利用神经表征进行音频操控

Scott H. Hawley[1,3] *, Christian J. Steinmetz[2]

[1]贝尔蒙特大学,纳什维尔,田纳西州, 美国[2]数字音乐中心, 伦敦玛丽女王大学, 英国[3]Harmonai, 美国

arXiv:2304.04394v1 [eess.AS] 10 Apr 2023

摘要

我们研究了使用基于预训练神经网络的自编码器进行音频操控，将其作为传统信号处理方法的替代方案，因为前者可能提供更强的语义或感知组织。为了确立这一方法的潜力，我们首先考察这些模型中的表征是否编码了有关操控的信息。我们使用来自两个不同预训练自编码器的表征开展实验并生成可视化结果。我们的结果表明，尽管其中编码了一些有关音频操控的信息，但这些信息既有限，其编码方式也并不简单。这一点得到了我们对这些表征进行可视化的尝试的支持：即使对于线性信号操控，常见操控对应的表征轨迹通常也是非线性的，并且取决于内容。因此，目前尚不清楚如何使用这些预训练自编码器来操控音频信号；然而，我们的结果表明，这可能是由于相对于常见音频操控缺乏解耦所致。

1 引言

音乐音频制作工作流程使用各种参数化变换来执行音频信号的处理和重新合成。示例包括多段均衡器上的滑块、通过调整增益或压缩实现的动态范围变化、频谱处理，以及通过更改 MIDI 参数所进行的调整。越来越“智能”的音乐接口的开发，可以被视为一种追求：寻找能够产生 表征的变换，使这些表征更加贴合音乐制作人所关注的感知或语义内容，例如使用更少的旋钮来控制声音的高层面特征 [1]。

基于神经网络的音频自动编码器已在许多应用中展现出前景，包括音频编码 [2, 3] 以及音乐风格特征的迁移，例如乐器类型 [4,5] 和音频制作细节 [6]。所选择的编码器类型将生成通常更适合某些任务，而不是其他任务。此类任务通常采用分类和/或音乐信息检索 [7, 8, 9] 的形式。我们通过冻结针对音频重建进行优化的预训练自编码器的权重，专注于音频信号的分析与合成。此类自编码器产生的潜在表征可能编码语义或风格信息 [9]。

为了将这项工作置于适当的背景下，可以考虑通过短时傅里叶变换（STFT）计算得到的频谱图，这些频谱图可以被视为多维空间中的“向量”列。神经网络系统生成的表征同样可以被视为一组向量，或视为“神经 [激-活] 频谱图”或“特征图”中的列。图 1 展示了这类类似频谱图的表征。

以往用于音频效果变换的神经网络方法 [10, 11, 12] 都是通过监督学习以端到端的方式创建的。本文探讨的变换仅作用于预训练自动编码器的潜在表征。通过利用已

*通讯作者: scott.hawley@belmont.edu

预印本： 2023年4月11日

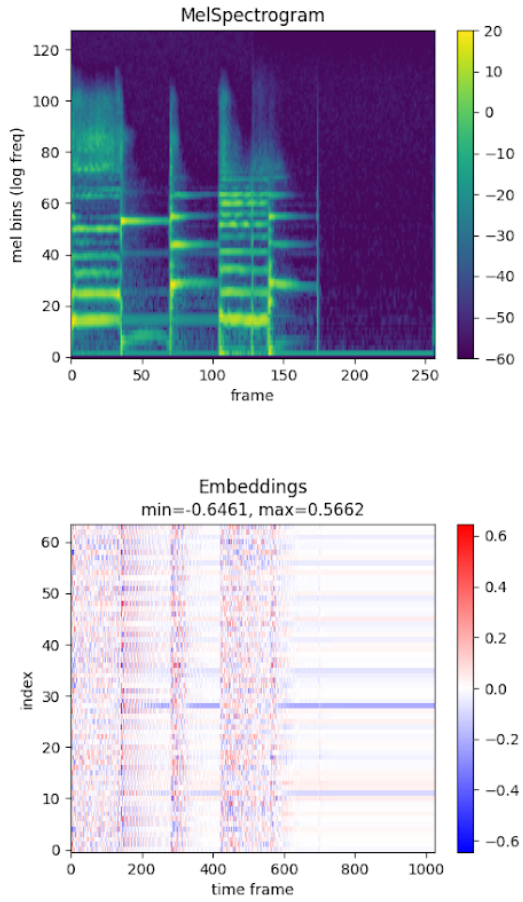

图 1： 梅尔频谱图（上）以及 DiffAE 扩散自编码器生成的潜在表征“频谱图”。

在大型数据集上完成预训练的系统中的编码器和解码器部分，人们或许能够从相对较小的数据集中发现潜在表征的变换。与通常会进一步更新模型权重的迁移学习方法不同，本研究旨在探索：仅操纵冻结的预训练模型的中间激活值或表征，是否能够实现有用的信号处理。

本文为基于自监督训练方案开发少样本或零样本音乐音频变换奠定了基础。我们最初尝试使用预训练自编码器表征中的向量代数运算，对音频制作效果进行零样本风格迁移 [1]，但结果大多不成功，这促使我们开展了本文中的可视化与分类测试：我们的目标是更好地理解这类表征的特征空间。

