arXiv:2601.06406v1 [cs.SD] 10 Jan 2026

# 将声音表示为神经振幅场：坐标多层感知机与傅里叶—柯尔莫哥洛夫–阿诺德框架的基准测试

**Linfei Li**[1], **Lin Zhang**[1*], **Zhong Wang**[2],**Fengyi Zhang**[3],**Zelin Li**[4],**Ying Shen**[1同 1]
济大学计算机科学与技术学院 [2] 上海交通大学自动化系[3] 昆士兰大学电气工程与计算机科学学院[4]西北大学麦考密克工程学院 cslinfeili@tongji.edu.cn，cslinzhang@tongji.edu.cn，cszhongwang@sjtu.edu.cn，fengyi.zhang@uq.edu.au，zelinli2025@u.northwestern.edu，yingshen@tongji.edu.cn

**摘要**

尽管基于坐标多层感知机的隐式神经表示在辐射场、三维形状和图像表示方面表现出色，但其在音频信号中的应用仍未得到充分探索。为填补这一空白，我们研究了现有的隐式神经表示，并从中提取出 3 种位置编码和 16 种常用激活函数。通过组合设计，我们建立了首个针对音频信号表示的坐标多层感知机基准测试。基准测试结果表明，坐标多层感知机需要复杂的超参数调优和频率相关初始化，限制了其鲁棒性。为解决这些问题，我们提出了 Fourier-ASR，这是一种基于傅里叶级数定理和柯尔莫哥洛夫-阿诺德表示定理的新型框架。Fourier-ASR 引入了傅里叶柯尔莫哥洛夫-阿诺德网络（Fourier-KAN），利用周期性和强非线性来表示音频信号，无需额外的位置编码。此外，我们提出了频率自适应学习策略（FaLS），通过捕获高频分量并防止低频信号过拟合，提升 Fourier-KAN 的收敛性。在自然语音和音乐数据集上开展的大量实验表明：(1) 坐标多层感知机中经过精心设计的位置编码和激活函数能够有效提升音频表示质量；(2) Fourier-ASR 无需大量超参数调优，即可鲁棒地表示复杂音频信号。展望未来，隐式音频表示的连续性和无限分辨率使我们的研究在音频压缩、合成和生成等任务中极具应用前景。为确保研究结果可复现，源代码将公开发布。代码可在 https://github.com/lif314/Fourier-ASR 获取。

## 引言

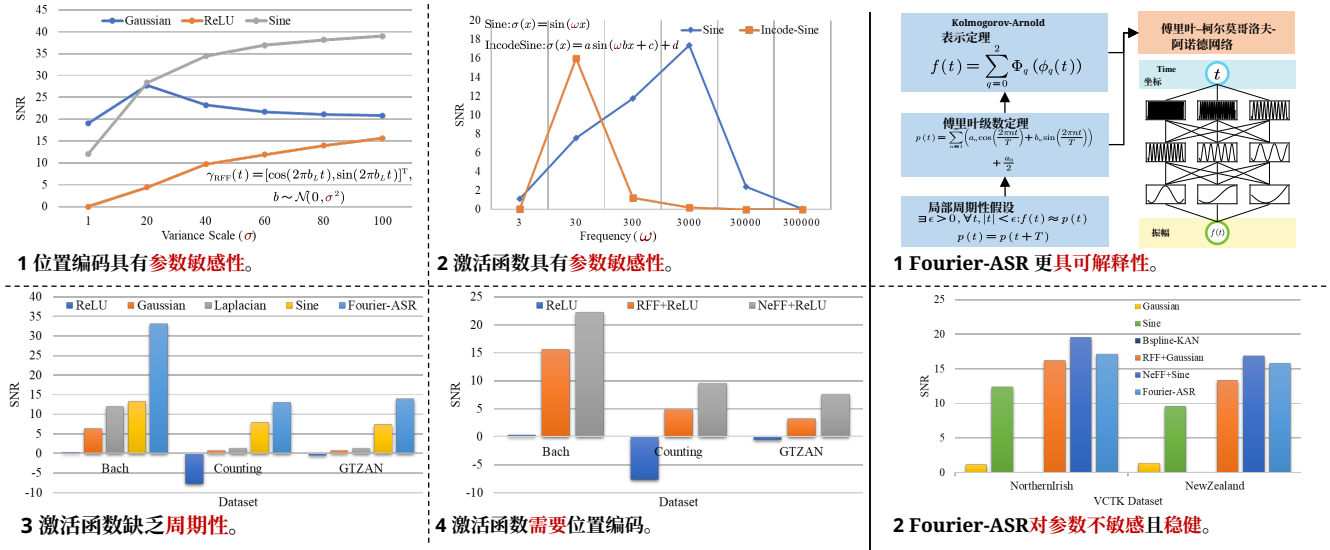
隐式神经表示（INR）通过将任意离散信号表示为连续函数，为信号参数化提供了一种创新方法。这些函数将信号的定义域（坐标，例如音频中的时间戳）映射到相应坐标上的内容（例如音频信号的振幅）。通常，这些函数由神经网络近似，而由于当前的神经网络主要由多层感知机（MLP）构成，因此这类隐式神经表示称为坐标多层感知机。

