2.*左图：在训练过程中，我们*从纯音频*训练集中提取MuLan音频标记、语义标记和声学标记。在语义建模阶段，我们用MuLan音频标记作为条件预测语义标记。在随后的声学建模阶段，鉴于MuLan音频标记和语义标记，我们预测声学标记。每个阶段都被建模为一个序列对序列的任务，使用仅有解码器的变形器。右边。在推理过程中，我们使用从文本提示中计算出的MuLan文本标记作为调节信号，并使用SoundStream解码器将生成的音频标记转换成波形。

4. 实验设置

4.1. 模型

我们使用纯解码器的Transformers来为AudioLM的半音阶段和声学阶段建模。这些模型具有相同的结构，由24层、16个注意力头、1024的嵌入维度、维度为4096的前馈层、0.1的剔除和相对位置嵌入组成（Raffel等人，2020），每个阶段有430M的参数。

4.2. 训练和推理

通过依赖预训练和冻结的MuLan，我们只需要音频数据来训练MusicLM的其他组件。我们在自由音乐档案（FMA）数据集上训练SoundStream和w2v-BERT（Defferrard等人，2017），而标记器和语义和声学建模阶段的自回归模型是在包含500万个音频片段的数据集上训练的，相当于28万小时24kHz的音乐。每个阶段的训练都是通过对训练数据的多次处理。我们在语义阶段和声学阶段分别使用30秒和10秒的目标音频的随机剪裁。AudioLM的精细声学建模阶段是在3秒的片段上进行训练。

在推理过程中，我们利用MuLan学到的音频和文本之间的联合嵌入空间，也就是说，我们用M_T 来代替M_A。然后，我们按照上述阶段，在给定的M_T ，得到A。我们在所有阶段都使用温度取样来进行自回归取样，语义建模阶段的温度为1.0，0.95和0.1。在粗略的和精细的声学建模阶段分别为0.4。这些温度值是根据次级检测选择的，以便在生成的音乐的多样性和时间一致性之间提供一个良好的权衡。

4.3. 评价数据集

为了评估MusicLM，我们准备了MusicCaps，这是一个高质量的音乐标题数据集，我们将其公开提供。[1]该数据集包括来自AudioSet（Gemmeke等人，2017年）的5.5千条音乐片段，每条音乐片段都有对应的英文文本描述，由10位专业音乐家撰写。对于每个10秒的音乐片段，MusicCaps提供：（1）一个平均由四句话组成的自由文本*标题*，描述音乐；（2）一个音乐*方面的*列表，描述流派、情绪、节奏、歌手声音、乐器、不协调性、节奏等。平均而言，每个片段的数据集包括11个方面。一些标题和内容列表的例子见附录A。

MusicCaps是对AudioCaps（Kim等人，2019）的补充，因为它们都包含来自AudioSet的音频片段和相关的文本描述。然而，AudioCaps包含非音乐内容，而MusicCaps只关注音乐，并包括高度详细的专家提供的注释。这些例子是从AudioSet的训练和评估部分中提取的，涵盖了各种类型的分布，详见附录A。MusicCaps还提供了一个*流派平衡的*分割数据，其中有1千个例子。

4.4. 度量衡

我们计算了不同的指标来评估MusicLM，抓住了音乐生成的两个重要方面：音频质量和对文本描述的遵守。

Fre´chet Audio Distance（FAD）。Fre´chet Audio Distance（Kilgour等人，2019年）是一个无参考文献的音频质量指标，与人类的感知很相关。其他产生FAD低分的样本的模式被认为是

[1]kaggle.com/datasets/googleai/musiccaps

来产生可信的音频。然而，生成的样本不一定符合作为条件提供的文本描述。

我们报告了基于两个音频嵌入模型的FAD，这两个模型都是公开可用的：（1）Trill[2](Shor et al., 2020)，它是在语音数据上训练的，和(2) VGGish[3]，（Hershey等人，2017），它是在YouTube-8M音频事件数据集（Abu-El-Haija等人，2016）上训练的。由于训练数据的不同，我们希望这些模型能够测量音频质量的不同方面（分别是语音和非语音）。

KL分歧（KLD）。文本描述和与之相符的音乐片段之间存在着多对多的关系。因此，不可能在音频波形的层面上直接比较生成的音乐和参考文献。为了评估对输入文本描述的遵守情况，我们采用了类似于Yang等人（2022）；Kreuk等人（2022）提出的代理方法。具体来说，我们使用一个LEAF（Zeghidour等人，2021）分类器，该分类器为AudioSet上的多标签分类而训练，为生成的音乐和参考音乐计算类别预测，并测量类别预测的概率分布之间的KL分歧。当KL-分歧较低时，根据分类器，预计生成的音乐将具有与参考音乐类似的声学特征。

