arXiv:2303.02396v1 [cs.SD] 4 Mar 2023

用于学习环境声音程序化音频模型的通用框架

Danzel Serrano 和 *Mark Cartwright*

美国新泽西州纽瓦克市新泽西理工学院ds867@njit.edu

摘要

本文提出程序化（音频）变分自编码器（*ProVE*）框架，作为一种学习环境声音程序化音频（**PA**）模型的通用方法。该框架在提升合成声音真实感的同时，通过可调参数保留对生成声音的控制能力。框架包含两个阶段：（*i*）音频类别表征，通过训练音频自编码器定义潜在表征空间；以及（*ii*）控制映射，学习一个联合函数，该函数以从音频中提取的静态/时间控制变量和均匀噪声的随机样本为输入，取代音频编码器。我们以不同表面上的脚步声效果为例，展示**ProVE**的应用。结果表明，根据 Fr´echet 音频距离（FAD）、最大均值差异（MMD）和主观评估衡量，**ProVE** 模型在声音保真度方面优于传统**PA**模型和基于对抗的方法，因此可用于声音设计工作流程。

索引词— 程序化音频、神经音频合成、可控生成模型

1. 引言

声音设计师的工作是为电影、电子游戏和虚拟环境等沉浸式媒介录制、合成和处理音频刺激，使其与视觉或其他刺激相辅相成。目前的典型做法是遍历庞大的音频数据库，寻找某一声音类别的录制声音集合 C ，并对其进行预处理以满足项目需求 [1]。这一过程耗时较长；如果找不到或无法处理合适的声音，还可能导致真实感不足。存储限制也可能成为问题，因为对于变化多样的声音类别，需要大量不同的录音；在电子游戏和虚拟环境等交互式媒介中尤其如此，因为声音事件会受到用户所处情境的影响。

为缓解上述问题，程序化音频（**物理动画**）是传统声音设计的一种不同方法其中，声音类别被建模为一个数学过程 [2]。由此得到的模型能够减少存储限制，具备更强的可控性和表现力，并能够创造独特的听觉体验。**物理动画** 模型不同于物理建模合成 [3],，后者通过计算物理模拟来生成声音。这些基于物理的模型能够产生逼真且富有动态变化的声音，但计算开销大，开发时还需要大量领域知识。另一方面，**物理动画**模型更加简单且计算效率更高，它们使用算法，根据静态和时间控制变量生成声音，通常还会加入随机噪声，以涵盖合成过程中的变化。尽管计算效率更高，当前经典的**物理动画** 模型合成的声音质量仍低于使用真实样本或物理建模合成所获得的声音——这也是它们尚未在声音设计中普遍应用的主要原因之一 [2, 4]。

提升声音合成质量的最先进方法是数据驱动神经音频合成，它是用于生成音频的深度学习技术的一个子集。神经音频合成已在语音合成 [5–7] 和音乐生成 [8, 9], 领域得到广泛研究，但直到最近，人们才开始探究其在特定环境声音或沉浸式媒体拟音合成中的应用 [10]。此外，这些模型的可控性在语音 [11] 和音乐 [8, 9, 12] 领域已得到充分研究，但在环境声音领域尚未如此。近期一篇论文采用生成式对抗方法 [13, 14] ，对不同表面材料上的脚步声这一环境声音类别进行建模 [15]。尽管该模型生成的脚步声保真度很高，且样本空间丰富，但其可控性仅限于选择表面材料，表达能力也未考虑时间上下文，例如移动速度或脚部施加于地面的作用力大小。

2. PROVE框架

2.1. 概述

物理动画 模型根据静态和时变控制变量生成声音，通常还会加入随机噪声。一个例子是 Andy 开发的经典脚步声**物理动画**模型

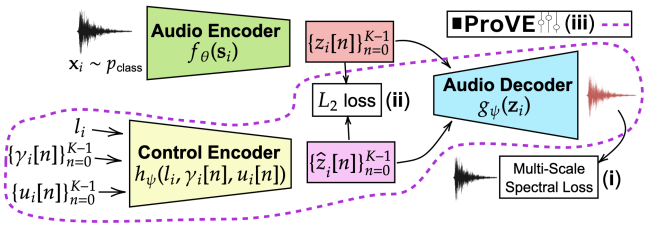


图 1。**(i)** 和 **(ii)** 展示了 **ProVE** 的两阶段框架，其中 **(i)** 在音频数据集 $\mathbf{s}_i \sim p_{class}$ 上训练类别专属自编码器，而 **(ii)** 则训练控制参数编码器，其输入包括静态条件标签、从音频 $\hat{}\boldsymbol{\gamma}_i = C\{\mathbf{s}_i\}$ 中提取的时间控制代理以及噪声序列。**(iii)** 是最终的 **ProVE** 模型，利用训练好的控制编码器和音频解码器，根据控制参数合成音频。