与传统的离散信号表示方案相比，隐式神经表示提供了一种连续隐式表示，它不受空间分辨率限制，并可实现无限分辨率。因此，参数化信号所需的存储空间与空间分辨率无关，这些信号可以按任意所需分辨率进行采样。得益于这些优势，坐标多层感知机已成功应用于各种数据模态，包括神经辐射场（Mildenhall等，2020）、三维占据网格（Mescheder 等人 2019）、有符号距离函数（Park 等人 2019）、图像（Sitzmann 等人 2020）、二维计算机断层扫描和三维磁共振成像（Tancik等，2020；Saragadam等，2023；Kazerouni 等人 2024）。

对于音频信号而言，连续表示具有无限分辨率的优势，能够实现自然生成、高效压缩和平滑处理。然而，使用坐标多层感知机表示连续音频信号面临着巨大挑战，因为音频信号本身具有噪声高、频率高、非线性和局部周期性等特点。根据韦伯-费希纳定律，由于人类听觉感知具有对数特性，即使音频信号中的重构误差相对较小，也可能被人感知到，因此对音频重建质量提出了很高的要求。此外，多层感知机网络由线性变换和非线性激活函数简单组合而成，难以捕捉音频信号的周期性和高频分量。经过全面调研，我们发现，迄今为止，只有 SIREN（Sitzmann 等人 2020）尝试使用正弦激活函数表示音频信号，并与 ReLU多层感知机进行了简单比较，但此后没有开展进一步研究。

为填补这一空白，据我们所知，我们建立了**首个**开源基准测试框架，以全面探索坐标多层感知机在连续音频信号表示中的潜力与局限。具体而言，由于坐标多层感知机的性能主要取决于激活函数的选择以及是否采用位置编码，我们从现有的坐标多层感知机方法中确定了 3 种位置编码映射和 16 种常用激活函数

[*]通讯作者 版权所有 © 2025，人工智能促进协会（www.aaai.org）。保留所有权利。




（a）音频信号表示中的坐标多层感知机基准测试。

（b）Fourier-ASR。

图 1：坐标多层感知机和傅里叶-ASR 的属性。验证结果见附录（附录 A）。

由此得到 48 种可用于音频信号表示的网络配置，我们在语音和音乐数据集上对其进行评估，以衡量其性能。如图 1(a) 所示，我们的基准测试得出以下发现。(1) 除了线性特征显著（如高斯）或具有周期性（如正弦）的激活函数外，大多数激活函数都无法有效表示音频信号。(2) 尽管有些激活函数（如高斯和正弦）旨在克服频谱偏差以及位置编码所需的繁琐参数调优，但位置编码对于表示音频信号仍不可或缺。它能将时间坐标高效映射到高维空间，使网络能够捕捉音频信号中的高频分量。(3) 由于音频信号具有局部周期性，周期性激活函数（如正弦）在表示能力方面显著优于其他激活函数。此外，加入基于傅里叶特征的位置编码可进一步增强其能力。(4) 正弦型激活函数凭借其周期性，在表示音频信号方面表现出色，但遗憾的是，它们需要对超参数敏感的位置编码和依赖频率的初始化方案，从而对其稳健性和泛化能力产生负面影响。