MuLan循环一致性（MCC）。作为一个联合的音乐-文本嵌入模型，MuLan可以被用来量化音乐-文本对之间的相似性。我们从MusicCaps中的文本描述以及基于它们生成的音乐中计算出MuLan嵌入，并将MCC指标定义为这些嵌入之间的平均余弦相似度。

定性评价。最终，我们依靠主观测试来评估生成的样本与文本描述的一致性。我们设置了一个A-vs-B的人类评分任务，在这个任务中，评分者会看到文本描述和由两个不同模型生成的两个音乐样本，或者一个模型和参考音乐。有五种可能的评价者：对A或B有强烈或微弱的偏好，以及没有偏好。评审员被要求在做决定时不要考虑音乐质量，因为FAD指标已经涵盖了评价的这一环节。

除了参考音乐之外，我们还考虑了n个不同模型的输出，因此总共有$n+1$个条件和$n(n+1)/2$对。为了汇总成对测试的结果并对条件进行排序，我们计算"胜利"的数量。

[2]tfhub.dev/google/nonsemantic-speech-benchmark/trill/3[3]tfhub.dev/google/vggish/1

也就是说，一个条件被强烈或弱化的频率。这些样本是从我们的评估数据的流派平衡的1k子集中选出的。

训练数据的记忆。大型语言模型有能力记忆训练数据中的模式（Carlini等人，2020）。我们采用Carlini等人（2022）的方法来研究MusicLM对音乐片段的记忆程度。我们专注于第一阶段，负责语义建模。我们从训练集中随机选择N个例子。对于每个例子，我们向模型提供一个提示，其中包括MuLan音频标记M_A，然后是第一个T语义标记S的序列。与$T\in\{0,\ldots,250\}$，对应的时间最长为10秒。我们使用贪婪解码来生成125个se-的续集。语气词（5秒），我们将生成的语气词与数据集中的目标语气词进行比较。我们用在整个采样段上生成的标记和目标标记相同的例子的百分比来衡量完全匹配。

此外，我们提出了一种检测近似匹配的方法，其依据是看似不同的标记序列可能导致声学上相似的音频片段。也就是说，我们计算了相应词汇的语义标记计数的直方图。准则$\{0,\ldots,1023\}$，从生成的和目标的都是符号，并定义一个匹配的成本措施，在他的图如下。首先，我们计算语义标记对之间的距离矩阵，该矩阵由用于将w2v-BERT量化为语义标记的相应k-means中心点之间的欧几里得距离填充（见第3.1节）。然后，我们使用Sinkhorn算法（Cuturi，2013）解决一个最佳传输概率，以找到一对直方图之间的匹配成本，只考虑两个直方图中非零标记计数对应的子矩阵。为了校准用于确定两个序列是否可能是近似匹配的阈值，我们通过对带有目标标记的例子进行置换来构建负对，并测量这种负对的匹配成本的经验分布。我们将匹配阈值τ设定为0.85，这将导致低于0.01%的假阳性近似匹配。

5. 结果

我们通过将MusicLM与最近两个从描述性文本生成音乐的基线，即Mubert（Mubert-Inc, 2022）和Riffusion（Forsgren & Mar- tiros, 2022）进行比较来评估它。特别是，我们通过查询Mubert的API来生成音频。[4]并在Riffusion模型上运行推理。[5]我们在MusicCaps上进行评估，这是我们与本文一起公开发布的评估数据集。

[4]github.com/MubertAI（2022年12月和2023年1月访问）。
[5]github.com/riffusion/riffusion-app（2022年12月27日访问）。

*表1.*使用MusicCaps数据集的标题对生成的样本进行评估。通过Fre´chet Audio Distance（FAD）对模型的音频质量进行了比较，通过Kullback-Leibler Divergence（KLD）和MuLan Cycle Consistency（MCC）对模型对文本描述的忠实度进行了比较，以及在成对的人类听力测试中的获胜次数（Wins）。

模型	$\text{FAD}_{\text{TRILL}}$ ↓	FAD_{VGG} ↓	KLD ↓	MCC ↑	赢了 ↑
裂变	0.76	13.4	1.19	0.34	158
穆贝尔	0.45	9.6	1.58	0.32	97
音乐MusicLM	0.44	4.0	1.01	0.51	312
音乐帽	-	-	-	-	472

与基线的比较。表1报告了本文的主要定量和定性结果。就FAD指标所反映的au-dio质量而言，在FAD_{VGG}，MusicLM取得了比Mubert和Riffusion更好的分数。在$\text{FAD}_{\text{Trill}}$，MusicLM的得分与Mubert相似（0.44对0.45），比Riffusion更好（0.76）。我们注意到，根据这些指标，MusicLM能够生成与Mubert相当的高质量音乐，Mubert依赖于音乐家和声音签名者准备的预先录制的声音。在对输入文本描述的忠实度方面，正如KLD和MCC所捕获的那样，MusicLM获得了最好的分数，这表明与基线相比，它能够从文本描述中捕获更多信息。