Farnell，见 [16]。Farnell 模型的核心是一组纹理合成模块，每种表面各有一个。静态的表面选择控制用于选择要启用的纹理合成模块，而地面反作用力（GRF）曲线生成器则通过时间控制来驱动当前启用的纹理合成模块，将均匀噪声转换为脚步声。Farnell 通过分析双足运动开发了 GRF 曲线生成器。该生成器是由 3 个多项式分段构成的函数，其数值每隔固定时间重复一次。这 3 个多项式分段描述了人类迈步各阶段的地面反作用力，即脚跟着地、从脚跟滚动至前脚掌，以及前脚掌着地，从而生成可解释的合成控制信号（见图 2）。

我们提出了**Procedural**（音频）**V**ariational autoEncoder（**ProVE**），这是一种用于学习**PA** 模型的两阶段分析-合成方法，旨在提高保真度，同时保留可控性和表现力。这种方法采用了由可微数字信号处理（**DDSP**）层组成的神经音频合成解码器 [17]。更具体地说，我们使用可微分滤波噪声模块合成一个环境声音类别，该模块由通过分析获得的控制信号控制；不同于在频谱域预测傅里叶系数的模型 [5,18], 、自回归时域方法 [19,20],以及前文提到的对抗式方法，因为后者通常难以稳定训练 [21]。

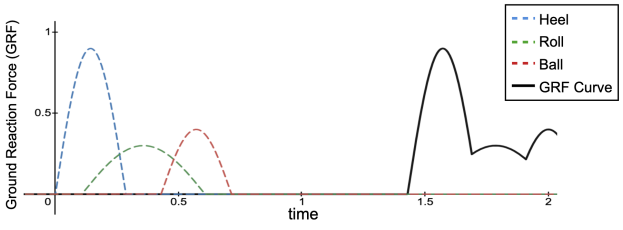


图 2。Farnell 的 GRF 曲线生成器生成的示例曲线[16]。

鉴于环境声音具有高度可变的特性，**ProVE** 旨在学习一个受控随机过程。在这里，$s[n] = \{s_n[m]\}_{m=0}^{\delta-1}$ 是一个长度为 δ的离散音频小窗口，采样自：

$$s[n] \sim q(l, \gamma[n], u[n]) \tag{1}$$

其中，q 是采样随机噪声（$u[n] \sim \mathcal{U}[-1,1]^{n_\gamma}$）的函数，取决于静态（$l \in \mathbb{R}^{n_l}$）和时间（$\gamma[n] \in \mathbb{R}^{n_\gamma}$）确定性控制。令 p_{class} 表示某个声音类别中长度为L的离散信号先验。通过重叠并相加一系列采样的小窗口$\mathbf{s} = \text{OLA}(\{s[n]\}_{n=0}^{K-1})$ 来采样离散信号，其中 $K \geq \frac{L}{\delta}$。理想情况下，我们希望 $\mathbf{s} \sim p_{class}$。这与直接根据静态条件标签输出固定长度 L 采样数组的模型不同，例如对抗方法；这种方式使我们能够通过 l 和 $\boldsymbol{\gamma} = \{\gamma[n]\}_{n=0}^{K-1}$ 定义自己的控制实例。

ProVE 框架包含两个步骤：第一步是学习一个潜在音频表示空间，并将其视为“机器可解释”的控制空间；第二步是学习从“人类可解释”的控制（即带有表面选择的 GRF 曲线）到相应“机器可解释”控制的映射。我们的训练数据集以 i 为索引，不区分声音类别，格式为 4 元组：$(\mathbf{s}_i,\ l_i,\ \hat{}\boldsymbol{\gamma}_i,\ \mathbf{u}_i)$。$\mathbf{s}_i$ 是来自目标声音类别（$\mathbf{s}_i \sim p_{class}$）的离散时间音频信号，长度为 L；l_i 是与$\mathbf{s}_i$ 关联的静态条件标签；$\hat{}\boldsymbol{\gamma}_i$ 则是通过对音频$\mathbf{s}_i$进行信号分析而得到的近似时间控制变量。也就是说，$\hat{}\boldsymbol{\gamma}_i = \{\hat{}\gamma_i[n]\}_{n=0}^{K-1}$是实际 $\boldsymbol{\gamma}_i$ 的代理变量，该变量决定音频$\mathbf{s}_i$的时间上下文，并根据 p_{class}的先验知识推导得出。最后，$\mathbf{u}_i$ 是一个假定分配给$\mathbf{s}_i$的随机噪声序列。