上述坐标多层感知机的问题，根本上源于多层感知机固有的非线性能力不足和缺乏周期性。如图 1(b) 所示，为增强神经网络的非线性表示能力和周期性表示能力，我们基于傅里叶级数定理和 Kolmogorov-Arnold 表示定理，提出了一种新颖的隐式音频表示框架 Fourier-ASR。首先，我们引入 Kolmogorov-Arnold 网络（Fourier-KAN），利用傅里叶基函数表示音频信号。该网络将任意复杂音频信号隐式分解为一系列局部周期性的傅里叶级数。与多层感知机不同，Fourier-KAN 无需额外的位置编码或激活函数，从而避免了繁琐的超参数调优。此外，由于使用了傅里叶基函数，它能够更有效地捕捉信号的高频成分和局部周期性。其次，为加速 Fourier-KAN 的收敛，我们提出了一种频率自适应学习策略（FaLS）。FaLS 采用反向频率金字塔配置来捕捉不同频率的信号，并基于前向传播理论采用频率自适应权重初始化方案，以缓解梯度爆炸或梯度消失问题，从而加快收敛。实验结果表明，Fourier-ASR 不仅具有更强的可解释性，而且对超参数变化具有较强的鲁棒性，能够有效表示复杂音频信号。

总而言之，我们的贡献总结如下：

- 我们为音频表示中的坐标多层感知机引入了**首个基准测试**，涵盖 3 种位置编码和 16 种常用激活函数。我们的基准测试深入分析了位置编码和激活函数对连续音频信号表示的影响。

- 为避免位置编码带来的频谱偏差，以及激活函数复杂的参数调优，我们提出了一种基于傅里叶级数定理和柯尔莫哥洛夫-阿诺德定理的新型音频信号表示框架，**Fourier-ASR**。Fourier-ASR 包含傅里叶柯尔莫哥洛夫-阿诺德网络（Fourier-KAN）和频率自适应学习策略（FaLS）。凭借傅里叶基函数的周期性和强非线性，Fourier-ASR 能够有效表示音频信号，并增强可解释性。

- 如图 1 所示，我们在

语音和音乐数据集上开展的大量实验表明：(1) 仔细调整位置编码和激活函数参数，可以显著增强坐标多层感知机表示音频信号的能力；(2) Fourier-ASR 无需繁琐的参数调整，便能稳健地表示音频信号。

## 相关工作

**坐标多层感知机**。 坐标多层感知机的使用方式与传统多层感知机存在两方面的显著差异：(a) 传统多层感知机通常处理高维输入，例如图像、声音或三维形状；(b) 传统多层感知机主要用作分类头，其决策边界不必具备光滑性。相比之下，坐标多层感知机将信号编码到权重中，其中输入是低维坐标，而输出必须保持光滑性。NeRF（Mildenhall等，2020）的成功表明，在有限数量的透视图像上进行训练后，坐标多层感知机能够从任意角度、以任意分辨率重建光度投影。这一突破推动了坐标多层感知机在众多领域中的应用，包括辐射场重建（Barron 等人 2022；Chen 等人 2022；M¨uller 等人 2022）、三维形状表示（Wang 等人 2021；Yu 等人 2022；Yariv 等人 2023）、二维图像回归（Tancik等，2020；Saragadam 等人 2023；Lindell 等人 2022；Ramasinghe和Lucey 2022）、音频信号回归（Sitzmann 等人 2020；Kazerouni 等人 2024），以及二维计算机断层扫描和三维磁共振成像中的逆向渲染问题（Tancik等，2020）。

**位置编码**。 位置编码有助于在辐射场、图像和三维形状中学习高频表示。NeRF（Mildenhall等，2020）通过将输入坐标映射到高维傅里叶空间，提升了ReLU多层感知机捕获高频信号的能力。在NeRF的基础上，FFN（Tancik等，2020）引入高斯噪声，以增强ReLU多层感知机的鲁棒性。尽管位置编码使多层感知机能够表示高频分量，但选择合适的频率尺度至关重要，而且通常需要繁琐的参数调优。具体而言，当信号带宽增加过多时，坐标多层感知机往往会产生带噪的信号插值（Ramasinghe、MacDonald和Lucey，2022；Hertz等，2021）。