我们用人类的听觉测试来进一步补充我们对文本忠实度的评估。参与者会看到两个10秒的片段和一个文字说明，并被问及哪一个片段最能被说明的文字所描述，评分标准为5分Likert。我们收集了1200个评分，每个来源都参与了600个配对的比较。表1报告了"胜利"的总数，也就是说，计算人类评分者在并排比较中喜欢一个模型的频率。MusicLM显然比两个基线都要好，而与地面真实参考音乐仍有可观的差距。听力研究的全部细节可以在附录B中找到。

听听那些通过MusicLM进行预推理的例子，可以发现以下模式：（1）标题非常详细，提到了五个以上的乐器或者描述了非音乐方面的内容，比如*"风，人们在说话"*；（2）标题描述了正在播放的音频的时间顺序；（3）使用了否定词，MuLan不能很好地捕捉这些否定词。

总的来说，我们得出的结论是。(1)我们的方法能够从MusicCaps丰富的自由文本标题中捕捉到细粒度的信息；(2)KLD和MCC指标提供了对文本描述忠实度的定量测量，这与人类评分研究是一致的。

语义标记的重要性。为了了解将语义模型与声学模型脱钩的用处，我们可以将语义模型与声学模型脱钩。我们训练了一个Transformer模型，通过对MuLan标记的建模，直接预测粗略的声学标记。$p(A_t \mid A_{<t}, M_A)$。我们观察到，虽然FAD度量是相当的（0.42 $\text{FAD}_{\text{Trill}}$ 和 4.0 FAD_{VGG}），KLD和移除语义建模阶段后，MCC得分会恶化。特别是KLD得分从1.01上升到1.05，而MCC得分从0.51下降到0.49，这表明语义代币促进了对文本描述的依从。我们还通过聆听样本来证实这一点。此外，我们观察到长期结构的退化。

由音频标记代表的信息。我们进行额外的实验来研究语义和声学标记所捕获的信息。在第一项研究中，我们固定了MuLan文本标记以及语义标记，多次运行声学建模阶段以生成几个样本。在这种情况下，通过聆听生成的音乐，可以观察到样本是多种多样的，但它们往往具有相同的流派、节奏特性（如鼓声）和部分主旋律。它们在具体的声学特性（如混响水平、失真）方面有所不同，在某些情况下，不同的例子中可以合成具有相似音高范围的不同乐器。在第二项研究中，我们只固定MuLan文本标记，同时生成语义和声学标记。在这种情况下，我们观察到在旋律和节奏特性方面有更高水平的多样性，仍然与文本描述相一致。我们在随附的材料中提供了这项研究的样本。

记忆分析。图3报告了当语义符号提示的长度在0到10秒之间变化时的准确和近似匹配。我们观察到，准确匹配的比例始终非常小（<0.2%），即使是在使用10秒的提示来产生5秒的延续时也是如此。图3也包含了近似匹配的结果，使用$\tau=0.85$。我们可以看到用这种方法检测到的匹配数量较多，在只使用MuLan标记作为输入时也是如此（提示长度$T=0$），而且随着提示长度的增加，匹配样本的比例也在增加。我们更仔细地检查这些匹配，发现那些匹配分数最低的序列对应于低水平的标记多样性。也就是说，125个语义符号样本的平均经验熵是4.6比特，而当考虑到被检测为匹配分数小于0.5的近似匹配的序列时，它下降到1.0比特。我们在随附的材料中包括了一个用$T=0$获得的近似匹配样本。请注意，由第二阶段进行的声学建模在生成的样本中引入了进一步的多样性，在语义符号完全匹配时也是如此。

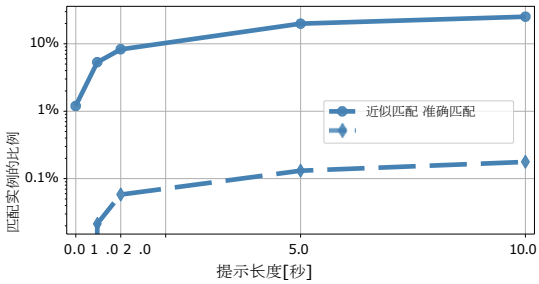


*图 3.*语义建模阶段的记忆结果。我们将5秒钟的音频所产生的语义标记与训练集中的相应标记进行比较，考虑精确和近似的匹配。

6. 延伸

旋律调节。我们对MusicLM进行了扩展，使其能够根据文字描述和旋律生成音乐，而旋律是以哼唱、歌唱、吹口哨或演奏乐器的形式提供的。这需要扩展调节信号的方式，以捕捉目标旋律。为此，我们创建了一个合成数据集，由具有匹配旋律但不同声学的音频对组成。为了创建这样的配对，我们使用同一音乐片段的不同版本，如翻唱、指导或人声。此外，我们还获得了人们哼唱和歌唱的数据对。然后，我们训练一个联合嵌入模型，当两个音频片段包含相同的旋律时，相应的嵌入是相互接近的。关于实施细节，我们参考附录C。