我们假设，“良好”的嵌入空间也能够与基于文本的控制系统实现紧密耦合，正如文本到图像模型中所见 [13, 14, 15]。识别潜在空间中具有感知意义的变换，还有助于发现和设计新颖的音频操控，从而为新的音频效果设计方法铺平道路，并使用户能够更直观地探索和发现新颖的声音操控 [16]。

使用潜在空间表征的一项关键挑战，是如何解耦不同的维度，例如，让旋钮和滑块等用户控制项只产生一种主要的（感知或语义）效果 [17, 18, 5]。维度解耦也推动了少样本语音风格迁移的改进 [19]。对比方法 [20, 21]的应用表明，在潜在空间中填充语义信息与解耦维度可以兼得。我们选择使用现有的自编码器，它们未必针对潜在维度的解耦进行过优化，以此探索其解耦程度。

为探究这一问题，我们使用主成分分析和 UMAP 将表征投影到二维和三维空间并进行可视化 [22], ，同时开展分类实验，以量化表征中编码的操作信息量及其编码方式。我们希望这些研究能推动音乐音频制作的发展，让基于内容的操作和语义操作比目前更容易实现。

2 方法

2.1 模型

我们研究的模型是由 Harmonai [2] 内部开发的两个预训练扩散自编码器。第一个模型我们称为“DiffAE”，它使用 64 维潜在表示空间；图 1 下方的图像来自 DiffAE 模型。第二个模型是一个两阶段级联潜在扩散模型 [23,

[1] Colab 笔记本：https://tinyurl.com/Destructo-ipynb，2022年10月

[2] 这些模型尚未发表，但与 https://github.com/Harmonai-org 上正在开发的模型类似

表 1： 分类测试的参数设置。除压缩器使用的“比率”值为 5 外，其他设置均为默认值。

Abbr	效果名称	参数	值
CLN	干净	-	-
CHS	合唱	速率（Hz）	1
CMP	压缩器	阈值（dB）	-50
DLY	延迟	延迟时间（秒）	0.5
DIS	失真	驱动（dB）	25
HPF	高通滤波器	截止频率（Hz）	2000
LPF	低通滤波器	截止频率（Hz）	70
PS	音高移调	半音	4
RVB	混响	房间大小	0.8
TRV	时间反转	-	-

表 2： 参数变化测试的设置，使用 32 个增量值，从最小值到最大值，其中高通滤波器和低通滤波器的取值按对数变化。其他所有设置均为默认值。

效果名称	参数	最小值 - 最大值
失真（DIS）	驱动（dB）	0 - 30
混响（RVB）	房间大小	0.01 - 0.99
高通滤波器（HPF）	截止频率（Hz）	50 - 10000
低通滤波器（LPF）	截止频率（Hz）	50 - 10000

24] 我们将其称为“堆叠式DiffAE”，或简称为“Stacked”模型。两个阶段的表征均为 32 维，但第2阶段的表征在时间上更加紧凑，其时间分辨率比第一阶段更粗。除非另有说明，本文始终使用规模更小、更加紧凑的阶段2表征。本文使用的（规模更大的）第一阶段表征，是在解码器一侧通过对（更加紧凑的）第二阶段表征进行上采样并执行额外的扩散而获得的。我们没有使用任何来自编码器一侧的第一阶段表征。除了架构差异之外，这两个模型是在不同的数据集上训练的，这些数据集取样自一个包含多种音乐的未发布仓库。在本研究中，我们的目的并不是确定表征中的哪些差异源于预训练自编码器的具体差异；相反，我们使用多个模型，以避免仅考察单个自编码器模型可能导致的过度泛化。

2.2 数据集

我们构建了一个包含 1024 个 48kHz 音频样本的数据集，每个样本长 5.4 秒（2^{18} 样本）。我们从 GuitarSet [25], 中获取了 512 个吉他音，其中包括独奏音符和扫弦和弦的演奏。其余 512 个样本是从 MAESTRO 数据集 [26]的 2018 子集中采样的钢琴音。我们选择只使用两种乐器进行独奏，这样音频效果处理的结果便能比应用于乐器种类更丰富的音乐时更加清晰地呈现。

音频数据集。所有输入声音均为单声道录音，并被“加倍”为立体声（因为模型需要立体声输入），且在通过音频效果处理前后，均使用 pyloudnorm[3] [27]进行响度归一化，以消除音量差异。我们选取了九种使用固定参数的不同音频处理方式，以探索按效果进行聚类的情况；此外，还选取了四种效果，研究改变一个关键参数时的结果。除“干净”旁路和时间反转外，所有音频效果均通过 Pedalboard[4] 应用；分类测试的设置见表 1，参数变化测试的设置见表 2。

