中观结构：超越可微时频分析中的频谱图损失

CYRUSVAHIDI[1],HAN HAN[2],CHANGHONGWANG[2], MATHIEU LAGRANGE[2],
GYÖRGY FAZEKAS[1],和 VINCENT LOSTANLEN[2](c.vahidi@qmul.ac.uk)

[1]伦敦玛丽女王大学数字音乐中心，英国[2]南特大学、南特中央理工学院、法国国家科学研究中心、*LS2N*，*UMR 6004*，*F-44000* 南特，法国

arXiv:2301.10183v1 [cs.SD] 24 Jan 2023

计算机音乐家将中观结构视为介于波形的微观结构与音乐形式的宏观结构之间的中间层次的组织形式。中观结构的例子包括旋律、琶音、切分音、复调分组和织体对比。尽管中观结构在音乐表达中发挥着核心作用，但深度学习领域对其关注有限。目前，自编码器和神经音频合成器仅在微观结构的尺度上进行训练和评估：即约 100 毫秒以内的局部振幅变化。本文通过组合可微琶音器与时频散射，提出并解决中观结构音频建模问题。我们通过实验证明，时频散射可作为一种可微模型，用于衡量控制中观结构的合成参数之间的相似度。通过揭示短时谱距离对时间对齐的敏感度，我们说明了为何需要一种时移不变、多尺度的可微时频相似度模型，以同时描述局部频谱和时频调制。

0 引言

0.1 可微时频分析

时频表示（TFR），例如短时傅里叶变换（STFT）或常数Q变换（CQT），在音乐信号处理中发挥着关键作用 [1, 2]，因为它们能够解调缓慢变化的复音调的相位。因此，任何两个时频表示幅值（即频谱图）相同的声音 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，在人类听者听来都相同，即使其底层波形可能不同。因此，频谱图不仅可用于可视化，也可用于相似性检索。将频谱图算子记为 $\boldsymbol{\Phi}$，则欧几里得距离 $\|\boldsymbol{\Phi}(\boldsymbol{y}) - \boldsymbol{\Phi}(\boldsymbol{x})\|_2$ 比波形距离 $\|\boldsymbol{y} - \boldsymbol{x}\|_2$ 更有信息量，因为即使相位差很小，波形距离也会迅速增大。

近年来，已有的 STFT 和 CQT 算法已被移植到 PyTorch、TensorFlow、MXNet 和 JAX 等深度学习框架中 [3, 4, 5]。为此，开发者利用了可微编程这一范式，其定义是通过反向模式自动微分计算数学函数梯度的能力。在音频处理领域，可微编程可用于训练神经网络，以实现音频编码、解码或两者兼有。因此，我们可以创造一个统称：可微时频分析 （DTFA），用来描述深度学习中一个新兴的子领域。在该领域中，随机梯度下降涉及神经网络层与时频表示的组合。过去，时频表示主要局限于分析前端，如今则已成为音频生成学习架构中不可或缺的一部分。

可微时频分析的最简单示例是自编码。给定输入波形 $\boldsymbol{x}$，自编码器是一种权重为$\mathbf{W}$的神经网络架构 f，它会返回另一个波形 $\boldsymbol{y}$ [6, 7]。在训练过程中，神经网络 $f_{\boldsymbol{W}}$ 旨在最小化以下损失函数：

$$\mathcal{L}_{\boldsymbol{x}}(\mathbf{W}) = \|(\boldsymbol{\Phi} \circ f_{\boldsymbol{W}})(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x})\|_2, \quad (1)$$

对无标签数据集中的每个样本 $\boldsymbol{x}$ 取平均值。上面的函数称为频谱图损失，因为$\boldsymbol{\Phi}$将 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 映射到时频域。

可微时频分析的另一个例子见于音频修复。此时，$f_{\boldsymbol{W}}$ 的输入不是 $\boldsymbol{x}$ 本身，而是其某种退化版本$\boldsymbol{h}(\boldsymbol{x})$——例如含噪声或带限的版本 [8, 9]。 $f_{\boldsymbol{W}}$ 的目标是通过生成在频谱图损失方面与$\boldsymbol{x}$ 接近的修复后声音（$f_{\boldsymbol{W}} \circ \boldsymbol{h}$）$(\boldsymbol{x})$，来逆转退化算子 $\boldsymbol{h}$：

$$\mathcal{L}_{\boldsymbol{x}}(\mathbf{W}) = \|(\boldsymbol{\Phi} \circ f_{\boldsymbol{W}} \circ h)(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x})\|_2. \quad (2)$$

第三，可微时频分析可用于声音匹配，也称为合成器参数反演 [6, 10, 11]。给定

一个参数合成器 g 和一个音频查询 x，这项任务旨在找出参数设置 θ ，使得$y = g(\theta)$与 x相似。实践中，可以通过对 θ进行随机采样，生成$x = g(\theta)$，并测量x与 y之间的频谱图损失，基于合成数据训练声音匹配：

$$\mathcal{L}_\theta(\mathbf{W}) = \|(\mathbf{\Phi} \circ g \circ f_W \circ g)(\theta) - (\mathbf{\Phi} \circ g)(\theta)\|_2. \quad (3)$$

0.2 频谱图损失的局限

尽管频谱图损失在生成式音频建模中已被证明颇有价值，但当事件在时间或音高上未对齐时，它会表现出违背直觉的特性 [12]。较低的频谱图距离意味着感知相似度较高，但反过来却不成立：有些例子中，$\mathbf{\Phi}(x)$与 $\mathbf{\Phi}(y)$相差甚远，人类听众却认为它们在音乐上相似。首先， $\mathbf{\Phi}$ 仅对频谱图窗口 T 尺度内的时间偏移敏感，也就是约 10–100 毫秒。在自编码的情况下，如果 $f_W(x)(t) = x(t-\tau)$与 $\tau \gg T$之间存在差异，那么 $L_x(\mathbf{W})$ 可能会很大，即使 f_W 的输出很容易通过互相关重新对齐到 x 上。在有音高声音的音频修复中，听众对起音附近的伪影（例如预回声）更为敏感 [13], ，尽管频谱能量大多集中在时间轮廓的延音和释音部分。

最后，在声音匹配的情况下，某些合成器包含控制较大时间尺度上周期结构的参数，而这些参数独立于局部频谱变化。在加法合成中，颤音、震音或波音等周期调制技术都有一个“速率”参数，它既无法从孤立的频谱图帧中预测，也无法简化为一系列离散声音事件。对 ε 的合成参数进行微小扰动，会导致 $g(\theta + \varepsilon)$ 在全局上扩展或压缩、但在局部时间上错位，从而使 $\|(\mathbf{\Phi} \circ g)(\theta + \varepsilon) - (\mathbf{\Phi} \circ g)(\theta)\|$ 无法反映 ε 的大小。模块化合成器通过控制模块（音序器、函数发生器）与声音处理和生成模块（振荡器、滤波器、波形整形器）之间的交互来塑造声音 [14]。在一个“音色补丁”中，音序器决定播放速度并触发事件，而振幅包络、振荡器波形和滤波器则塑造音色。改变音色补丁的时钟速度会导致事件在时间上错位，但不会改变孤立事件的频谱构成。因此，无法再在孤立频谱图帧的时间尺度上比较音色相似度。