为了提取MusicLM的旋律条件，我们用RVQ对旋律嵌入进行量化，并将得到的标记序列与MuLan音频标记M_A。在推理过程中，我们从输入音频片段中计算出旋律标记，并将它们与MuLan文本标记M_T。基于这一条件，MusicLM可以成功地完全生成遵循输入音频片段中的旋律的音乐，同时遵守文本描述。

长的生成和故事模式。在MusicLM中，基因定量在时间维度上是自回归的，这使得它有可能生成比训练时所用的序列更长的序列。在实践中，语义建模阶段是在30秒的序列上训练的。为了生成更长的序列，我们以15秒的步幅前进，用15秒作为前缀来生成额外的15秒，始终以同一文本描述为条件。通过这种方法，我们可以生成在几分钟内连贯的长音频序列。

通过一个小的修改，我们可以生成长的音频句子，同时随着时间的推移改变文字描述。根据Villegas等人（2022）在视频生成方面的经验，我们把这种方法称为*故事模式*。构成具体来说，我们从多个文本描述中计算出M_T，并每15秒改变一次调节信号。该模型产生平滑的过渡，这些过渡与节奏一致且语义合理，同时根据文本描述改变音乐背景。

7.结论

我们介绍了MusicLM，一个以文本为条件的生成性的该模型能以24kHz的频率产生高质量的音乐，持续数分钟，同时忠实于文本调节信号。我们在MusicCaps上证明了我们的方法优于基线，MusicCaps是一个由音乐家准备的5.5千首音乐-文本对的手工整理的高质量数据集。

我们的方法的一些局限性继承自MuLan，因为我们的模型误解了否定句，并且没有坚持文本中描述的精确的时间排序。此外，我们的定量评估还需要进一步改进。具体来说，由于MCC也依赖于MuLan，MCC的分数对我们的方法是有利的。

未来的工作可能会集中在歌词的生成上，同时改进文本调节和声乐质量。另一个方面是对高级别歌曲结构的建模，如引子、诗句和合唱。以更高的采样率对音乐进行建模是一个额外的目标。

8.更广泛的影响

MusicLM根据文本描述生成高质量的音乐，因此它进一步扩展了协助人类完成创造性音乐任务的工具集。然而，我们的模型和它所处理的用例也有一些风险。生成的样本将反映训练数据中存在的偏见，提出了对训练数据中未被充分理解的文化进行音乐生成的适当性问题，同时也提出了对文化盗用的担忧。

我们承认有可能盗用与本案例有关的创造性内容的风险。根据负责任的模型开发实践，我们对记忆进行了彻底的研究，改编和扩展了基于文本的LLMs的方法，重点是语义建模阶段。我们发现，只有极小部分的例子被完全记住了，而对于1%的例子，我们可以确定一个近似的匹配。我们强烈强调，在解决这些与音乐基因相关的风险方面，未来需要更多的工作--我们目前还没有计划发布模型。

参考文献

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., T oderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube- 8m。一个大规模的视频分类基准。*arXiv:1609.08675*，2016。

Borsos, Z., Marinier, R., Vincent, D., Kharitono v, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: a language modeling approach to audio generation. *arXiv: 2209.03143*, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. M., 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. Language models are few-shot learners。在*Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herber t- Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models, 2020。URL https://arxiv.org/abs/2012.07805。

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F ., and Zhang, C. Quantifying memorization across neural language models, 2022.URL https://arxiv.org/ abs/2202.07646.

Chung, Y., Zhang, Y., Han, W., Chiu, C., Qin, J., P ang, R., and Wu, Y. W2v-bert: Combining contrastive learn- ing and masked language modeling for self-supervised speech pre-training. *arXiv: 2108.06209*, 2021.