3 结果

3.1 可视化

图 2 展示了表示向量随时间变化的轨迹，并将其投影到三个

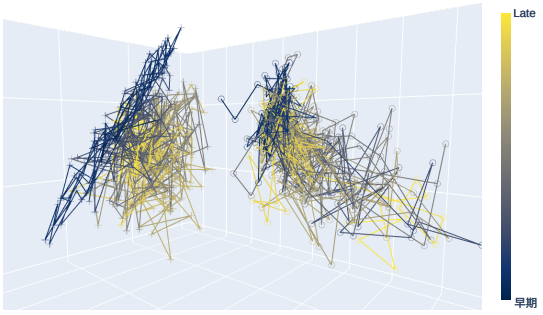

图 2： 一个吉他样本（圆圈）和一个钢琴样本（叉号）的时间轨迹主成分分析图，采用 32 维堆叠式 DiffAE 模型。

[3] https://github.com/csteinmetz1/pyloudnorm[4] https://github.com/spotify/pedalboard

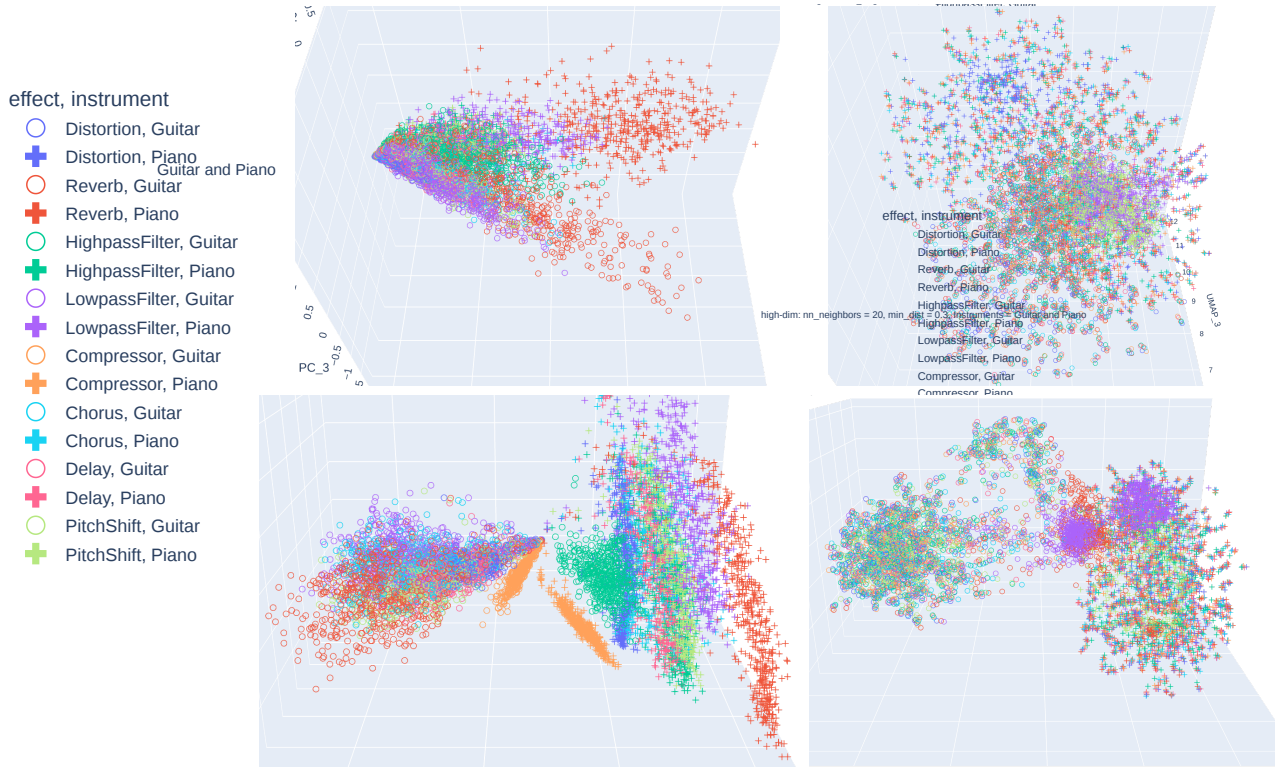

图3： DiffAE 自编码器（上排）和堆叠模型（下排）的展平表征的三维投影[5]，展示了主成分分析（左列）和 UMAP（右列）映射。UMAP 图中可以看到，由演奏表现造成的变化通常比由音频效果造成的变化更显著。两类乐器声音映射到潜在空间中基本互不重叠的区域。

使用主成分分析（主成分分析）对维度进行降维。图中展示的两条轨迹是使用从整个数据集得出的主成分分析变换投影得到的。这些表示向量的"运动"描述了一条复杂的路径。虽然诸如 RAVE [5] 之类的方法会采用"先验模型"，学习以自回归方式预测此类轨迹，从而实现音乐生成，但在本研究中，我们仅指出，从三维 PCA 投影中观察到的轨迹似乎并不遵循任何容易直观理解的模式。

另一种视角来自将所有维度和时间的表征展平为单个向量，这些向量存在于高维空间中。图 3 展示了在主成分分析和UMAP [22] 3D投影中按音频效果类别着色的展平表征。左列中DiffAE（顶部）和 Stacked（底部）模型的主成分分析投影显示出了一些按音频效果类别的分离，其中后者按音频效果呈现出更强的聚类。然而，右列中的 UMAP 图像的