0.3 音乐时间尺度：微观、中观、宏观

仅在短时频谱的微观尺度上建模音乐相似度的不足，可从算法作曲中使用的音乐结构术语窥见一斑。计算机音乐家会按照时间尺度的层级来指称音乐结构。层级的一端是微观尺度，涵盖从由少量采样点构成的声音粒子，到短时频谱分析所涉及的毫秒级时间尺度 [15]。沿着时间层级再往上是中观尺度；这一层级的结构由声音对象的组合及其复杂的频谱—时间演变而产生 [16]。而宏观尺度 则大体涵盖整首乐曲或整场演出的编排。Curtis Roads 指出了在算法作曲中对多尺度结构进行连贯建模所面临的挑战[17]。在颗粒合成中，微观结构由单个音粒构成，而它们的播放速率则形成中观结构层级上的纹理云。超越微观尺度和频谱图分析，还存在由复杂的频谱与时间包络涌现出的声音结构，例如声音纹理和乐器演奏技法 [18]。

0.4 贡献

本文为中观结构的可微时频分析铺平了道路。核心思路是在声音 x 的时频域中计算二维小波分解（“散射”）。所得结果称为联合时频散射变换（JTFS），对音乐事件之间的相对时间滞后和频率间隔敏感。同时，JTFS 对全局时间平移保持稳定：回到自编码的例子，$f_W(x)(t) = x(t-\tau)$ 会得到 $(\mathbf{\Phi} \circ f_W)(x) \approx \mathbf{\Phi}(x)$，这与人类感知相符。

为了展示 JTFS 在可微时频分析中的潜力，我们给出一个可微声音匹配的例子，其中微观尺度距离并不能很好地反映参数距离。在这个例子中，目标声音 $x = g(\theta)$ 是由短滑音事件（“啁啾小波”）组成的琶音，其跨度为两个八度。问题中的两个未知量分别是单位时间内啁啾小波的数量和琶音的总时长。我们展示了，无需任何特征工程，只需在联合时频散射变换空间中构造如下形式的最小二乘逆问题，就能求出这两个未知量：

$$\begin{aligned}\theta^* &= \arg\min_{\widetilde{\theta}} \mathcal{L}_\theta(\widetilde{\theta}) \\ &= \arg\min_{\widetilde{\theta}} \|(\mathbf{\Phi} \circ g)(\widetilde{\theta}) - (\mathbf{\Phi} \circ g)(\theta)\|_2^2 \end{aligned} \quad (4)$$

直观而言，要使上述逆问题能够通过梯度下降求解，在任意初始猜测 θ 处计算时，L_θ 的梯度都应指向 $\sim\theta$。我们的主要发现是，当 $\mathbf{\Phi}$ 为 JTFS 时确实如此，但当 $\mathbf{\Phi}$ 为多尺度频谱图（MSS）时则不然。此外，我们发现，即使目标声音存在数百毫秒的随机时间滞后，L_θ 的梯度仍然具有指导意义。为了解释这一差异，我们将可微中尺度结构算子 定义为在 θ 处，$(\mathbf{\Phi} \circ x)$ 的雅可比矩阵，即在所关注的参数设置下，音频合成与 JTFS 分析的乐曲。这个概念不仅适用于声音匹配，在训练用于自编码和音频修复的神经网络时也有对应的概念。

我们在 Kymatio v0.4[1]，中释音了 JTFS 的可微实现；Kymatio 是一款面向图形处理器上可微时频分析的开源软件，可与现代深度学习库互操作

[1]Kymatio v0.4：https://github.com/kymatio/kymatio

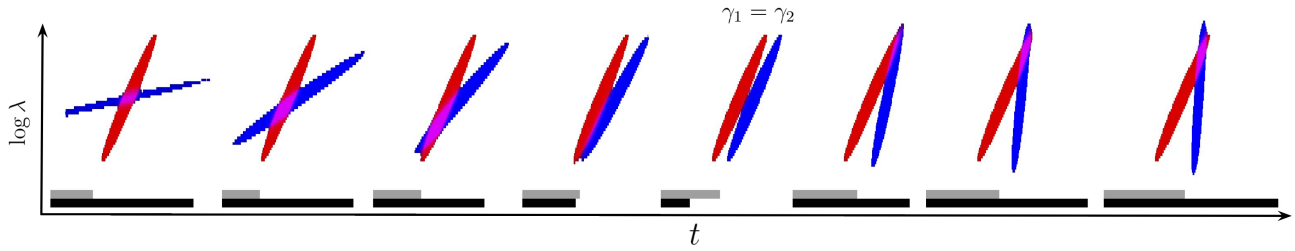

图1. 啁啾信号在时间和对数频率上重叠的示意图。红色啁啾信号的啁啾率均为 γ。蓝色啁啾信号相对于红色信号在时间上发生偏移，且其 γ 逐渐增大（从左到右）。这些线段分别表示两个啁啾信号在多尺度频谱图（灰色）和时频散射（黑色）域中的距离。我们观察到，当决定中观结构 的啁啾率 γ 相同时，JTFS 距离达到最小值，而频谱图距离则接近最大值。JTFS 距离与 γ 中的距离具有良好的相关性。我们将在第3节更详细地讨论时移不变的可微中尺度结构算子的重要性。

[19]。为鼓励数值实验的可复现性，我们还为本文提供了开源代码[2]。

1 动机示例

1.1 比较存在时间延迟的啁啾信号

图（1）展示了在可微时频分析中，如何可靠地计算由 $\boldsymbol{g}$ 合成的啁啾信号之间相似度这一挑战。在该示例中，无论调频率如何，两个啁啾信号在时频域中的一阶矩都相等。考虑两个在时间上彼此错开的啁啾信号。当中观结构相同时，也就是说调频率相等且两个信号互不重叠时，它们的频谱图距离达到最大值。随着调频率增大，两个啁啾信号在时频域中发生重叠，导致频谱图距离减小；这种变化并不能正确反映 $\boldsymbol{\theta}$ 的变化。随着 γ 变化，频谱图损失几乎没有变化。此外，在 γ 下，啁啾信号的局部微观片段会在时间和频率上周期性偏移，这意味着比较微观结构不足以衡量相似度。一种可能的解决方案是动态重新对齐啁啾信号，但这种操作在数值上不稳定且不可微。接下来的章节中，我们将介绍一种能够对 $\boldsymbol{\theta}$ 中的距离进行建模且对时间平移保持稳定的可微算子。能够区分这三个变化因素的表示，应当能够提供声学空间中的邻域距离度量，使其反映参数空间中的距离。