Cohen, A. D., Roberts, A., Molina, A., Butryna, A., Jin, A., Kulshreshtha, A., Hutchinson, B., Zevenbergen, B., Aguera-Arcas, B. H., ching Chang, C., Cui, Du, C., Adiwardana, D. D. F., Chen, D., Lepikhin, D. D. Chi。E.H., Hoffman-John, E., Cheng, H.-T., Lee, H., Kri-T. vokon, I., Qin, J., Hall, J., Fenton, J., Soraker, J., Meier-Hellstern, K., Olson, K., Aroyo, L. M., Bosma, M. P., Pickett, M. J., Menegali, M. A., Croak, M., D´ıaz, M. 。Lamm, M., Krikun, M., Morris, M. R., Shazeer, N., Le, Q.V., Bernstein, R., Rajakumar, R., Kurzweil, R., Thop- pilan, R., Zheng, S., Bos, T., Duke, T., Doshi, T., Zhao。V.Y., Prabhakaran, V., Rusch, W., Li, Y., Huang, Y., Zhou, Y., Xu, Y., and Chen, Z. Lamda。*arXiv:2201.08239*，2022年，用于对话应用的语言模型。

Cuturi, M. Sinkhorn distances:最佳运输的光速计算。在*神经信息处理系统的进展（NeurIPS）*中，2013年。

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X.FMA：一个用于音乐分析的数据集。In *International Society for Music Information Retrieval Conference (IS- MIR)*, 2017.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding.在*NAACL-HLT*，2019年。

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox:*arXiv:2005.00341*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby。
N.一张图片价值16x16个字。规模化图像识别的变形器。In *International Conference on Learning Representations (ICLR)*, 2021.

De´fossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv:2210.13438*, 2022.

Engel, J. H., Hantrakul, L., Gu, C., and Roberts, A. DDSP：可微分数字信号处理。In *International Conference on Learning Representations (ICLR)*, 2020.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for highresolution image synthesis.在*IEEE计算机视觉和模式识别会议（CVPR）上*，2021年。

Forsgren, S. and Martiros, H. Riffusion - Stable diffusion for real-time music generation, 2022.URL https://riffusion.com/about。

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. 音频集。一个本体论和人类标记的音频事件数据集。在*IEEE声学、语音和信号处理国际会议（ICASSP）上*。IEEE，2017。

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.A., Dieleman, S., Elsen, E., Engel, J. H., and Eck, D. Enabling factorized piano music modeling and generation with the MAESTRO dataset。In *International Conference on Learning Representations (ICLR)*, 2019.

Hawthorne, C., Jaegle, A., Cangea, C., Borgeaud, S., Nash, C., Malinowski, M., Dieleman, S., Vinyals, O., Botvinick,
M.M., Simon, I., Sheahan, H., Zeghidour, N., Alayrac, J., Carreira, J., and Engel, J. H. General-purpose, long-context autoregressive modeling with perceiver AR.在Chaudhuri, K., Jegelka, S., Song, L., Szepesva´ri, C., Niu, G., and Sabato, S.（编辑），*国际机器学习会议（ICML）*，2022a。

Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., and Engel, J. H. Multi-Instrument Music synthesis with spectrogram diffusion. *arXiv:2206.05408*, 2022b.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R., and Wilson, K.用于大规模音频分类的Cnn架构。In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models.*Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv:2204.03458*, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo:*arXiv:2205.15868*, 2022年，通过变换器对文本-视频基因进行大规模预训练。

Huang, C. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. Music transformer。具有长期结构的基因评级音乐。In *International Conference on Learning Representations (ICLR)*, 2019.

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. W. Mulan。音乐音频和自然语言的联合嵌入。在*国际音乐信息检索协会会议（ISMIR）上*，2022年。

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fre´chet audio distance:用于评估音乐增强算法的无参考指标。在*INTERSPEECH*，2019年。

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps:为野外的音频生成字幕。In *NAACL- HLT*, 2019.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., De´fossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audio- gen:文本指导的音频生成，2022年。

穆博特-公司 穆博特。 https://mubert.com/。 https://github.com/MubertAI/穆贝尔-文字转音乐，2022年。

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and edit- ing with text-guided diffusion models。在*国际机器学习会议（ICML）上*，2022年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision。在*国际机器学习会议（ICML）上*，2021年。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S. 。Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer.*机器学习研究杂志（JMLR）*, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Rad- ford, A., Chen, M., and Sutskever, I. Zero-shot text to image generation.In Meila, M. and Zhang, T. (eds.), *Inter- national Conference on Machine Learning (ICML)*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M.*ArXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Om- mer, B. High-resolution image synthesis with latent diffu- sion models.In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684-10695, June 2022a.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Om- mer, B. High-resolution image synthesis with latent diffu- sion models.*在IEEE/CVF计算机视觉和模式识别会议论文集中*，2022b。

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Den-. ton, E. L., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S.S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion mod- els with deep language understanding. *arXiv:2205.11487*, 2022.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering.在*IEEE计算机视觉和模式识别会议（CVPR）论文集中*，2015年。

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chau- mont Quitry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. Towards Learning a Universal Non Semantic Representation of Speech。在*INTERSPEECH*，2020。

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. Wavenet: A Generative model for raw audio.在*ISCA*，2016年。