2.2. 步骤 1：音频类别表征

步骤 1 是训练一个自编码器模型，以学习音频$\mathbf{z}_i = \{z_i[n]\}_{n=0}^{K-1}$的潜在序列表征$\mathbf{s}_i \sim p_{class}$。由于这是一个通用框架，编码器$\mathbf{z}_i = f_\theta(\mathbf{s}_i)$和解码器 $\hat{}\mathbf{s}_i = g_\phi(\mathbf{z}_i)$模型均不受特定架构限制。不过，在第 3 节介绍的实验中，鉴于 magenta [17] 的 **DDSP**库在学习具有表现力且可控的语音 [22] 和乐器 [12], 生成模型方面表现出色，并且能够使用较小的数据集合成高质量、保真度相对较高的音频，我们在实验中采用了该库中的编码器和解码器模型。训练完成的自编码器会学习一个潜在序列空间，可以将其视为合成模块的“机器可理解”控制空间。自编码器使用 [17]中提到的多尺度频谱损失进行优化。

2.3. 步骤 2：控制映射

由于步骤 1 得到的潜在控制序列空间并非“人类可解释”，因此需要第二步来实

现可控模型。在这一步中，我们训练控制编码器 $\hat{z}_i[n] = h_\psi(l_i, \hat{\gamma}_i[n], u_i[n])$，使其根据元组 $(l_i, \hat{\boldsymbol{\gamma}}_i, \mathbf{u}_i)$，通过优化 h_ψ 上关于 L_2 的损失，来近似“机器可解释”的控制序列 $\mathbf{z}_i$ 与 $\hat{\boldsymbol{z}}_i = \{\hat{z}_i[n]\}_{n=0}^{K-1}$ 之间的关系。训练完成后，我们将控制编码器 h_ψ 与音频解码器 g_ϕ 组合，构成合成模型 $q(l, \gamma[n], u[n]) = (g_\phi \circ h_\psi)(l, \gamma[n], u[n])$。

当前 ProVE 框架的一个局限是需要通过对音频 $\hat{\boldsymbol{\gamma}}_i = C\{\mathbf{s}_i\}$ 进行某种分析过程 C，得到一个简单的控制变量代理。未来，我们计划通过针对 p_{class}，利用现有可控模型开发一种过程C，或在没有关于 p_{class} 的先验知识时，直接从$\mathbf{s}_i$中获取，以缓解这一局限。

3. 实验

本节展示如何训练用于脚步声效果的 **ProVE** 模型，并将其与 Farnell 的经典**物理动画** 模型以及 [15] 中的生成对抗网络方法进行客观和主观比较。此外，我们还提供了一个交互式 Web 应用[1]，用户可在其中输入控制变量，从预训练的**ProVE** 脚步声模型中合成声音。

在脚步声效果数据集上训练的 **ProVE** 模型具有静态控制，用于选择表面材料，与对抗方法类似。时间控制变量 $\boldsymbol{\gamma}_i$ 是 GRF 曲线，其近似值 $\hat{\boldsymbol{\gamma}}_i$ 定义为脚步声的平滑包络。也就是说，$\hat{\boldsymbol{\gamma}}_i = C\{\mathbf{s}_i\}$，其中 C 是对离散希尔伯特变换的幅度进行一维拉普拉斯平滑的结果。

选择该代理量的动机在于，施力的幅度会影响能量转换的幅度，进而影响冲击类声音的响度。控制代理量还显示，对于发生在 x_i中的每个脚步声，其具有 a3 多项式结构，如图 3 所示。