1.2 啁啾小波合成器

啁啾小波是一种短暂的声音事件，会在时频平面上产生一条对角线。一般而言，啁啾小波遵循以下形式的方程：$\boldsymbol{x}(t) = \boldsymbol{a}(t)\cos(2\pi\boldsymbol{\varphi}(t))$，其中 $\boldsymbol{a}$ 和 $\boldsymbol{\varphi}$ 分别表示瞬时振幅和相位。本文生成的啁啾小波，其瞬时频率随时间呈指数增长，因此其感知音高（大致与对数频率成正比）呈线性增长。我们用啁啾率 γ来表征这种频率调制（FM），单位为每秒八度。令 f_c 表示啁啾小波起音时的瞬时频率，可得：

$$\boldsymbol{\varphi}(t) = \frac{f_c}{\gamma \log 2} 2^{\gamma t}. \tag{5}$$

接下来，我们将啁啾小波的瞬时振幅 $\boldsymbol{a}$ 定义为正弦函数在时间支撑 δ^t上的半周期。我们用调幅（AM）频率 $f_m = \frac{1}{2}\delta^t$来参数化这个半周期。因此：

$$\boldsymbol{a}(t) = \sin(2\pi f_m t) \text{ if } 0 \leq f_m t < \tfrac{1}{2} \text{ and } 0 \text{ otherwise.} \tag{6}$$

在其终止处，啁啾小波的瞬时频率等于 $f_m = f_c 2^{\gamma\delta^t} = f_m 2^{\gamma/f_m}$。我们用记号 $\boldsymbol{\theta}$作为调幅/调频元组（f_m，γ）的简写。

1.3 可微琶音器

现在，我们定义一个递升的“琶音”，使前一个事件的结束时间与下一个事件在时频域中的起音时间重合。为此，我们将啁啾小波在时间上平移 $n\delta^t$，并将其相位乘以 $2^{n\delta^t} = 2^{n\gamma\delta^t}$，其中 n 为整数。最后，我们通过宽度为 γw 的高斯窗（$t \mapsto \boldsymbol{\phi}_w(\gamma t)/\gamma$）为琶音施加一个全局时间包络，其中带宽参数 w 以八度表示。因此：

$$\begin{aligned}\boldsymbol{x}(t) &= \frac{1}{\gamma}\boldsymbol{\phi}_w(\gamma t) \sum_{n=-\infty}^{+\infty} \boldsymbol{a}\left(t - \frac{n}{f_m}\right) \cos\left(2^{\gamma \frac{n}{f_m}} \boldsymbol{\varphi}\left(t - \frac{n}{f_m}\right)\right) \\ &= \boldsymbol{g}_{\boldsymbol{\theta}}(t), \text{ where } \boldsymbol{\theta} = (f_m, \gamma).\end{aligned} \tag{7}$$

在上述公式中，非零能量事件的数量与以下因素成正比：

$$\nu(\boldsymbol{\theta}) = \frac{f_m w}{\gamma}, \tag{8}$$

由于它随 $\boldsymbol{\theta}$ 连续变化，因此不一定是整数。这里可以看到，我们的参数模型 $\boldsymbol{g}$ 尽管十分简单，却能控制一种只有在中观尺度上才有意义的听觉感知：即构成序列流的琶音中的音符数量 ν。此外，这一数量源于调幅（f_m）与调频（γ）的交织，并且在时间平移（将 t 替换为（$t - \tau$））或频率移调（改变 f_c）后仍保持不变。因此，尽管这个可微琶音器的灵活性有限——

[2]实验代码库：https://github.com/cyrusvahidi/meso-dtfa

我们认为，它为中观结构的可微时频分析提供了一个富有启发性的试验平台。

2 时频散射

联合时频散射（JTFS）是一种时频域中的卷积算子 [20]。JTFS 通过在时频域中以不同尺度和速率应用二维小波滤波器，从数字音频中提取多尺度时频调制。作为二维卷积神经网络的前端使用时，JTFS 只需有限的标注训练数据，即可实现最先进的乐器分类性能 [21]。Florian Hecker 的乐曲，例如 2016 年的 *FAVN*，展现了 JTFS 在计算机音乐再合成方面的能力（乐曲完整列表见 [22]）。

2.1 小波标度图

令 $\psi \in \mathbf{L}^2(\mathbb{R},\mathbb{C})$ 为一个中心频率为 1、带宽为 $1/Q_1$ 的复值零均值小波滤波器。我们定义一个常数-Q 滤波器组，其膨胀因子取自 ψ，形式为 ψ_λ：$t \longmapsto \lambda\psi(\lambda t)$，品质因数为常数 Q_1。每个小波的中心频率为 λ，带宽为 λ/Q_1。我们将频率变量 λ 按公比为 2^{Q_1} 的几何级数离散化，起始值为 λ/Q_1。当品质因数为 $Q_1 = 1$ 时，后续小波的中心频率间隔为一个八度，即构成二进小波滤波器组。

将滤波器组 ψ 与波形 $x \in \mathbf{L}^2(\mathbb{R})$ 卷积，并逐点应用复模，即可得到小波标度图 $\mathbf{U}_1$：

$$\mathbf{U}_1 x(t,\lambda) = |x * \psi_\lambda|(t) \tag{9}$$

$\mathbf{U}_1$ 以时间和对数频率为索引，对应于时频分析中常用的常数Q变换。

2.2 时频小波

与第 2.1 节类似，我们沿时间轴和对数频率轴定义另外两个小波ψ^{t} 和 ψ^{f}，其品质因数分别等同于 Q_2 和 Q_{fr}。然后

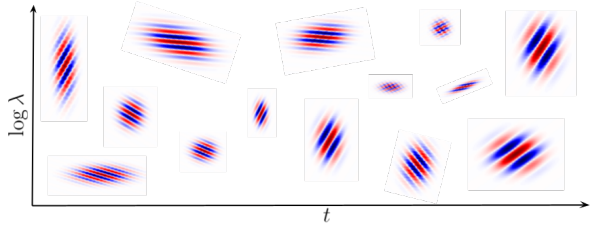

图 2. 二维时频小波（第二阶 JTFS）形状示意图。红色和蓝色分别表示较高的正值和较低的负值。每种图案均展示二维滤波器实部的响应；这些滤波器由不同速率 α和尺度 β（分别对应）的 1D 小波 $\psi_\alpha(t)$ 与 $\psi_\beta(\log\lambda)$ 的外积构成。方向由 β 的符号决定，该变量也称为自旋 变量，取值范围为 $\{-1,1\}$。有关 JTFS 的详情，请参阅第 2 节。

推导出两个滤波器组 ψ^{t}_α 和 ψ^{f}_β，其中心频率分别为 α 和 β，其中

$$\psi^{\mathrm{t}}_\alpha(t) = \alpha\psi^{\mathrm{t}}(\alpha t) \tag{10}$$

$$\psi^{\mathrm{f}}_\beta(\log_2\lambda) = \beta\psi^{\mathrm{f}}(\beta\log_2\lambda) \tag{11}$$