**激活函数**。随后，坐标多层感知机的非线性表示能力主要源自激活函数。在隐式神经表示领域，人们采用了各种激活函数来近似不同类型的信号。由于简单且初始化方案有效，ReLU 常被用于与 NeRF 相关的研究中（Mildenhall等，2020；Barron 等，2022；Yu 等，2021；Chen 等，2022）。然而，ReLU 难以捕捉辐射场中的高频信息，因此需要额外的位置编码。为避免位置编码所带来的繁琐参数调优，GARF（Chng等，2022）使用高斯激活函数，能够有效捕捉高频信息，但无法捕捉周期信号，并且倾向于对噪声和信号同等过拟合。为解决这些问题，WIRE（Saragadam等，2023）采用复数伽博小波激活函数来提升鲁棒性。SIREN（Sitzmann 等人 2020）采用正弦激活函数来捕捉信号的周期性，但对初始化方案较为敏感，限制了其在音频重建任务中的泛化能力。在 SIREN 的基础上，INCODE（Kazerouni 等人 2024）使正弦激活函数的参数可学习，从而在一定程度上降低了参数敏感性，但仍依赖频率感知初始化方案。

## 方法

### 问题表述

如图2(a)所示，自然音频信号是关于时间的连续函数，表示声音信号振幅随时间的变化。为了将这一连续信号转换为便于存储和处理的数字格式，需要对信号进行离散采样，从而得到相对于时间坐标 $t$ 的离散信号$a(t)$。然而，在音频超分辨率、合成和压缩等领域，研究人员希望利用隐式神经表示技术，尽可能保留信号的连续性和可微性。具体而言，神经网络接收离散时间坐标 $t$，并回归与 $t$ 对应的振幅 $f(t)$，从而将音频信号编码到网络权重中。我们将这种表示称为**神经振幅场（NeAF）**。通过使用均方误差损失函数，将 $f(t)$ 拟合到采样波形 $a(t)$，以进行优化，