[5] 本文中三维图的交互版本可在 https://api.wandb.ai/links/drscotthawley/36hi1day 查看

图 3表明，除了一两个效果之外，展平表征的聚类并非由类别决定：音乐演奏表现（吉他或钢琴曲目的演奏表现）所导致的变化在数据表征中起着更大的作用，而音频效果只是小幅扰动。

为了获得更强的"风格"信号，以研究按音频效果类别进行的分离，我们对表征进行时间平均，并将结果绘制在图4中。所得的主成分分析数据（左列）看起来与前一幅图相似；然而，UMAP 图（右列）显示，按音频效果分组的效果比前一幅图强得多。这表明，表征的时间平均值或可作为"风格"信息的有效代理，至少也说明这些自编码器模型的表征确实编码了与音频效果类别相关的一些"语义"结构。

最终的可视化研究探讨了由于效果参数值而导致的嵌入变-化。图6至8 展示了改变失真、混响、高通和低通滤波器效果的关键"旋钮"值所产生的影响。一个主要

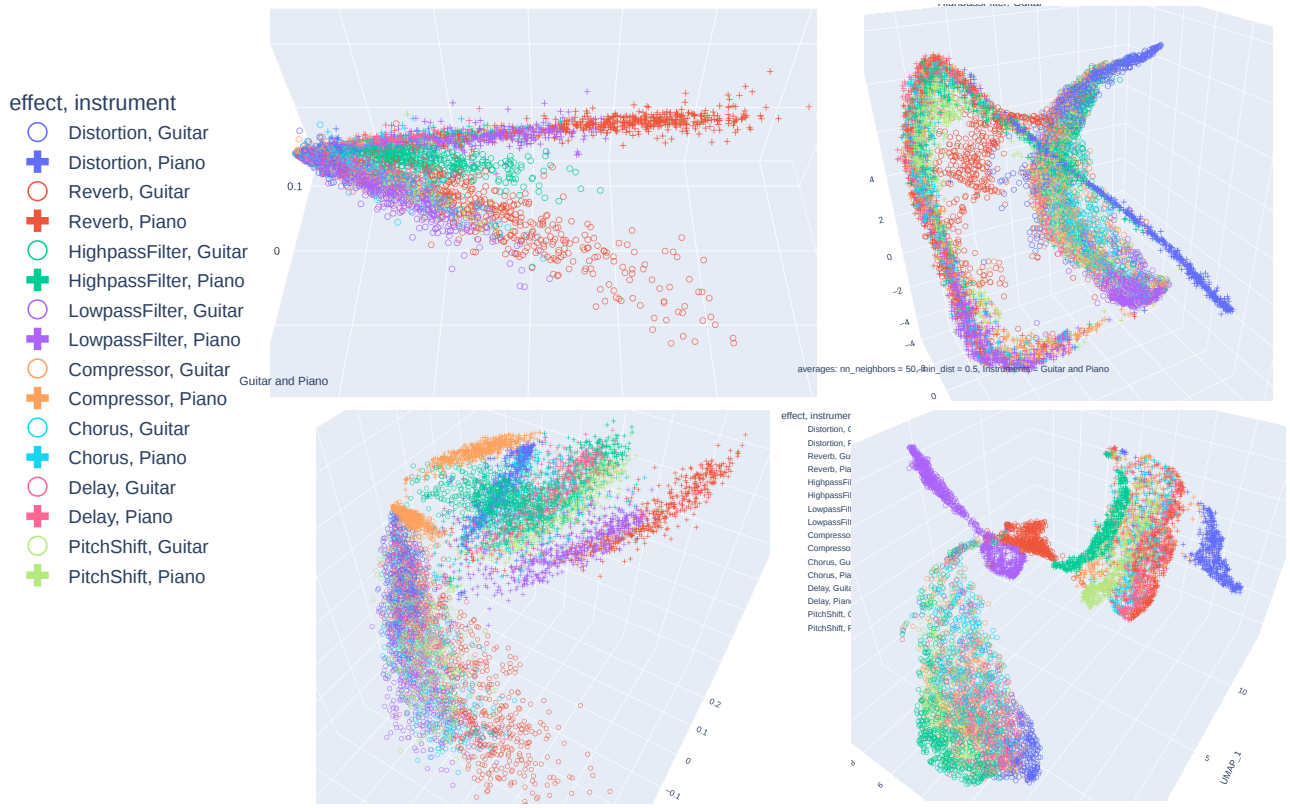

图4： DiffAE 自编码器（上排）和堆叠模型（下排）的时间平均表征的三维投影，展示了主成分分析（左列）和 UMAP（右列）映射。可以看出，与图3中的展平表征相比，时间平均表征通常能更好地依据音频效果进行聚类和分离。与上一幅图一样，我们观察到大规模聚类是按照乐器类别进行的。