Van Den Oord, A., Vinyals, O., et al. Neural discrete repre- sentation learning.*Advances in neural information pro- cessing systems (NeurIPS)*, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need.*Advances in neural information processing systems (NeurIPS)*, 2017.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki。*arXiv:2210.02399*, 2022年，从开放域文本描述中生成可变长度的视频。

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. Nu¨wa:用于神经视觉世界创建的视觉合成预训练。在*欧洲计算机视觉会议（ECCV）上*，2022a。

Wu, H., Seetharaman, P., Kumar, K., and Bello, J. P. Wav2clip:从CLIP中学习强大的音频表征。在*国际声学、语音和信号处理会议（ICASSP）上*，2022b。

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound:用于文本到声音生成的离散扩散模型。*arXiv:2207.09983*, 2022。

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y. Scaling autoregressive models for content-rich text toimage generation, 2022。

Zeghidour, N., Teboul, O., de Chaumont Quitry, F., and Tagliasacchi, M. LEAF: A learnable frontend for audio classification.在*第九届国际学习代表会议上，ICLR 2021，虚拟活动，奥地利，2021年5月3-7日*。OpenReview.net, 2021.URL https://openreview.net/forum?id=jM76BCb6F9m。

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream。一个端到端的神经音频编解码器。*IEEE ACM Trans.Audio Speech Lang.Process.*, 30, 2022.

Zen, H., Senior, A., and Schuster, M. Statistical parametric speech synthesis using deep neural networks.In*ternational Conference on Acoustics, Speech and Signal Processing（ICASSP）*, 2013.

A. 音乐帽数据集

与本文一起，我们发布了MusicCaps，一个高质量的音乐字幕数据集。[6]这个数据集包括来自AudioSet（Gemmeke等人，2017）的音乐片段，与相应的英文文本描述配对。它总共包含5,521个例子，其中2,858个来自AudioSet评估，2,663个来自AudioSet训练分割。我们进一步将1,000个例子标记为我们的数据集的平衡子集，它在所包含的音乐流派方面是平衡的。平衡子集中的所有例子都来自AudioSet的评估部分。

自由文本标题的例子。

- *"这首民歌由一个男声在情绪激动的情况下演唱主旋律。伴随着手风琴在背景中演奏的填充物。一把小提琴奏出的旋律悠扬动听。这首歌没有打击乐器。这首歌曲可以在中亚古典音乐会上演奏"。*
- *"这是一个键盘手在电子琴上演奏十二小节蓝调的现场录音。演奏者在和弦变化之间加入点缀，这首曲子听起来很有节奏感，很有蓝调，很有灵魂。*
- *"一个合成器正在演奏一个琶音弹奏，有大量的混响在速度上上升和下降。另一个合成器的声音正在演奏垫子和一个次要的基音。这首歌充满了合成器的声音，创造了一种舒缓和冒险的气氛。这首歌可能会在音乐节上播放，在两首歌曲中进行铺垫"。*
- *"一个低沉的男声在快节奏的鼓点上说唱，与贝斯一起演奏雷鬼的节拍。像吉他一样的东西正在演奏旋律。这段录音的音频质量很差。在背景中可以注意到一阵笑声。这首歌可能是在一个酒吧里播放的。*
- *"电子音乐的特点是大约每两秒钟重复一次的部分。它包括一个由踢鼓和拍子组成的节拍。一个嗡嗡作响的合成器通过每两拍播放一次来设定音乐的脉动。整个音乐听起来就像一个循环被反复播放。在节选的最后，可以听到类似高潮的嗡嗡声，增加了紧张感"。*

方面清单的例子。

- *"流行音乐、尖锐的宽高跟鞋、圆润的钢琴旋律、高亢的女声旋律、持续脉动的合成器主音、柔和的女声、有力的踢腿、持续的合成器低音、拍手、情感、悲伤、激情"*
- *"业余录音，指头的片段，男中音歌唱，混响"*
- *"背景音乐，爵士乐，数字鼓，钢琴，电子低音提琴，小号，原声吉他，数字键盘歌曲，中等节奏"*
- *"Rubab乐器，不同八度的重复旋律，没有其他乐器，弹拨弦乐器，没有声音，器乐，快节奏"*
- *"器乐、白噪音、女性发声、三首不相关的曲目、电吉他和声、贝斯吉他、键盘和声、女性主唱、键盘和声、光滑的鼓声、低音落音、男声伴唱"*