与计算$\mathbf{U}_1$, 时一样，我们将 α 和 β 离散化为$_1$ $_1$ 等比数列，其公比分别为 2^{Q_2}和 $2^{Q_{\mathrm{fr}}}$。我们从听觉谱时感受野的角度来解释频率变量 α和 β [23]：α 是以赫兹为单位测量的时间调制速率，而 β是以每倍频程周期数为单位测量的频率调制尺度。

ψ^{t}_α 与 ψ^{f}_β 的外积形成一组具有不同速率 α 和尺度 β的二维小波。我们依次将ψ^{t}_α 和 ψ^{f}_β 与$\mathbf{U}_1 x$ 进行卷积，并逐点计算复模，得到一个以（t、 λ、 α、 β）为索引的四维张量：

$$\mathbf{U}_2 x(t,\lambda,\alpha,\beta) = |\mathbf{U}_1 * \psi^{\mathrm{t}}_\alpha * \psi^{\mathrm{f}}_\beta| \tag{12}$$

图 2 展示了时频域中二维小波滤波器的实部。这些小波的速率为 α、尺度为 β，并沿对数$_2$ λ轴呈向上或向下的方向，可捕捉时间和频率中的多尺度振荡模式。

2.3 局部平均

我们将式 (9) 中的小波标度图$\mathbf{U}_1 x$ 与宽度为 T的高斯低通滤波器 ϕ_T 进行卷积，再沿对数频率轴与 ψ_β ($\beta \geq 0$)卷积，随后逐点计算复模，从而计算一阶联合时频散射系数：

$$\mathbf{S}_1 x(t,\lambda,\alpha=0,\beta) = |\mathbf{U}_1 x(t,\lambda) * \phi_T * \psi_\beta| \tag{13}$$

在与 ψ_β卷积之前，我们沿时间对$\mathbf{U}_1 x(t,\ \lambda) * \phi_T$ 的输出进行下采样，得到与 $1/T$成比例的采样率。事实上，式 (13) 是式 (12) 的一个特例，其中通过使用 ϕ_T来调制速率 $\alpha = 0$。

我们定义二阶联合时频散射变换 x 如下：

$$\mathbf{S}_2 x(t,\lambda,\alpha,\beta) = \mathbf{U}_2 x(t,\lambda) * \phi_T * \phi_F \tag{14}$$

其中， ϕ_F 是沿对数频率维度、宽度为 F的高斯低通滤波器。对于式（12）中特殊情况$\beta = 0$， ψ_β 发挥 ϕ_F的作用，得到：

$$\mathbf{S}_2 x(t,\lambda,\alpha,\beta=0) = |\mathbf{U}_1 x(t,\lambda) * \psi^{\mathrm{t}}_\alpha * \phi_F| * \phi_T \tag{15}$$

在式（14）和（15）中，我们分别沿时间轴和对数频率轴对$\mathbf{S}_2 x$ 进行降采样，使其采样率分别为 T^{-1} 和 F^{-1}。使用 ϕ_T 和 ϕ_F 进行低通滤波，可分别获得对时间平移和频率移调的不变性，其尺度分别为 T 和 F。组合$\mathbf{S}_1 x$与$\mathbf{S}_2 x$，即$\mathbf{S}x = \{\mathbf{S}_1 x,\ \mathbf{S}_2 x\}$，使我们能够涵盖组合变量（$\lambda$， α， β）的所有路径。在第3节中，我们将介绍如何将$\mathbf{S}x$ 用作处理中观结构的可微时频分析算子。

在图 1 中，我们强调了需要一种能够建模中观结构的算子。啁啾小波流以特定速率发生频移。在二阶层面，JTFS 描述了 $\mathbf{S}_1$ 未能捕捉到的较大尺度时频结构。此外，JTFS 具有时移不变性，

因此，在时间尺度 T范围内，它是衡量介观结构相似性的可靠指标。

3 可微中尺度结构算子

本节中，我们介绍一种用于时频分析的可微中观结构算子。这种算子适用于需要可微相似度度量的优化场景，例如自编码。

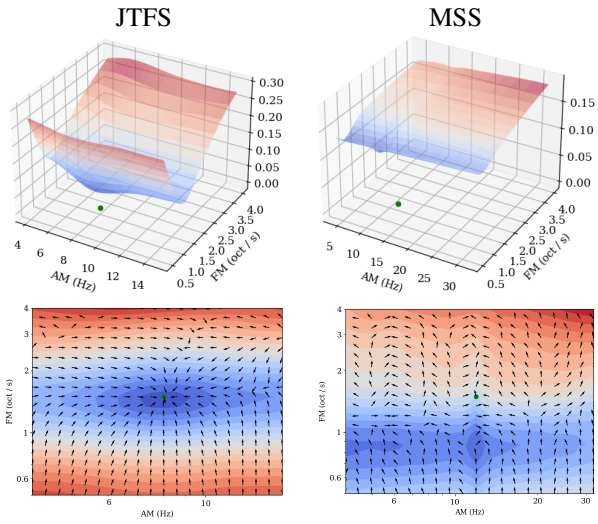

图 3. 在 $\mathbf{\Phi}$ 作为 JTFS（左）和 MSS（右）的情况下，对由 $\boldsymbol{g}$ 合成的声音进行损失曲面和梯度场可视化（见第1节）。声音取自 f_{m} 和 γ上的对数间隔网格。我们将目标声音绘制为绿色圆点，并计算目标与网格上每个点生成的声音之间的损失。将生成的声音相对于目标声音进行恒定 $\tau = 2^{10}$ 个采样点的时间偏移。在矢量场图中，我们计算损失算子相对于合成参数 f_{m} 和 γ的梯度。箭头的方向表明，基于 $\mathbf{\Phi}\circ\ \boldsymbol{g}$ 计算的距离在多大程度上能提供关于 $\boldsymbol{\theta}$的信息。对于 $\mathbf{\Phi}_{\mathrm{JTFS}}$，我们观察到一个三维损失曲面，其全局最小值位于目标声音附近，且梯度指向目标。相反， $\mathbf{\Phi}_{\mathrm{MSS}}$ 的全局最小值既不位于目标附近，也不为 0。在存在小幅时间偏移的情况下，MSS 损失似乎对调幅差异不敏感，也无法提供关于 $\boldsymbol{\theta}$的有效信息。

在第1节中，我们定义了一个可微琶音器 $\boldsymbol{g}$，其参数 $\boldsymbol{\theta}$ 控制 $\boldsymbol{x}$ 中的中观结构 。现在，我们寻求一个可微算子 $\mathbf{\Phi}\circ\ \boldsymbol{g}$，以此建立一个模型来控制低维参数空间 $\boldsymbol{\theta}$。通过在 $\mathbf{\Phi}\circ\ \boldsymbol{g}$ 下对距离和梯度进行可视化，我们着手评估 $\mathbf{\Phi}$ 是否适合在声音匹配任务中对 $\boldsymbol{\theta}$ 建模。