观察结果表明，即使对于滤波器这类“线性”效果，改变效果参数通常也会在该空间中形成一条非线性路径。这有助于解释早期尝试基于表示向量的代数运算进行少样本神经音频风格迁移为何未能成功。图9展示了对几个音频样本应用多个效果参数的情况。同样，即使对于 HPF 这类“线性”效果，也能看到弯曲的路径；而或许出乎意料的是，失真路径往往显得相对单向。

3.2 操控分类

为了定量测量我们的自编码器所捕获的有关信号操控的信息程度，我们使用操控分类作为代理任务。与自监督表征学习中采用的评估方法-论类似，我们在冻结的预训练编码器的嵌-入之上使用线性探针对分类 [28] 进行分类。在我们的设置中，我们训练线-

在十分类任务上的探针，预测对源音频施加了9 种不同音频操控中的哪一种，或者是否未进行任何操控。线性探针由一个单层组成，将嵌入空间映射到类别数量。这使得参数量与嵌入维度成正比。

由于我们的编码器会生成一系列表示（随时间变化），我们采用两种不同的方法，将提取出的嵌入输入线性探针。首先，我们计算所有嵌入在时间维度上的平均值，这样可以减少不同时间帧之间的方差。然而，如果音频效果在更大的时间尺度上表现出某种行为，例如混响，那么时间平均可能会移除检测音频效果所需的关键信息。为了解决这种情况，我们还考虑将完整的嵌入序列展平。

在我们的评估中，我们比较了来自堆叠自编码器（阶段1和阶段

编码器	Dim	探针	按类别准确率（%）										
			CHS	CLN	CMP	DLY	DST	高通滤波器	低通滤波器	PS	RVB	TRV	AVG
VGGish-T	128	1.3 k	98.0	80.6	90.1	81.3	91.4	94.5	96.9	86.7	90.4	94.5	92.9
VGGish-F	2176	21.8 千	98.0	76.5	85.2	76.8	92.1	90.1	96.9	83.3	88.7	95.1	90.3
Stacked2-T	32	0.3 千克	24.0	54.9	28.1	34.9	86.6	78.5	95.0	66.7	79.3	32.0	57.7
Stacked2-F	16384	163.9 k	45.2	42.2	30.2	65.1	84.7	80.9	94.4	68.2	71.3	39.2	63.1
Stacked1-T	32	0.3 k	29.9	59.2	27.1	26.5	88.6	71.4	96.9	54.2	74.8	36.8	57.4
Stacked1-F	262144	2621.5 k	14.6	37.6	9.4	29.2	41.9	59.1	88.7	24.5	58.5	48.6	41.1
DiffAE-T	64	0.7 千	11.2	32.5	3.6	9.3	92.9	46.1	86.9	31.0	63.5	27.1	40.3
DiffAE-F	131072	1310.7 千”】【	4.4	32.7	47.7	16.3	73.6	17.0	27.0	19.3	18.6	25.6	28.9

表3： 在预训练编码器的归一化表示之上训练线性探针时的效果分类准确率。T 表示时间平均嵌入，F 表示展平嵌入。表现最佳的模型以粗体标示。

2)，以及 DiffAE。作为非自编码基线，我们还使用从预训练的 VGGish 模型中提取的嵌入[6] [29]。与自编码器类似，VGGish 并未经过专门训练来捕捉有关音频操控的信息。然而，与自编码器不同，它无法根据编码序列重建原始信号。

所有探针均使用批量大小为32、AdamW优化器和学习率为 $3 \cdot 10^{-4}$ 进行训练，最多训练500个周期。我们采用早停策略，耐心值设为50个周期，并监控验证准确率。这些分类测试的结果见表3。虽然所有特征都能实现高于随机猜测的分类准确率，但我们发现，VGGish特征的表现显著优于所有自编码器表示。这表明，所有表征都编码了能够预测音频操控的某些信息；然而，在时间平均配置和展平配置下使用时，自编码器表示所捕获的信息相对较少。

3.3 维度掩蔽

虽然多分类任务实验为音频操控相关信息的程度提供了一些见解，但它并未揭示这些信息是如何编码的。要利用嵌入空间进行信号操控，深入理解信息的编码方式至关重要。为了.....

为探讨这一问题，我们设计了第二项分类实验。

在本实验中，我们研究一项二元分类任务：训练线性模型，以判断给定效果是否存在，或信号是否干净。为了了解信息如何编码在预训练表征中，我们屏蔽表征的一个维度，然后训练分类器。接着，对表征的每个维度和 9 种效果分别重复这一过程，并测量分类准确率的下降幅度。如果屏蔽某个维度会显著降低演奏表现，则说明能够预测相关操控的信息编码在该维度中。

鉴于 Stacked 自编码器模型在上一项实验中表现出色，我们选用该模型进行研究。该模型具有 32 维表征。因此，我们针对 9 种效果分别训练 32 个线性分类器，每次屏蔽表征的一个维度，总计训练 288 个模型。与之前的实验类似，我们使用批量大小 32、AdamW优化器以及 $3 \cdot 10^{-4}$ 的学习率。所有模型均训练 100 个轮次，并根据验证准确率选出最佳模型进行评估。