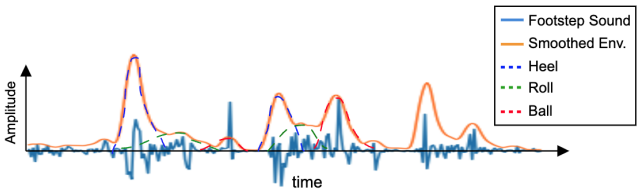


图 3。脚步声的平滑包络。

3.1. 数据集

我们在两个公开数据集上训练并评估用于脚步声的**ProVE** 模型。第一个数据集是来自 Zapsplat[2] 网站的一组高跟鞋脚步声。这与 [15], 中使用的数据集相同，我们用它将 ProVE 与 [15]中提出的对抗方法进行比较。该数据集包含单次声音（1 秒），并按地面材质分类；然而，评估部分的示例展示了一个 ProVE 模型，该模型使用同一数据集中在特定材质上行走的长段脚步声（10 秒）进行训练。之所以使用较长的音频片段而非单次声音进行训练，是因为单个脚步通常不会孤立出现，而是处于一连串具有时间结构的脚步声中。因此，为了提高物理动画模型的准确性，我们考虑能够捕捉这种时间结构的较长音频片段。双足加速的声音在感知上与双足减速不同，这是因为控制变量（GRF 曲线）及其对应的声音会随时间变化。

另一组数据集来自 *PremiumBeat.com*[3]，包含一系列长录音（~15-20 秒），并按表面材质和移动速度（步行、跑步等）分类。我们使用这组数据集来测试***ProVE*** 学习更高保真度合成的能力，同时保留经典**物理动画**的同等表现力。

3.2. 训练

为了进行训练，我们将输入音频重采样至 $f_s = 16$ kHz 采样率，以便与基线模型进行公平比较，但 DDSP 很容易扩展到更高的采样率。我们对 $\mathbf{z}_i$ 和 $(\boldsymbol{\gamma}_i, \mathbf{u}_i)$ 使用 $f_c = 250$ Hz 的控制率。用于训练和评估的主要音频编码器 $\mathbf{z}_i = f_\theta(\mathbf{s}_i)$ 是一种循环神经网络，源自提出 DDSP 的论文 [17], ，用于处理归一化的梅尔频率倒谱系数（MFCC）。随后，将前 13 个归一化 MFCC 输入门控循环单元（GRU），其输出再经过一个带双曲正切激活函数的线性层，得到潜在控制序列 $\mathbf{z}_i$，其中 $z_i[n] \in (-1,1)^{n_z}$，并选择 $n_z = 512$。音频解码器 $g_\theta(\mathbf{z}_i)$ 是一种用于参数化 DDSP减性噪声合成器的循环神经网络。控制编码器 $\hat{z}_i[n] = h_\psi(l_i, \hat{\gamma}_i[n], u_i[n])$ 与音频编码器类似：将 $(\boldsymbol{\gamma}_i, \mathbf{u}_i)$ 归一化后，与通过 l_i 获得的可学习嵌入进行广播相乘，再输入 GRU 和一个带双曲正切激活函数的线性层，从而得到 $\hat{z}_i[n] \in (-1, 1)^{n_z}$

在推理过程中，**ProVE**脚步声模型表示为 $\hat{\boldsymbol{s}} = \text{OLA}(\{\hat{s}[n]\}_{n=0}^{K-1})$，其中 $\hat{s}[n] \sim q(l, \gamma[n], u[n]) = (g_\phi \circ h_\psi)(l, \gamma[n], u[n])$。模型通过以下输入进行控制：(i) 表面材料条件标签 l；(ii) GRF曲线（例如，取自Farnell的GRF曲线生成器的样本），作为时间控制 $\boldsymbol{\gamma}$；(iii) 随机噪声**u**的样本序列。

3.3. 客观评估

我们使用数据驱动且基于感知的指标 Frech´et Audio Distance (FAD) [23] 和最大均值差异

[1]https://dependanz.github.io/projects/prove.html [2] https://www.zapsplat.com/sound-effect-packs/footsteps-in-high-heels

[3]https://www.premiumbeat.com/blog/40-free-footstep-foley-sound-effects/

（MMD）[24]，将 **ProVE** 脚步声模型的合成结果与经典的**物理动画** 模型以及近期的对抗式方法 [15]（记作 **生成对抗网络**）进行比较。这两种指标都会计算音频分布之间的距离。为了评估这些模型，我们计算模型输出、高跟鞋声训练集（**高跟鞋声**，来自 Zapsplat 网站）以及该网站上其余脚步声（**其他脚步声**）之间的 FAD 和 MMD。随着合成音频（**生成对抗网络**、**物理动画**、**ProVE**）质量的提高，其与真实音频数据集（**高跟鞋声**、**其他脚步声**）之间的分布距离应当减小。为了与**生成对抗网络**的结果进行比较，我们严格遵循了 [15]，相同的采样和评估流程：使用 Farnell 的 GRF 曲线采样，生成 12 万个一秒样本，并从 **物理动画** 和 **ProVE** 模型中进行采样。利用这两个模型生成的脚步声，我们计算 FAD 和 MMD。