我们考虑两种用于 $\mathbf{\Phi}$ 的 DTFA 算子：(i) 多尺度频谱图（MSS）（近似为 $\mathbf{U}_1\boldsymbol{x}$）和 (ii) 时频散射（$\mathbf{S}\boldsymbol{x} = \{\mathbf{S}_1\boldsymbol{x},\ \mathbf{S}_2\boldsymbol{x}\}$）（JTFS）。在情况 (i) 中，我们认为两个声音之间距离较小，意味着它们具有相同的微观结构。相反，在情况 (ii) 中，相似度则表明它们具有相同的中观结构。虽然相同的 $\mathbf{U}_1$ 意味着中观结构相同，但反之并不成立，例如存在时间平移和非平稳频率时。

此前，JTFS 已用于评估乐器演奏技巧之间的相似度，而这些技巧构成了中观结构。借助 DTFA 算子 $\mathbf{\Phi}$，我们有望根据原始音频波形、合成参数或神经网络权重所体现的相似度来建模中观结构。在颗粒合成等情况下，可能需要对中观结构进行控制，同时允许微观结构随机变化。

3.1 梯度计算与可视化

我们在算子$\mathbf{\Phi}\circ\ \boldsymbol{g}$下评估距离目标函数，作为 $\boldsymbol{\theta}$中距离的代理：

$$\mathcal{L}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}) = \|(\mathbf{\Phi}\circ\boldsymbol{g})(\boldsymbol{\theta}) - (\mathbf{\Phi}\circ\boldsymbol{g})(\tilde{\boldsymbol{\theta}})\|_2^2 \tag{16}$$

˜对于给定的参数估计值 $\boldsymbol{\theta}$，到目标 $\boldsymbol{\theta}$ 的距离的梯度 $\nabla \mathrm{L}_{\boldsymbol{\theta}}$ 为：

$$\nabla\mathcal{L}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}) = -2\Big((\mathbf{\Phi}\circ\boldsymbol{g})(\boldsymbol{\theta}) - (\mathbf{\Phi}\circ\boldsymbol{g})(\tilde{\boldsymbol{\theta}})\Big)^T \cdot \nabla(\mathbf{\Phi}\circ\boldsymbol{g})(\tilde{\boldsymbol{\theta}}) \tag{17}$$

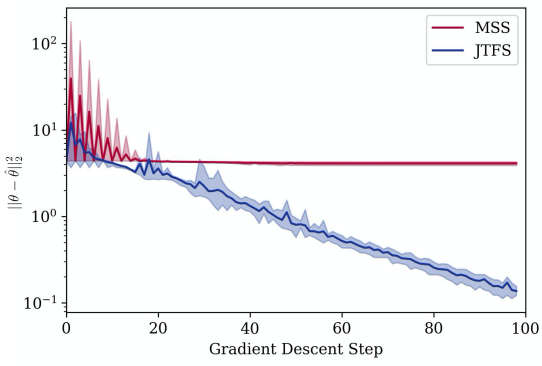

图 4. 梯度下降迭代过程中参数距离 $||\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}||_2$，分别使用 $\mathbf{\Phi}$ 作为 MSS 和 JTFS。目标声音的参数为 $\boldsymbol{\theta} = [8。49，1。49]$。我们将预测声音初始化为$\widetilde{\boldsymbol{\theta}}_0 = [4,0。5]$。折线图展示了多次运行中每次迭代的平均距离；这些运行分别将预测采样点在时间上偏移 $\tau = \{2^2，2^4，2^7，2^{10}\}$ 个采样点。阴影区域表示不同时间偏移下的范围。

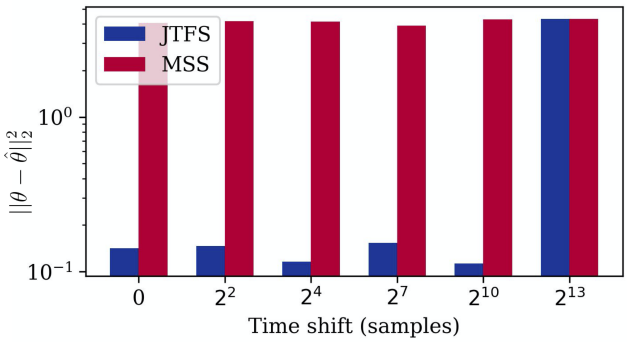

图 5. 对 $\boldsymbol{g}(\boldsymbol{\theta})(t)$和 $\boldsymbol{g}(\widetilde{\boldsymbol{\theta}})(t-\tau)$，在 $\boldsymbol{\theta} = [8.49,1.49]$, $\widetilde{\boldsymbol{\theta}}_0 = [4,0.5]$条件下，经过梯度下降后得到的最终参数距离 $||\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}||_2$ 。每次运行（横轴）都针对预测音频上的不同时间偏移 τ 进行优化。JTFS 在其低通滤波器支撑范围 $T = 2^{13}$ 内具有不变性。我们观察到，当 $\mathbf{\Phi}$ 为 JTFS 时，在可微中尺度结构算子$\mathbf{\Phi}\circ\ \boldsymbol{g}$下，参数恢复的收敛对时间偏移保持稳定。当 $\mathbf{\Phi}$ 为频谱图算子时，优化过程则不稳定。