我们在图 5 中报告结果，其中颜色图表示屏蔽相关维度后，与该效果下性能最佳的模型相比损失的百分点数。我们注意到，对于所有维度和效果，似乎没有任何单一维度特别具有预测性。在最坏的情况下，屏蔽维度——

[6] https://github.com/hearbenchmark/hear-baseline

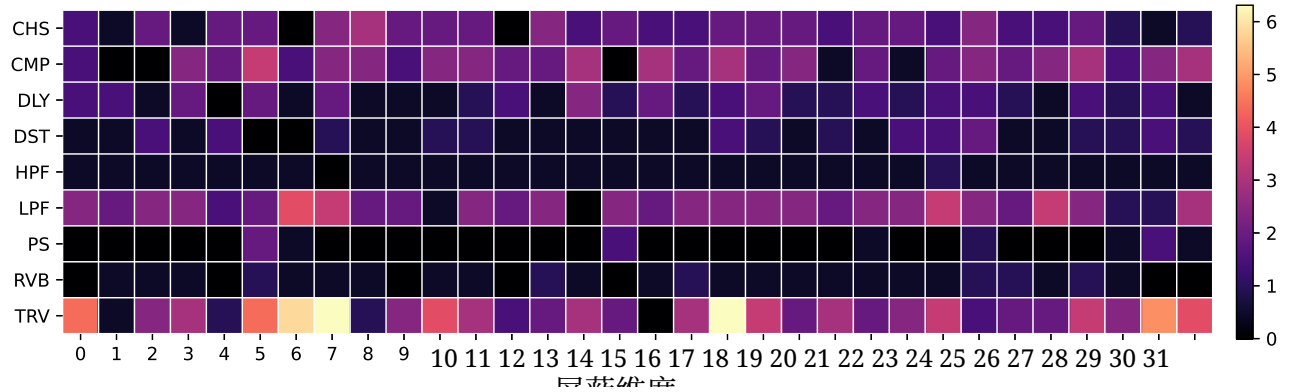

图 5： 评估使用时间平均嵌入检测每种音频效果是否存在时，堆叠自编码器（第 2 阶段）（$D = 32$）各维度的相对重要性。色彩图表示屏蔽表征的一个维度后，二元分类准确率相较于每种效果类型中性能最佳的配置下降的百分点数。

第 7 维和第 18 维使时间反转的分类准确率降低了 ≈ 6 个百分点。不过，似乎屏蔽表征中的任何维度，都会导致某些效果的准确率大幅下降，例如时间反转、低通滤波器、合唱和压缩器；而高通滤波器、移调和混响等效果的准确率变化不大。根据这些结果，我们得出结论：表征针对所研究的操控方式的解耦程度可能较低。

4 总结

我们通过多种类别的音频效果对音乐音频样本进行处理，探索了两个预训练音频自编码器的潜在空间。这些模型的内部表征显示，当考虑音频效果的时间平均值时，会按照音频效果呈现出一定的聚类现象；而以扁平化形式表示的完整表征，则往往会更明显地根据每次单独的音乐演奏表现进行聚类。在所有情况下，空间往往会按照乐器类型划分（例如，一侧是吉他样本，另一侧是钢琴样本）。

为了定量测量音频效果所实现的分离程度，我们使用线性探针进行了分类测试，并以 VGGish 模型作为基线进行比较。我们发现，在基于音频效果进行分类时，冻结的自编码器表征往往明显逊于 VGGish 模型，而时间平均表征的表现优于展平表征——这与我们对可视化结果的观察一致。

通过改变音频效果的参数，我们观察到，时间平均表征在潜在空间中的轨迹往往是非线性的，即使对于高通和低通滤波器等“线性”效果也是如此。因此，基于表征的方法在进行音频操控时必须考虑这种固有的非线性。

本文的研究发现或可为未来开发高效、精密的音乐音频制作方法提供参考，这些方法能够利用神经网络自动编码器倾向于编码语义内容这一特性。我们推测，进一步研究如何解耦潜在空间的各个维度，或可改善音频制作工作流的结果。

5 致谢

本研究由 Stability AI 通过 Harmonai 提供支持。我们感谢 Zach Evans 允许我们使用他预训练的自编码器。C.S. 部分经费由 UKRI 和 EPSRC 提供，资助项目为“英国人工智能与音乐 UKRI CDT”，拨款编号 EP/S022694/1。S.H. 感谢 Max Ortner 进行的有益讨论。