$$\mathcal{L} = \int \left\| \mathrm{II}_a(f(t)) - a(t) \right\|^2 dt, \tag{1}$$

其中 $\mathrm{II}_a$ 个样本 $f$（$t$）位于波形测量位置处。鉴于 NeAF 不受空间分辨率影响，音频可以按任意所需分辨率进行处理。

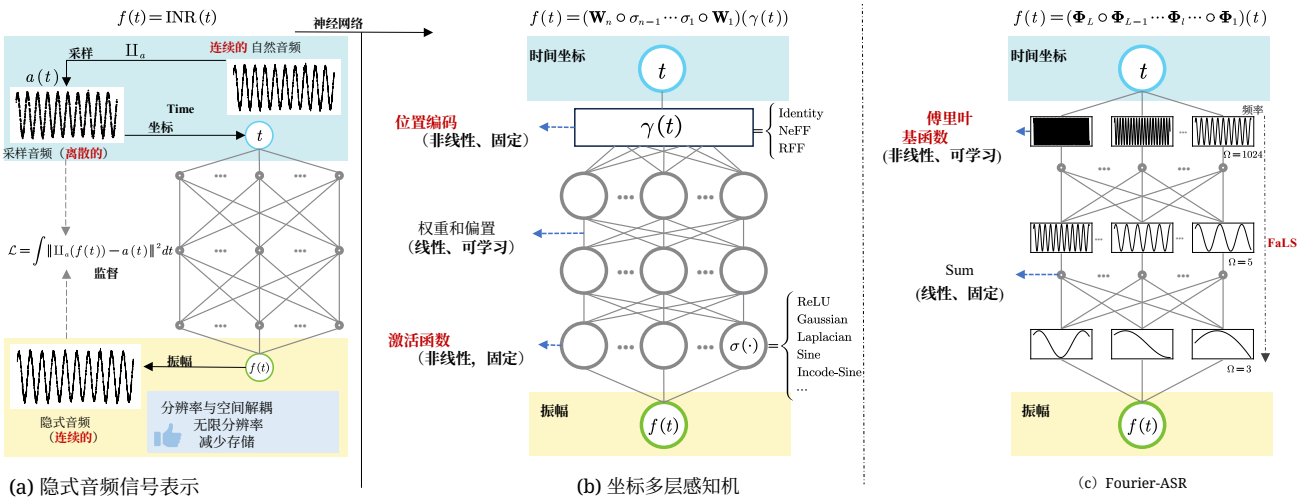
### 基于坐标多层感知机的 NeAF 基准测试

如图 2(b) 所示，为了表示任意复杂的音频信号，采用了一个 $k$ 层坐标多层感知机 $f: \mathbb{R} \to \mathbb{R}$，它以时间坐标 $t \in \mathbb{R}$ 作为输入，并输出振幅 $f$（$t$）$\in \mathbb{R}$。因此，$f$（$t$）可以通过以下递归关系定义：

$$\begin{aligned} \mathbf{z}^{(1)} &= \gamma(t) \\ \mathbf{z}^{(i+1)} &= \sigma\left(\mathbf{W}^{(i)}\mathbf{z}^{(i)} + \mathbf{b}^{(i)}\right), i = 1, \ldots, k-1 \\ f(t) &= \mathbf{W}^{(k)}\mathbf{z}^{(k)} + \mathbf{b}^{(k)}, \end{aligned} \tag{2}$$

其中，$\gamma$（·）表示一个可选的位置编码函数，它将输入坐标 $t$ 映射到高维空间；$\sigma$（·）表示逐元素应用的非线性激活函数；$\mathbf{W}^{(i)} \in \mathbb{R}^{d_{i+1} \times d_i}$ 和 $\mathbf{b}^{(i)} \in \mathbb{R}^{d_{i+1}}$ 分别表示第 $i$ 层的权重和偏置，而 $\mathbf{z}^{(i)} \in \mathbb{R}^{d_i}$ 表示第 $i$ 层的隐藏单元。

遵循坐标多层感知机的架构，我们对隐式神经表示进行了广泛综述




(a) 隐式音频信号表示　　(b) 坐标多层感知机　　(c) Fourier-ASR

图2：（a）隐式音频表示的问题定义；（b）基于坐标多层感知机的音频表示框架；（c）Fourier-ASR，一种基于 Fourier-KAN 的新型音频信号表示框架。

并确定了 3 种位置编码和 16 种潜在激活函数。具体而言，如表1所示，三种位置编码方案分别为恒等映射（Identity）、NeRF 傅里叶特征（NeFF）（Mildenhall 等人 2020）和随机傅里叶特征（RFF）（Tancik 等人 2020）。主要的激活函数包括 ReLU、高斯（Chng等，2022）、拉普拉斯（Ramasinghe 和 Lucey 2022）、正弦（Sitzmann 等人 2020）、Incode-Sine（Kazerouni 等人 2024）和 Gabor-小波（Saragadam等，2023）等（详情见表 2）。

| 位置编码（$\mathcal{P}$） | $\gamma\in\mathcal{P}$ | 参数 |
|---|---|---|
| 恒等函数 | $\gamma(t)=t$ | - |
| NeFF | $\gamma(t)=[\text{余弦}(2^L\pi t),\text{正弦}(2^L\pi t)]^{\mathrm{T}}$ | $[L]$ |
| RFF | $\gamma(t)=[\cos(2\pi b_L t),\ \sin(2\pi b_L t)]^{\mathrm{T}}$, $b_L\sim\mathcal{N}(0,\sigma^2)$ | $[\sigma,L]$ |
| **激活函数（$\mathcal{A}$）** | $\sigma\in\mathcal{A}$ | |
| ReLU | $\sigma(x)=\max(0,x)$ | - |
| 高斯 | $\sigma(x)=e^{-\frac{x^2}{2a^2}}$ | $[a]$ |
| 拉普拉斯 | $\sigma(x)=e^{-\frac{\lvert x\rvert}{a}}$ | $[a]$ |
| Sine | $\sigma(x)=\sin(\omega x)$ | $[\omega]$ |
| Incode-Sine | $\sigma(x)=a\sin(bx+c)+d$ | $a,b,c,d,[\omega]$ |
| ... | ... | ... |