[6]kaggle.com/datasets/googleai/musiccaps

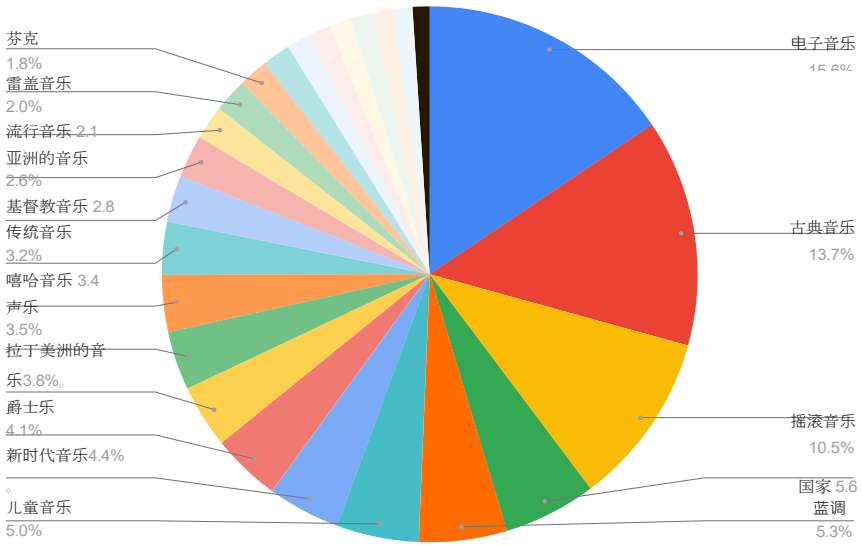


*图4.*根据AudioSet分类器，所有5.5千例MusicCaps的流派分布。

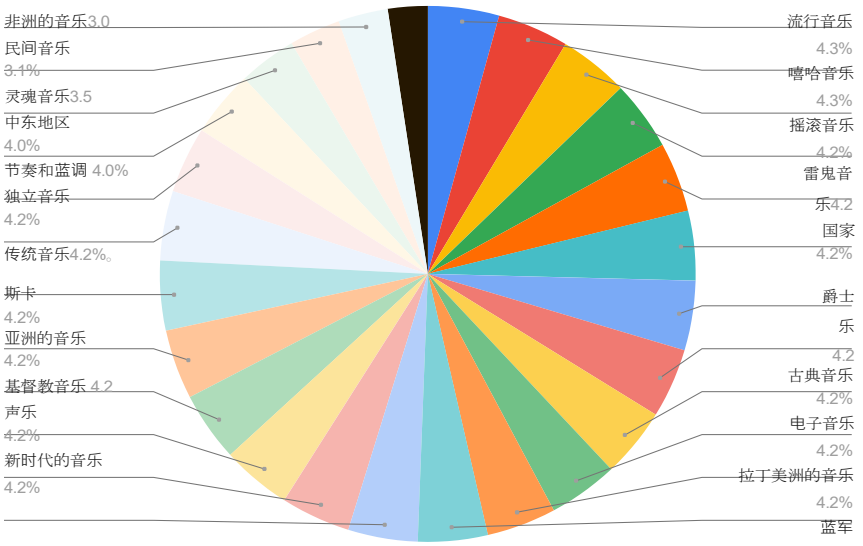


*图5.*根据AudioSet分类器，MusicCaps的一个平衡的1k例子子集的流派分布。

B. 定性评价

听力测试中的参与者会看到两个10秒钟的片段和一个文字说明，并被问及哪一个片段最能描述说明的文字，采用5分制的李克特评分。他们还被要求忽略音频质量，只关注文本与音乐的匹配程度（类似于MuLan评分）。图6显示了呈现给评分者的用户界面。

我们收集了1200个评分，每个来源参与了600个配对比较。图7和图8显示了各模型之间成对比较的细化结果。根据使用Wilcoxon符号秩检验和Bonferroni校正的事后分析（$P<0.01/15$），图8中显示的评分者的排序都有统计学意义。