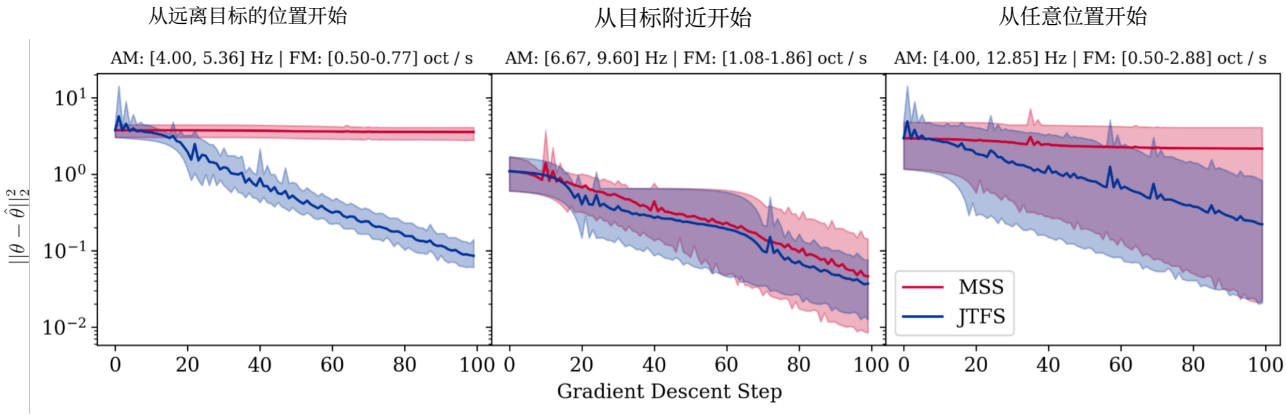

图 6. 梯度下降迭代过程中的参数距离 $||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||$（对数刻度），以 $\boldsymbol{\Phi}$ 作为 MSS 和 JTFS，在 3 种情形下进行比较：（左）5 个初始化位置 $\tilde{\boldsymbol{\theta}}$ 远离目标；（中）5 个初始化位置 $\tilde{\boldsymbol{\theta}}$ 位于目标邻域内；（右）5 个初始化位置 $\tilde{\boldsymbol{\theta}}$ 分布在网格范围内。我们不对预测声音应用时间偏移（梯度可视化见图 3）。目标声音的参数为 $\boldsymbol{\theta} = [8。49，1。49]$。曲线表示 5 次不同 $\tilde{\boldsymbol{\theta}}$ 初始化运行中每次迭代的平均距离。阴影区域表示 5 次初始化的范围。标题表示初始 $\tilde{\boldsymbol{\theta}}$ 的范围。需要强调的是，即使没有时间平移，MSS 也只有在 $\tilde{\boldsymbol{\theta}}$ 初始化于其局部邻域内时，才能很好地恢复 $\boldsymbol{\theta}$（中）。当 $\tilde{\boldsymbol{\theta}}$ 初始化位置远离目标时（左），MSS 无法收敛。从任意位置开始（右），在最佳情况下能够收敛，但平均而言无法收敛，结果接近最差情况。

公式（17）中的第一项是一个长度为$P = \dim\left((\boldsymbol{\Phi} \circ \boldsymbol{g})(\boldsymbol{\theta})\right)$的行向量，第二项是一个维度为 $P \times \dim(\tilde{\boldsymbol{\theta}})$ 的矩阵。第一项中的行向量与高维雅可比矩阵 $\nabla(\boldsymbol{\Phi} \circ \boldsymbol{g})$ 的每个列向量之间的点积，会得到一个维度为 $\dim(\boldsymbol{\theta})$的低维向量。雅可比矩阵的每一列都可以看作参数空间中使距离最小化的最速下降方向。因此，算子 $\boldsymbol{\Phi} \circ \boldsymbol{g}$ 应能得出反映 $\boldsymbol{\theta}$中变化的敏感度和方向的距离。

在公式 (16) 的 $\mathrm{L}_{\boldsymbol{\theta}}$ 中，我们采用时频散射（$\boldsymbol{Sx}$）（见第 2 节）作为 $\boldsymbol{\Phi}$。否则，使用多尺度频谱图（MSS）时，我们记为 $\mathrm{L}_{\boldsymbol{\theta}}^{MSS}$。在 JTFS 变换中，我们设置 $J = 12$、$J_{fr} = 5$、$Q_1 = 8$、$Q_2 = 2$、$Q_{fr} = 2$，并设置 $F = 0$ 以禁用频率平均。

另外，使用多尺度频谱图（MSS）时，我们记为 $\mathrm{L}_{\boldsymbol{\theta}}^{MSS}$。令 $\boldsymbol{\Phi}^{(n)}$ 短时傅里叶变换系数为使用窗长 2^n 计算得到的短时傅里叶变换系数。我们计算公式 (18) 中的 MSS 损失，即多个短时傅里叶变换分辨率下频谱图之间 L1 距离的平均值：

$$\mathcal{L}_{\boldsymbol{\theta}}^{MSS}(\tilde{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=5}^{10} |(\boldsymbol{\Phi}_{\mathrm{STFT}}^{(n)} \circ \boldsymbol{g})(\boldsymbol{\theta}) - (\boldsymbol{\Phi}_{\mathrm{STFT}}^{(n)} \circ \boldsymbol{g})(\tilde{\boldsymbol{\theta}})| \tag{18}$$

所选分辨率考虑了 $\boldsymbol{g}$ 使用的 8192 赫兹采样率。我们在之后的所有实验中均设置 $w = 2$ 个八度，并对每个 $\boldsymbol{g}_{\boldsymbol{\theta}}$ 的振幅进行归一化。

在本实验中，我们在对数刻度上均匀采样网格中的 20×20 调幅/调频速率（f_{m}，γ），范围为 4 至 16 赫兹和每秒 0.5 至 4 八度，共得到 400 个载波频率为 $f_{\mathrm{c}} = 512$赫兹的信号。我们将网格中心 $f_{\mathrm{m}} = 8$。29 赫兹、每秒$\gamma = 1$。49 八度处指定为目标声音。我们引入对目标声音施加恒定的时间偏移 $\tau = 2^{10}$ 采样点，以测试微观结构受到扰动时梯度的稳定性。我们评估每个声音在两个可微时频分析算子 $\boldsymbol{\Phi}_{\mathrm{STFT}}$ 和 $\boldsymbol{\Phi}_{\mathrm{JTFS}}$ 下对应的 $\mathrm{L}_{\boldsymbol{\theta}}$ 和 $\nabla\mathrm{L}_{\boldsymbol{\theta}}$。

我们在图 3 中可视化了损失曲面及其相对于 $\boldsymbol{\theta}$ 的梯度场~。我们观察到，JTFS 算子形成的损失曲面只有一个局部极小值，该极小值位于目标声音的 $\boldsymbol{\theta}$处。与此同时，尽管在较高 γ处存在一些例外（在听感上对应极高的调频率），但沿~采样的参数 $\boldsymbol{\theta}$的梯度始终指向目标~。相反，由于时间平移等变性，当 $\boldsymbol{\theta}$ 位于目标处时，多尺度频谱图损失的梯度会受到多个局部极小值的影响，且无法达到全局最小值。值得指出的是，MSS 距离对沿调幅方向的变化不敏感，因此不适合用于建模中观结构。

与我们的发现一致，先前的研究 [21] 发现，对 JTFS 最近邻图的流形嵌入进行三维可视化后，得到的三维网格的主成分与描述载波频率、调幅和调频的参数相关。此外，在 JTFS 空间中使用最近邻图进行 K 近邻回归时，三个参数各自的误差比都接近 1。

3.2 通过梯度下降进行声音匹配

~不同于经典声音匹配文献中通过可训练的 $f_{\boldsymbol{W}}$（即神经网络权重）进行 $\boldsymbol{\theta}$ 前向传播来估计 $\boldsymbol{\theta}$ 的做法，我们将声音匹配表述为$(\boldsymbol{\Phi} \circ \boldsymbol{g})$中的一个逆问题。为简单起见，我们不学习任何权重来近似 $\boldsymbol{\theta}$。

使用第3.1节推导出的梯度，我们尝试在 $\boldsymbol{\theta}$中对目标状态进行声音匹配，采用带有大胆驱动启发式方法的简单梯度下降方案。我们沿着梯度指示的方向，对 $\boldsymbol{\theta}$进行~加性更新