表1：坐标多层感知机中的非线性映射。注意，$a$ 表示可学习参数，而 $[a]$ 表示超参数。

值得注意的是，提出高斯、正弦和 Incode-Sine 激活函数，旨在消除辐射场和图像表示对位置编码的依赖。然而，高维位置编码映射可能有助于学习音频信号中的高频特征，以及这些特征在不同时间尺度上的结构变化。因此，在我们的基准测试中，我们对这三种激活函数均应用了位置编码映射。

根据表1，用于对音频信号表示进行基准测试的坐标多层感知机可表示如下：

$$f(t)=(\mathbf{W}_n\circ\sigma_{n-1}\cdots\sigma_1\circ\mathbf{W}_1)(\gamma(t)),\gamma(\cdot)\in\mathcal{P},\sigma_i(\cdot)\in\mathcal{A}, \tag{3}$$

其中，$t$ 表示归一化到区间 $[0,1]$的输入时间坐标，$\mathcal{P}$ 表示位置编码集合，$\mathcal{A}$ 表示激活函数集合。

**Fourier-ASR**

我们的基准测试表明，只有通过精心设计的位置编码和激活函数，一些坐标多层感知机才能有效表示音频信号。然而，由于参数调优复杂且对初始化高度敏感，它们的灵活性和通用性有所降低。为解决这一问题，如图2(c)所示，我们借鉴傅里叶级数定理和柯尔莫哥洛夫-阿诺德表示定理，提出了一种用于音频信号表示的新框架 Fourier-ASR，该框架结合了傅里叶柯尔莫哥洛夫-阿诺德网络（Fourier-KAN）和频率自适应学习策略（FaLS）。

**傅里叶柯尔莫哥洛夫-阿诺德网络（Fourier-KAN）。**不同于基于通用逼近定理（Hornik、Stinchcombe 和 White，1989 ）的坐标多层感知机（它们使用线性变换与非线性的组合），Fourier-ASR 遵循柯尔莫哥洛夫–阿诺德表示定理（Kolmogorov，1956；Arnold，1957），将任意连续函数表示为单变量函数与加法的有限复合。对于连续信号 $f$（$t$），这可简化为：

$$f(t)=\sum_{q=0}^{2}\Phi_q\left(\phi_q(t)\right), \tag{4}$$

其中，$\Phi_q:\mathbb{R}\to\mathbb{R}$ 和 $\phi_q:[0,1]\to\mathbb{R}$ 分别表示外函数和内函数。为了增强该表示的容量并

提升其可学习性，我们采用 KAN（Liu 等，2024）方法，将网络扩展为任意层数。

**假设 1** 局部周期性假设。 对于复杂的非平稳信号 $f(t)$，存在一个足够小的时间区间 $\epsilon > 0$，使得 $f(t)$ 可以近似为周期函数 $p(t)$：

$$\exists \epsilon > 0, \quad \forall t, \quad |t| < \epsilon : \quad f(t) \approx p(t)$$

其中，$p(t)$ 是周期为 $T$ 的周期函数，满足 $p(t+T) = p(t)$。

**定理 1** 傅里叶级数定理。 任何周期为 $T$ 的周期信号 $p(t)$ 都可以表示为正弦函数和余弦函数的无穷级数，

$$p(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right)$$

其中，$a_0$、$a_n$ 和 $b_n$ 是傅里叶系数。

具体而言，考虑一个形状为 $[n_0, n_1, \cdots, n_L]$ 的神经网络，其中 $n_l$ 表示计算图第 $l$ 层上的神经元数量。对于第 $l$ 层上的第 $i$ 个节点，记为 $(l, i)$，该神经元的激活值为 $t_{l,i}$。在第 $l$ 层与第 $(l+1)$ 层之间，有 $n_l \times n_{l+1}$ 个非线性基函数。根据假设 1 和定理 1，任何音频信号在短时间区间内都可以近似表示为余弦函数和正弦函数的组合。因此，不同于在 KAN 中使用 B 样条（Liu 等，2024），我们采用傅里叶基函数作为连接神经元 $(l, i)$ 和 $(l+1, j)$ 的非线性单元，

$$\begin{aligned} &\phi_{l,j,i}(t_{l,i}) = a_{l,i} \cos(\omega t_{l,i}) + b_{l,i} \sin(\omega t_{l,i}) + c_{l,i}, \\ &l = 0, \cdots, L-1, i = 1, \cdots, n_l, j = 1, \cdots, n_{l+1}, \end{aligned} \tag{5}$$