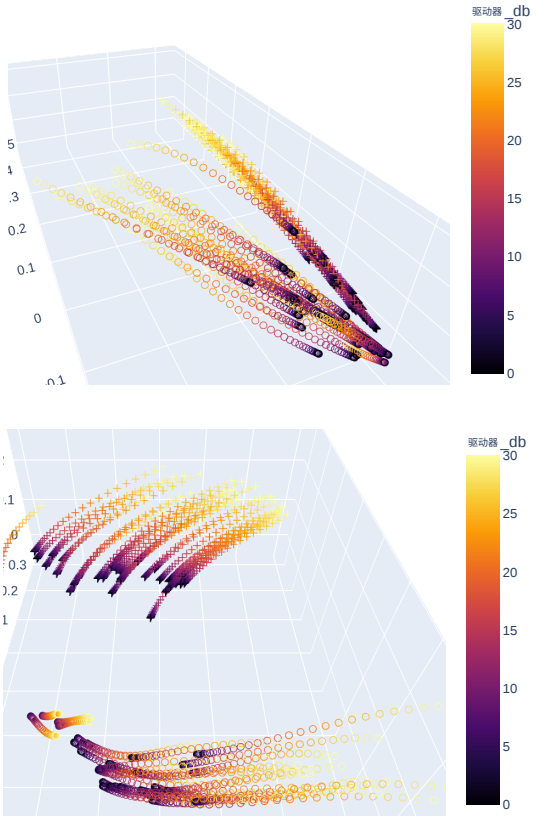

图 6： 对 16 个吉他样本和 16 个钢琴样本施加不同程度的失真效果，分别使用 DiffAE（上）和堆叠式 DiffAE（下）编码器。此处展示时间平均嵌入的主成分分析图。