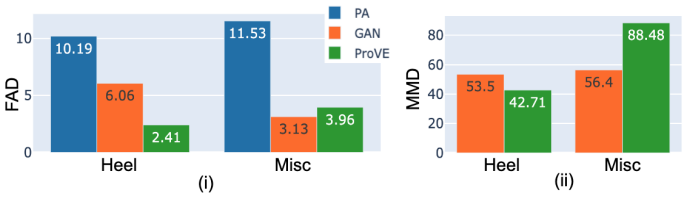


图 4. (i) 真实与合成脚步声数据集的 Frech´et 音频距离；(ii) 最大均值差异

图 4 的主要结果表明，经典**物理动画**模型与**高跟鞋声**和**其他脚步声**的差距都最大。计算 FAD 的结果还表明，**ProVE**模型比**生成对抗网络**模型更接近训练数据（**高跟鞋声**），FAD 为 2.41。然而，在两个模型与**其他脚步声**之间计算的 FAD 显示，**生成对抗网络**更接近，FAD 为 3.13，而 ProVE 模型的 FAD 为 3.96。MMD 的结果与 FAD 类似，**ProVE** 模型更接近**高跟鞋声**（42.71），而**生成对抗网络**模型的差距更大（53.5）。

这些发现表明，就保真度而言，**ProVE** 模型优于**物理动画**模型；**ProVE** 模型的 FAD 和 MMD 与两个数据集都显著更接近，进一步说明**物理动画**模型缺乏真实感。这也表明，ProVE 模型与两个数据集的距离表现一致：它与训练**高跟鞋声**数据集更接近，而与分布外数据集**其他脚步声**的距离则相应更远。然而，生成对抗网络和 ProVE 模型与其他脚步声数据集之间的 FAD 并无显著差异，因此仍可以认为 ProVE 模型也能够生成超出训练集范围的样本。

3.4. 主观评估

我们进行了主观听音测试，以比较**ProVE**模型、经典的**物理动画** 模型和训练数据集；该数据集由 PremiumBeat 网站上的长录音组成（在表2中记为 PB）。我们将这些录音片段（每段 8 秒）的平滑包络作为 GRF 曲线输入到这些录音**ProVE** 和**物理动画** 模型中。这样做是为了在听音测试中使合成样本与真实录音对齐。我们对可复现主观评估（*reseval*）框架 [25] 中的成对测试进行了修改，要求参与者选择两个脚步声中哪一个更符合描述性纹理提示。我们使用*reseval*通过亚马逊 Mechanical Turk 招募了 50 名参与者，每位参与者完成了 10 次合成分布与真实分布之间的成对比较。我们向参与者支付基本报酬，并在其完成整份问卷后给予奖金。表1和表2展示了参与者更偏好**ProVE**而非**物理动画** 或真实数据集（表2）的比例，其中 p 值来自双侧二项检验。

p 值 $= 6 \times 10^{-7}$	Dirt	草地	砾石	Wood
ProVE	**0.77**	**0.705**	0.371	**0.81**
PA	0.23	0.295	**0.629**	0.19

T表 1。主观评估结果（ProVE ，物理动画）

p 值 = 0.1	Dirt	草地	砾石	Wood
ProVE	0.5	0.486	0.431	0.387
真实 (PB)	0.5	0.514	0.569	0.613

表2。主观评估结果（ProVE，PB）

结果表明，ProVE 与经典程序音频模型在表面材料方面存在显著差异，而 ProVE 与真实数据集的比较结果则更为相似。对于泥土、草地和木材，参与者更喜欢 **ProVE** 而非经典**物理动画**模型；对于砾石，他们则更喜欢经典**物理动画**模型而非 ProVE。这可能是因为砾石中小卵石和岩石沿轨迹运动产生的细粒度声音，仅使用脚步声平滑包络的控制变量代理无法体现。

4. 结论

程序化音频变分自编码器（**ProVE**）被提出作为一个通用框架，用于学习高保真度的环境声音程序化音频模型。**ProVE**模型将控制空间从音频自编码器学习到的“人类可解释”映射为“机器可解释”，在保留受控随机过程可控性的同时，其保真度超越了经典的**物理动画** 模型。未来可能的研究方向包括缩短推理时间，以实现实时音频合成，以及使用**ProVE** 根据视频/动画自动进行拟音生成。**ProVE** 的影响不止于音频信号，还可扩展至一般的程序化信号生成，成为实现潜在空间可控性的一步。