其中，$a_{l,\ i}$、$b_{l,\ i}$ 是可学习的傅里叶系数，$c_{l,\ i}$ 是可学习的偏置项，$\omega$ 是频率超参数。随后，$(l+1,\ j)$神经元的激活值 $t_{l+1,\ j}$ 就是所有输入后激活值的求和，

$$t_{l+1,j} = \sum_{i=1}^{n_l} \phi_{l,j,i}\left(t_{l,i}\right), j = 1, \cdots, n_{l+1}. \tag{6}$$

对于第 $l$ 个 Fourier KAN 层，将公式 6 改写为矩阵形式，可得：

$$\mathbf{t}_{l+1} = \underbrace{\begin{pmatrix} \phi_{l,1,1}(\cdot) & \phi_{l,1,2}(\cdot) & \cdots & \phi_{l,1,n_l}(\cdot) \\ \phi_{l,2,1}(\cdot) & \phi_{l,2,2}(\cdot) & \cdots & \phi_{l,2,n_l}(\cdot) \\ \vdots & \vdots & & \vdots \\ \phi_{l,n_{l+1},1}(\cdot) & \phi_{l,n_{l+1},2}(\cdot) & \cdots & \phi_{l,n_{l+1},n_l}(\cdot) \end{pmatrix}}_{\mathbf{\Phi}_l} \mathbf{t}_l.$$

其中，$\mathbf{\Phi}_l$ 是 Fourier 层之间的转移矩阵。

总之，Fourier-ASR 使用 Fourier-KAN 从离散音频信号中推导连续表示，形式如下：

$$f(t) = (\mathbf{\Phi}_L \circ \mathbf{\Phi}_{L-1} \cdots \mathbf{\Phi}_l \cdots \circ \mathbf{\Phi}_1)(t). \tag{7}$$

与坐标多层感知机（公式 3）相比，我们的 Fourier-KAN 利用傅里叶基函数，不仅增强了非线性表征能力，还能够捕捉音频信号中的局部周期性。

**频率自适应学习策略（FaLS）**。由于音频信号在不同时间尺度上的频率分布不同，固定的频率超参数（公式 5 中的 $\omega$）可能导致 Fourier-KAN 主要学习特定的频率分量，从而阻碍收敛。为解决这一问题，我们提出了频率自适应学习策略（FaLS）。具体而言，我们为不同的 Fourier-KAN 层分配具有不同频率阈值的基函数。于是，Fourier-KAN 可表示为：

$$\begin{aligned} \mathbf{z}^{(0)} &= t \\ \mathbf{z}^{(l+1)} &= \sum_{\omega=1}^{\Omega_l} \left[ \mathbf{a}^{(l,\omega)} \cos(\omega \mathbf{z}^{(l)}) + \mathbf{b}^{(l,\omega)} \sin(\omega \mathbf{z}^{(l)}) \right] + \mathbf{c}^{(l)} \\ f(t) &= \mathbf{z}^{(n_L)} \left( \mathbf{z}^{(n_L-1)} \left( \mathbf{z}^{(l)} \left( ...(\mathbf{z}^{(0)}) \right) \right) \right), l = n_0, \ldots, n_L, \end{aligned}$$

其中，$\mathbf{a}^{(l,\omega)}$ 和 $\mathbf{b}^{(l,\omega)} \in \mathbb{R}^{d_{l+1} \times d_l}$ 表示第 $l$ 层在频率 $\omega$ 处的傅里叶系数权重，$\mathbf{c}^{(l)} \in \mathbb{R}^{d_{l+1}}$是第 $l$ 层的偏置项，$z^{(l)} \in \mathbb{R}^{d_l}$ 表示第 $l$ 层的隐藏单元，$\Omega_l$ 是一个超参数，表示第 $l$ 层的最大频率阈值。