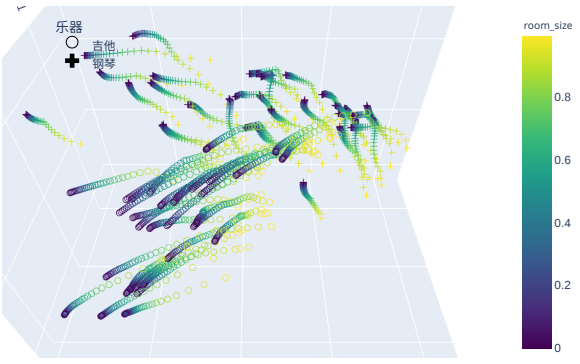

图 7： 使用堆叠模型的时间平均表征绘制的混响效果参数调整路径主成分分析图，其中吉他以圆圈表示，钢琴以叉号表示。

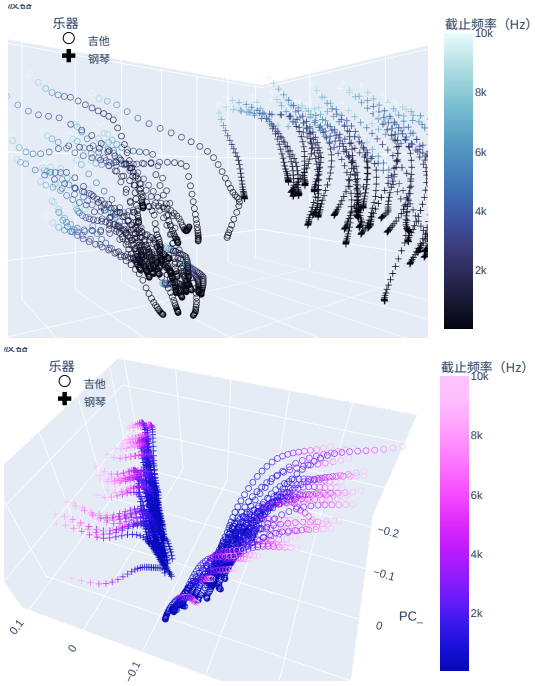

图 8： 堆叠模型中时间平均表示的主成分分析图，针对吉他（圆圈）和钢琴（叉号）调整高通（上）和低通（下）滤波器

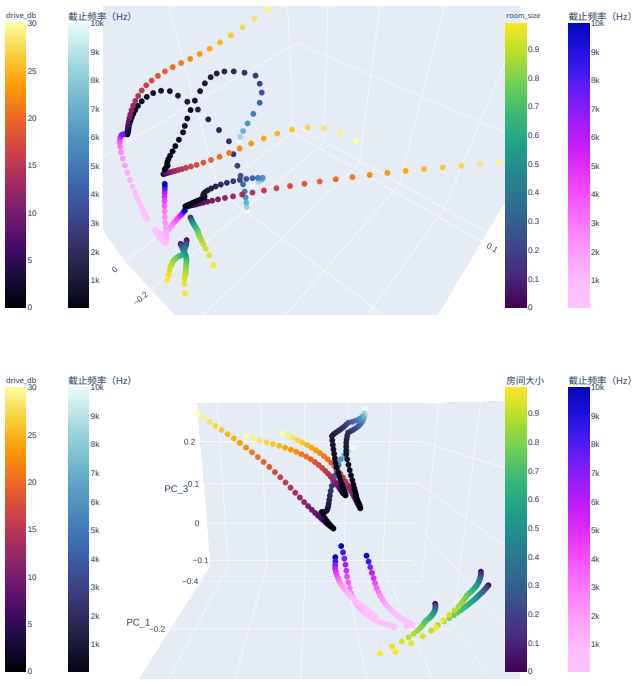

图9： DiffAE（上）和堆叠式 DiffAE（下）模型的多效果轨迹。

参考文献

[1] Ryan Stables 等人。“SAFE：用于提取和检索语义音频描述符的系统”。发表于：(2014)。[2] Neil Zeghidour 等人。“SoundStream：端到端神经音频编解码器”。发表于：*CoRR* abs/2107.03312 (2021)。arXiv：2107 . 03312。[3] Alexandre Défossez 等人。“高保真神经音频压缩”。发表于：*arXiv*预印本 *arXiv:2210.13438* (2022)。[4] Jesse Engel 等人。*DDSP*：可微分数字信号处理。2020。doi：10 . 48550 / ARXIV.2001.04643。[5] Antoine Caillon 和 Philippe Esling。*RAVE*：用于快速、高质量神经音频合成的变分自编码器。2021。doi：10. 48550/ARXIV.2111.05011。[6] Christian J. Steinmetz、Nicholas J. Bryan 和 Joshua D. Reiss。“使用可微分信号处理进行音频效果风格迁移”。发表于：音频工程学会期刊 (2022)。[7] Joseph Turian 等人。*HEAR*：音频表征的整体评估。2022。doi：10 . 48550/ARXIV.2203.03022。[8] Prafulla Dhariwal 等人。*Jukebox*：音乐生成模型。2020。doi：10 .48550 / ARXIV.2005.00341。[9] Rodrigo Castellon、Chris Donahue 和 Percy Liang。编码音频语言建模可学习用于音乐信息检索的实用表征。2021。doi：10 . 48550 / ARXIV . 2107.05677。[10] Marco Martinez Ramirez、Emmanouil Benetos 和 Joshua Reiss。“用于音频效果黑盒建模的深度学习”。发表于：应用科学 10 (2020年1月)，第638页。doi：10.3390/app10020638。[11] Scott H. Hawley、Benjamin Colburn 和 Stylianos Ioannis Mimilakis。“利用深度神经网络分析音频压缩器特性”。发表于：音频工程学会期刊 (2019年10月)。[12] Christian J. Steinmetz 和 Joshua D. Reiss。“用于实时模拟动态范围压缩建模的高效神经网络”。发表于：第*152*届*AES*大会。2022年。[13]Aditya Ramesh 等。“零样本文本到图像生成”。载于：*CoRR*abs/2102.12092（2021）。arXiv: 2102.12092。[14] RobinRombach 等。“使用潜在扩散模型进行高分辨率图像合成”。载于：*CoRR* abs/2112.10752（2021）。arXiv: 2112.10752。[15] Tim Brooks、Aleksander Holynski 和 AlexeiA.Efros。“InstructPix2Pix：学习遵循图像编辑指令”。2022 年 11 月。[16] Christian J.Steinmetz 和Joshua D. Reiss。“神经音频效果的可控发现”。载于：*NeurIPS*第 *5* 届创意与设计研讨会。2021。[17] Erik Härkönen 等。“GANSpace：发现可解释的GAN 控制”。2020 年。doi: 10.48550/ARXIV.2004.02546。[18] Erik Härkönen 等。“GANSpace：发现可解释的 GAN 控制”。载于：*CoRR*abs/2004.02546（2020）。arXiv: 2004.02546。[19] Si yangYuan 等。“通过解耦表征学习改进零样本语音风格迁移”。载于：国际学习表征会议。2021。doi: 10 . 48550 / arXiv . 2103 . 09420。url: https : // openreview .net /forum ? id = TgSVWXw22FQ。[20] Eduardo Fonseca 等。“声音事件表征的无监督对比学习”。载于：*ICASSP2021 -2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。2021，第 371–375 页。doi: 10 . 1109/ ICASSP39728 . 2021 . 9415009。[21] Junghyun Koo等。音乐混音风格迁移：一种解耦音频效果的对比学习方法。2022。doi: 10 . 48550/ARXIV.2211.02247。[22] Leland McInnes 等。“UMAP：统一流形近似与投影”。载于：开放源代码软件期刊3.29（2018），第 861 页。[23] Chitwan Saharia 等。具有深度语言理解能力的逼真文本到图像扩散模型。2022。arXiv: 2205 . 11487[cs.CV]。

[24] Zach Evans、Scott H.Hawley 和KatherineCrowson。“基于联合文本嵌入生成的音乐音频样本”。载于：美国声学学会期刊152.4（2022），A178–A178。doi：10.1121/10. 0015956。电子预印本：https: //doi . org / 10 . 1121/10.0015956。网址：https://doi.org/10.1121/10.0015956。[25] QingyangXi 等。“GuitarSet：用于吉他转录的数据集。”载于：*ISMIR*。2018，第 453–460 页。[26] CurtisHawthorne 等。“借助 MAESTRO 数据集实现因子化钢琴音乐建模与生成”。载于：学习表示国际会议。2019。网址：https://openreview.net/ forum?id=r1lYRjC9F7。[27] Christian J Steinmetz 和JoshuaDReiss。“pyloudnorm：Python中简单而灵活的响度计”。载于：音频工程学会第 *150* 届大会。音频工程学会。2021。[28] XinleiChen、Saining Xie和 KaimingHe。“自监督视觉 Transformer 训练的实证研究”。载于：*IEEE/CVF*国际计算机视觉会议论文集。2021，第9640–9649 页。[29] ShawnHershey 等。“用于大规模音频分类的 CNN 架构”。载于：*ICASSP*。IEEE。2017，第 131–135 页。