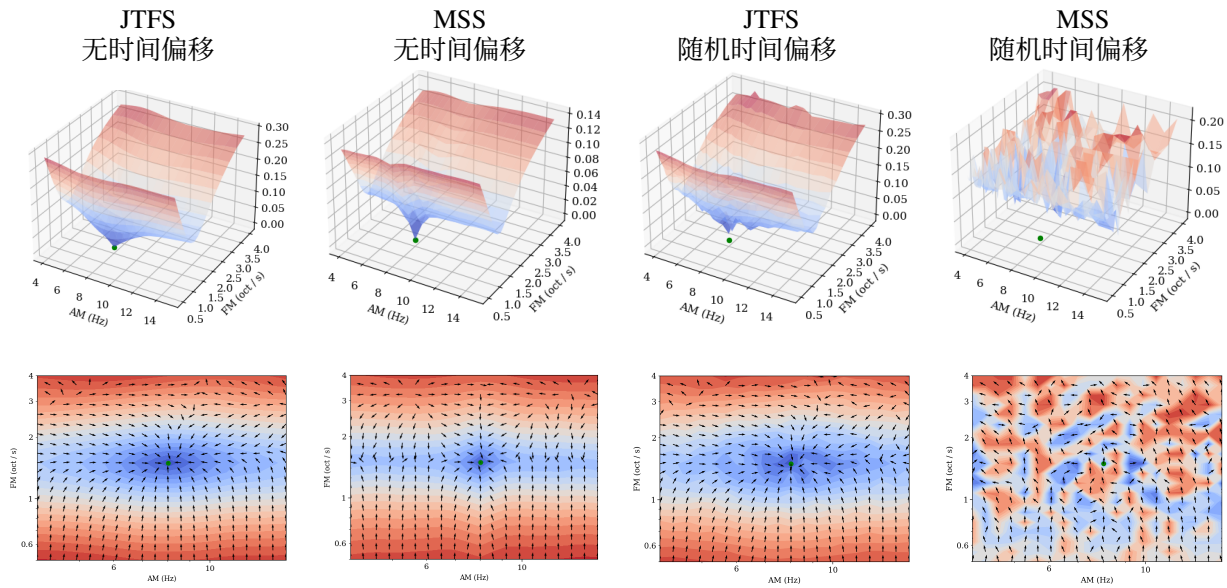

图 7. 在 $\mathbf{\Phi}_{\text{JTFS}}$ 和 $\mathbf{\Phi}_{\text{MSS}}$ 下，由 $\boldsymbol{g}$ 合成的声音（见第1节）对应的损失曲面（上）和梯度场（下）。这些声音从 f_{m} 和 γ 上的对数间隔网格中采样得到。每个声音相对于目标声音都随机进行了 2^n 个采样点的时间偏移，其中 n 在 $[8, 12]$之间均匀采样。我们将目标声音绘制为绿色圆点，并计算每个声音与目标声音之间在 $\mathbf{\Phi}_{\text{JTFS}}$ 和 $\mathbf{\Phi}_{\text{MSS}}$ 下的损失。在矢量场图中，我们计算生成声音的合成参数 f_{m} 和 γ 对应的损失算子的梯度。在两种无时间偏移的情况下，JTFS 的梯度都指向目标；当目标位于位置 0 时，周围的距离为 0。没有时间偏移时，MSS 会计算时频域中相交对象之间的距离。它的梯度似乎指向目标，但沿调幅方向存在局部极小值，图 6 所示的收敛情况证实了这一点。存在随机时间偏移时，JTFS 似乎具有鲁棒性，而 MSS 则高度不稳定，且容易陷入局部极小值。

$\nabla_{\tilde{\theta}} \mathcal{L}_{\theta}$:

$$\tilde{\theta} \leftarrow \tilde{\theta} - \alpha \nabla_{\tilde{\theta}} \mathcal{L}_{\theta} \tag{19}$$

当 L_{θ} 下降时，我们的大胆驱动启发式方法将学习率 α 提高 1.2 倍；否则则将其降低一半。我们在参数空间中的评估指标定义如下：

$$\mathcal{L}_{\theta}(\widetilde{\theta}) = \|\theta - \widetilde{\theta}\|_2^2 \tag{20}$$

图 4 展示了每个 $\mathbf{\Phi}$ 的梯度下降过程中，L2 参数误差的均值。我们选取固定的目标和初始预测，并进行了多次优化，在目标音频上考虑 0 到 2^{10} 个采样点的时间偏移。在 $\mathbf{\Phi}_{JTFS}$ 中低通滤波器的支撑范围 T 内进行时间偏移时，优化过程稳定收敛至接近 0。我们~观察到，MSS 无法收敛，而 $\text{L}_{\theta}(\theta)$ 未能从初始值大幅改善，包括不进行时间偏移的情况。在图 5 中，我们进一步展示了时间偏移对可微时频分析的影响，验证 JTFS 是一种在支撑范围 T内具有时间不变性的介观结构算子。

3.3 时间不变性

在图 6 中，我们探究了 $\boldsymbol{\theta}$ 采用不同初始化时的梯度收敛情况，且不对预测声音进行时间平移。在每张图中，我们针对 $\boldsymbol{\theta}$ 的 5 种不同初始化执行梯度下降：（i）远离目标声音；（ii）位于目标声音的局部邻域内；（iii）广泛分布于参数网格上。值得注意的是，在这 3 种初始化方案中，JTFS 均能收敛到解，这一点也得到了图 7 中梯度的印证。我们观察到，即使不进行时间偏移，当参数初始化远离目标时，MSS 确实无法恢复目标声音。若 $\boldsymbol{\theta}$ 初始化在目标的邻域内，MSS 确实能够恢复目标声音。尽管从任意位置开始时，MSS 在最佳情况下确实能够收敛，但平均而言，其表现接近无法收敛的最差情况。

图 7 展示了在不对预测声音施加时间偏移以及施加随机时间偏移的情况下，$\mathbf{\Phi}_{\text{JTFS}}$ 和 $\mathbf{\Phi}_{\text{MSS}}$ 的损失曲面与梯度场。尽管当预测声音与目标居中对齐时，MSS 能达到全局最小值，但我们的梯度下降实验表明，只有当 $\boldsymbol{\theta}$ 初始化于目标 $\boldsymbol{\theta}$ 的局部区域内时，~才能保持稳定。对预测声音施加随机时间偏移时，多尺度频谱图损失会变得非常不稳定，并产生许多不位于目标声音处的局部极小值。正如预期，不施加时间偏移时，JTFS 的梯度非常稳定。即使存在随机时间偏移，JTFS 仍是时频调制的一种不变表征，其不变性适用于时间偏移 T。

4 结论

可微时频分析（DTFA）是音频深度学习任务中一个新兴的方向。目前，自编码、音频修复和声音匹配领域的最先进方法主要在频谱图域中执行可微时频分析。然而，在以下情形中计算相似度时，频谱图损失会受到数值不稳定性的影响：(i) 时间偏移超出频谱图窗口的尺度；(ii) 合成参数引起的非平稳性。这些问题使频谱图损失无法可靠地用作.....