5. 参考文献

[1] Mats Liljedahl 和 Johan Fagerl¨onn，“声音设计方法：综述及其对研究与实践的启示”，载于第五届*Audio Mostly*会议论文集：声音交互会议，美国纽约州纽约市，2010年，AM '10，美国计算机协会。

[2] Andy Farnell，《程序化音频简介及其在电脑游戏中的应用》，2007年。

[3]Julius O. Smith，物理音频信号处理，http://ccrma.stanford.edu/~jos/pasp/，访问日期：2023年2月12日，在线著作，2010年版。

[4] Niels B¨ottcher，《电脑游戏中程序化音频的当前问题与未来可能性》，《游戏与虚拟世界杂志》，第5卷，第3期，第215–234页，2013年9月。

[5] Yuxuan Wang、RJ Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J. Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio、Quoc Le、Yannis Agiomyrgiannakis、Rob Clark 和 Rif A. Saurous，《Tacotron：迈向端到端语音合成》，2017年。

[6] Jonathan Shen、Ruoming Pang、Ron J. Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、YuZhang、Yuxuan Wang、RJ Skerry-Ryan、Rif A. Saurous、Yannis Agiomyrgiannakis 和 Yonghui Wu，《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》，2017 年。

[7] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速、高质量的端到端文本转语音》，2020 年。

[8] GautamMittal、Jesse Engel、Curtis Hawthorne 和 Ian Simon，《使用扩散模型进行符号音乐生成》，2021 年。

[9]Prafulla Dhariwal、Heewoo Jun、Christine Payne、Jong Wook Kim、Alec Radford 和 Ilya Sutskever，《Jukebox：一种音乐生成模型》，2020 年。

[10] Keunwoo Choi、SangshinOh、Minsung Kang 和 Brian McFee，“拟音合成挑战提案”，2022。

[11] Meiying Chen 和 Zhiyao Duan，“Controlvc：通过音高和节奏的时变控制实现零样本语音转换”，2022。

[12] Yusong Wu、Ethan Manilow、Yi Deng、Rigel S wavely、Kyle Kastner、Tim Cooijmans、Aaron Courville、Cheng- Zhi Anna Huang 和 Jesse Engel，《Midi-DDSP：通过分层建模实现对音乐演奏的精细控制》，2021。

[13] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae，《Hifi-gan：用于高效、高保真度语音合成的生成对抗网络》，2020。

[14] Chris Donahue、Julian McAuley 和 Miller Puckette，“对抗性音频合成”，2019。

[15]Marco Comunit`a、Huy Phan 和 Joshua D. Reiss，“使用生成对抗网络合成脚步声效”，2021。

[16] Andy Farnell，“迈步向前：面向电子游戏和动画的程序化合成脚步声”，2007。

[17]Jesse Engel、Lamtharn Hantrakul、Chenjie Gu 和 Adam Roberts，“Ddsp：可微分数字信号处理”，2020。

[18]Jesse Engel、Kumar Krishna Agrawal、Shuo Chen、Ishaan Gulrajani、Chris Donahue 和 Adam Roberts，《GANSynth：对抗式神经音频合成》，2019。

[19] Nal Kalchbrenner、Erich Elsen、Karen Simonyan、Seb Noury、Norman Casagrande、Edward Lockhart、Florian Stimberg、Aaron van den Oord、Sander Dieleman 和 Koray Kavukcuoglu，《高效神经音频合成》，2018。

[20] Aaron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior 和 KorayKavukcuoglu，《WaveNet：一种用于原始音频的生成模型》，2016。

[21] Divya Saxena 和 Jiannong Cao，《生成对抗网络（GAN）：挑战、解决方案与未来方向》，2020。

[22] Giorgio Fabbro、VladimirGolkov、ThomasKemp 和 Daniel Cremers，《使用可微数字信号处理进行语音合成与控制》，2020 年。

[23] Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi，《Fr´echet 音频距离：一种评估音乐增强算法的指标》，2018 年。

[24] Arthur Gretton、Karsten M. Borgwardt、Malte J. Rasch、Bernhard Sch¨olkopf 和 Alexander Smola，《核两样本检验》，机器学习研究期刊，第 13 卷，期号 null，页码 723–773，2012 年 3 月。

[25] Max Morrison、Brian Tang、Gefei Tan 和 BryanPardo，《可复现主观评估》，发表于*ICLR* 机器学习评估标准研讨会，2022 年 4 月。