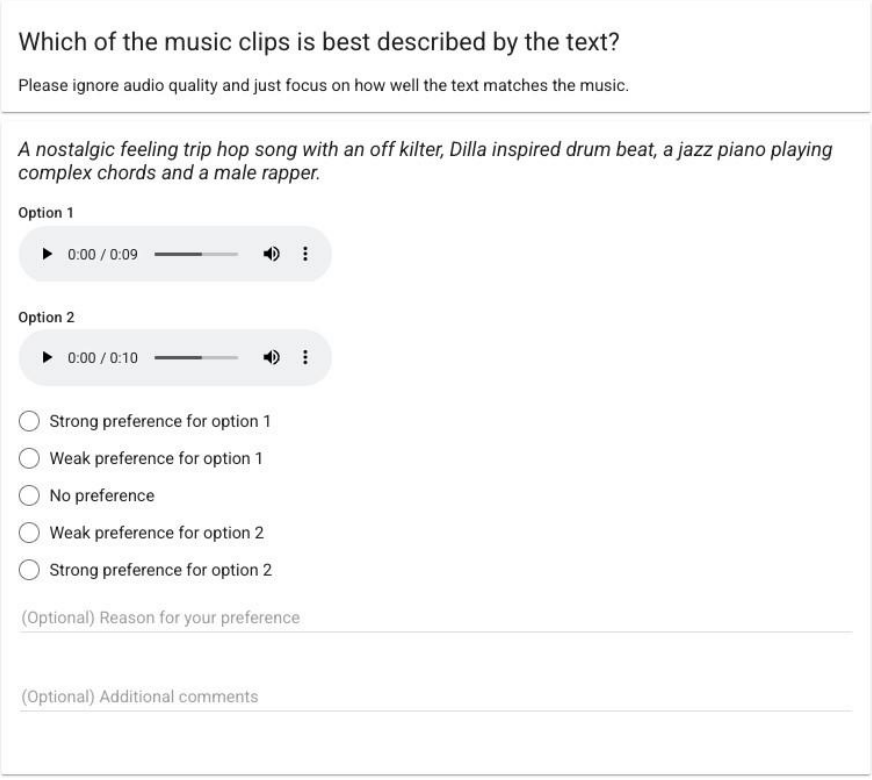


*图6.*人类听众研究的用户界面。

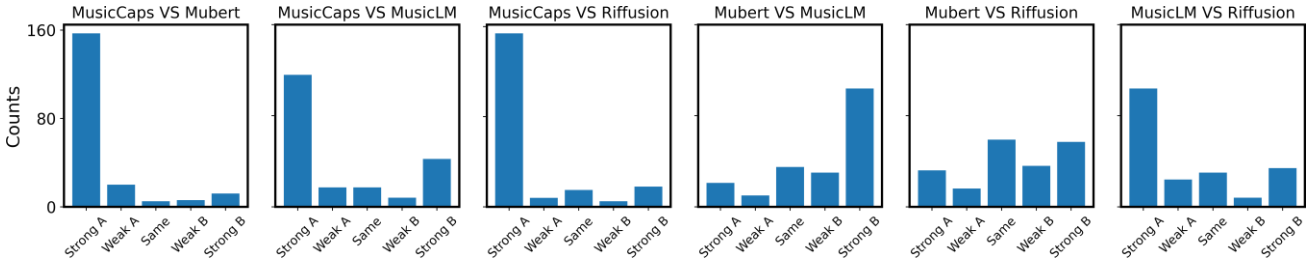


*图7.*人类听众研究中的配对比较。每一对都是以5分的李克特量表进行比较。除了Mubert与Riffusion的比较，评分者在所有情况下都有决定性的模型偏好。

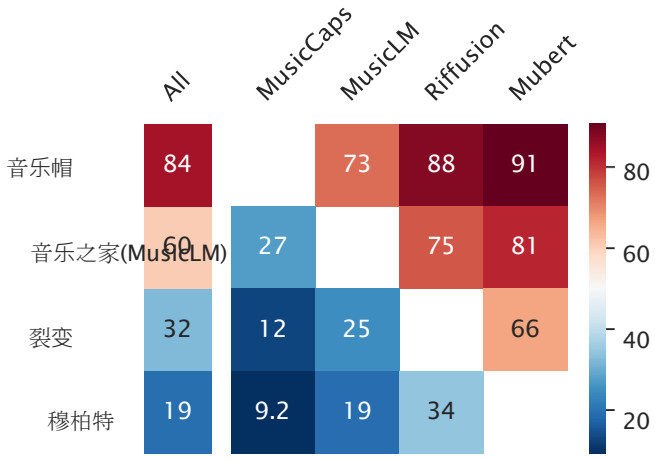


*图 8.*人类听众研究中的获胜百分比。每一行表示听众发现该系统的音乐与任何其他系统（第一列，N=1200）和每个系统单独（其他列，N=600）的标题更好地匹配的次数。地面真相数据（MusicCaps）显然是与字幕最匹配的，但紧随其后的是MusicLM，它甚至在27%的比较中击败了地面真相。

C. 美乐迪调理

我们在此提供用于调节旋律的音乐生成的模型的实施细节。该模型是基于一个小型的ViT（Dosovitskiy等人，2021），由12层、6个注意头、512的嵌入维度和1024维度的前馈层组成。该模型的输入是音频的融化谱图的时间帧。我们使用半硬性三连音损失（Schroff等人，2015）来训练旋律嵌入模型，为每4秒的音频生成192维的嵌入。该模型学习生成代表旋律的嵌入，同时不受与所演奏乐器相关的声学特性影响。这是特别有利的，因为这个代表是对MuLan嵌入学习的代表的补充。因此，我们的旋律嵌入和MuLan可以共同和互补地用于调节音乐生成过程。在训练过程中，我们考虑了持续时间为10秒的输入音频。我们提取三个跳长为3秒的旋律嵌入，用残余矢量量化（RVQ）将每个嵌入离散为标记，并将得到的标记序列与MuLan音频标记M_A。我们使用由24个量化器组成的RVQ，每个量化器的词汇量为512。