……用于建模多尺度音乐结构的相似性度量。

本文介绍了可微中观结构算子，由 DTFA 和琶音合成器组成，用于中观结构的时频分析。对于声音匹配任务，我们使用联合时频散射（JTFS）对合成参数建模，以通过 DTFA 分析超越微观结构局部性的可识别结构；即啁啾小波合成器的振幅和频率调制。值得注意的是，JTFS 提供了一种可微且可扩展的实现方式，可用于听觉谱时感受野、时频域中的多尺度分析，以及对时间平移的不变性。

然而，尽管已有证据表明，JTFS 能够准确地对包含时频调制的信号中的相似度进行建模，但 JTFS 在逆问题和声音合成控制的 DTFA 中仍有待评估。通过分析 DTFA算子相对于合成参数的梯度，我们表明，与频谱图损失不同，JTFS 距离适用于对描述中观结构的合成参数之间的相似度进行建模。我们通过梯度下降进行声音匹配，证明了 JTFS 作为 DTFA 算子的稳定性，尤其是在存在时间平移的情况下。

本研究为进一步探索可微时频分析在自编码、声音匹配、再合成和计算机音乐创作中的应用奠定了基础。事实上，我们的可微中尺度结构算子可以直接作为原始音频波形的模型，但这种方法容易产生再合成伪影 [24, 22]。我们已证明，借助可微时频分析，我们可以对塑造序列音频事件的低维合成参数进行建模。未来的一个研究方向是可微参数化纹理合成，其中纹理相似性可以通过由更大尺度结构导出的参数进行优化，例如，不再局限于颗粒合成中对单个音粒的定义。

5 致谢

Cyrus Vahidi 是 UKRI 人工智能与音乐博士培训中心的研究人员，该中心由 UKRI（资助编号 EP/S022694/1）和 Music Tribe 共同资助。本研究是在法国国家科学研究中心（CNRS）的 LS2N 进行研究访问期间开展的。CHANGHONG WANG 获得了 Atlanstic2020 项目 Trainable Acoustic Sensors（TrAcS）的资助。

6 参考文献

[1] C. Schörkhuber 和 A. Klapuri，《用于音乐处理的恒定 Q 变换工具箱》，发表于西班牙巴塞罗那的声音与音乐计算（*SMC*）会议，第 3–64 页（2010年）。

[2] M. Müller，《音乐处理基础：音频、分析、算法与应用》，第 5 卷（施普林格，2015年）。

[3] K. W. Cheuk、H. Anderson、K. Agres 和 D. Herremans，“nnAudio：一种利用一维卷积神经网络实现音频到频谱图实时转换的图形处理器工具箱”，*IEEEAccess*，第 8 卷，第 161981–162003 页（2020年）。

[4] M. Andreux 和 S. Mallat，“基于矩匹配—散射逆变换的音乐生成与变换
Networks》，发表于国际音乐信息检索学会（*ISMIR*）（2018年）。[5] Y.-Y. Yang、M. Hira、Z. Ni、A. Chourdia、A. Asta- furov、C. Chen，等，《TorchAudio：音频与语音处理构件》，*arXiv* 预印本 *arXiv:2110.15018*（2021年）。[6] J. Engel、C. Gu、A. Roberts、等，《DDSP：可微数字信号处理》，发表于国际学习表征会议（*ICLR*）（2019年）。

[7] 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，第 30 卷，495–507 页（2021年）。

[8] P. Manocha、A. Finkelstein、R. Zhang、N. J. Bryan、G. J. Mysore 和 Z. Jin，“基于最小可觉差异学习的可微感知音频度量”，发表于*INTERSPEECH*（2020年）。

[9] J. Su、Y. Wang、A. Finkelstein 和 Z. Jin，“带宽扩展即你所需的一切”，发表于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），696–700 页（2021年）。

[10] P. Esling、N. Masuda、A. Bardet、R. Despres 和 A. Chemla-Romeu-Santos，《流合成器：使用归一化流实现通用音频合成器控制》，《应用科学》，第 10 卷，第 1 期，第 302 页（1999年）。

[11]N. Masuda 和 D. Saito，《使用可微数字信号处理进行合成器声音匹配》。发表于国际音乐信息检索学会（*ISMIR*）会议，第 428–434 页（2021年）。

[12] J. Turian 和 M. Henry，《很遗憾你遭遇了损失：基于频谱的音频距离不擅长处理音高》，发表于“我不敢相信它竟然没更好！”*NeurIPS 2020* 年研讨会（2020年）。

[13] K. Brandenburg，《MP3 和 AAC 详解》，发表于音频工程学会（*AES*）会议（1999年）。

[14] M. Subotnick，《Buchla合成器在音乐创作中的应用》，发表于第*38*届音频工程学会年会 (1970年)。[15] C. Roads，《微观声音》(麻省理工学院出版社，2004年)。[16] C. Roads，《电子音乐中的节奏过程》，发表于国际计算机音乐会议 *(ICMC)*(2014年)。[17] C. Roads，《从音粒到形式》，发表于伊阿尼斯·泽纳基斯国际研讨会，第8卷 (2012年)。[18] V. Lostanlen、J. Andén 和 M. Lagrange，《傅里叶在计算机音乐的核心：从谐波声音到纹理》，《物理学报告》，第20卷，第5期，第461–473页 (2019年)。[19] M. Andreux、T. Angles、G. Exarchakis、R. Leonar- duzzi、G. Rochette、L. Thiry、等，《Kymatio：Python中的散射变换》。《机器学习研究期刊》，第21卷，第60期，第1–6页 (2020年)。[20] J. Andén、V. Lostanlen 和 S. Mallat，《联合时频散射》，《*IEEE*信号处理汇刊》，第67卷，第14期，第3704–3718页 (2019年)。

[21]J. Muradeli、C. Vahidi、C. Wang、H. Han、V. Lostanlen、M. Lagrange、等，《基于图形处理器的可微时频散射》，发表于数字音频效果会议（*DAFx*）(2022年)。

[22]V. Lostanlen 和 F. Hecker，《RemiXXXes 的未来形态：基于时频散射的音频纹理合成》，发表于国际数字音频效果会议（*DAFx*）（2019年 06月）。

[23] T. Chi、P. Ru 和 S. A. Shamma，《复杂声音的多分辨率时频分析》，美国声学学会期刊，第 118 卷，第 2 期，第 887–906 页（2005年）。

[24]J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, 等，“使用 WaveNet 自编码器进行神经音频合成，生成音符”，发表于国际机器学习会议 *(ICML)*，第 1068–1077 页（2017 年）。

作者