参数初始化。遵循 Xavier（Glorot 和 Bengio, 2010）和 Kaiming 的研究（He 等，2015）所确立的原则，我们推导出 Fourier-KAN 的初始化方案。具体而言，在 Fourier-KAN 第 $l$ 层的前向传播过程中，傅里叶基函数的对称性确保输入和输出的期望值均为零，即 $E[\mathbf{z}^{(l)}] = E[\mathbf{z}^{(l+1)}] = 0$。根据 Kaiming 初始化（He 等，2015），我们作出如下假设：（1）傅里叶参数 $\mathbf{a}^{(l,\omega)}$ 和$\mathbf{b}^{(l,\omega)}$ 的期望值均为零，并省略偏置项 $\mathbf{c}^l$；（2）输入 $\mathbf{z}^{(l)}$ 和输出$\mathbf{z}^{(l+1)}$ 的方差均为 1。因此，我们可以确定第 $l$ 层输出的方差为：

$$\begin{aligned} Var[\mathbf{z}^{(l+1)}] &= Var[\sum_{\omega=1}^{\Omega_l} \left[ \mathbf{a}^{(l,\omega)} \cos(\omega \mathbf{z}^{(l)}) + \mathbf{b}^{(l,\omega)} \sin(\omega \mathbf{z}^{(l)}) \right]] \\ &= \sum_{\omega=1}^{\Omega_l} \left( \cos^2(\omega \mathbf{z}^{(l)}) Var[\mathbf{a}^{(l,\omega)}] + \sin^2(\omega \mathbf{z}^{(l)}) Var[\mathbf{b}^{(l,\omega)}] \right). \end{aligned}$$

假设傅里叶系数的方差相等，则可得：

$$Var[\mathbf{a}^{(l)}] = Var[\mathbf{b}^{(l)}] = \frac{1}{\Omega_l}. \tag{8}$$

因此，每个独立的傅里叶系数 $a_i^{(l)}$（以及 $b_i^{(l)}$）均采用以下正态分布进行初始化，

$$a_i^{(l)}, b_i^{(l)} \sim \mathcal{N}(0, \frac{1}{\Omega_l d_{in}^{(l)}}), \tag{9}$$

其中，$d_{in}^{(l)}$ 表示层 $l$ 输入的维数。

| 激活函数 $\sigma(\cdot)$ | 方程 | 参数 | 位置编码 $\gamma(\cdot)$ | 巴赫（7秒） | | 计数（7秒） | | 布鲁斯（30 秒） | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ |
| PReLU | $\begin{cases} x, & \text{如果 } x > 0 \\ ax, & \text{否则} \end{cases}$ | $[a]$ | 恒等函数 | 0.00 | 4.724 | 0.00 | 4.630 | 0.00 | 7.031 | 0.00 | 5.462 |
| | | | RFF | 13.42 | 1.010 | 3.38 | 1.437 | 2.50 | 2.035 | 6.43 | 1.494 |
| | | | NeFF | 17.50 | 1.133 | 7.88 | 1.575 | 5.20 | 1.539 | 10.19 | 1.416 |
| ReLU | $\max(0, x)$ | | 恒等函数 | 0.00 | 4.623 | -7.66 | 4.546 | 0.00 | 6.774 | -2.55 | 5.314 |
| | | | RFF | 15.62 | 0.978 | 4.93 | **1.400** | 3.23 | 1.862 | 7.93 | 1.413 |
| | | | NeFF | 22.29 | 1.129 | 9.57 | 1.538 | 7.64 | 1.324 | 13.17 | 1.330 |
| 高斯 | $e^{\frac{-x^2}{2a^2}}$ | $[a]$ | 恒等函数 | 6.35 | 1.130 | 0.74 | 2.165 | 0.68 | 3.059 | 2.59 | 2.118 |
| | | | RFF | 20.85 | 2.046 | 12.14 | 3.195 | 11.80 | 1.346 | 14.93 | 2.196 |
| | | | NeFF | 19.68 | 2.127 | 9.20 | 3.438 | 7.74 | 1.597 | 12.21 | 2.387 |
| 拉普拉斯 | $e^{\frac{-\lvert x\rvert}{a}}$ | $[a]$ | 恒等函数 | 12.04 | 0.932 | 1.34 | **1.561** | 1.37 | 2.434 | 4.92 | 1.642 |
| | | | RFF | 15.57 | 2.386 | 10.97 | 2.632 | 14.74 | **1.112** | 13.76 | 2.043 |
| | | | NeFF | 15.26 | 2.434 | 8.67 | 3.191 | 8.16 | 1.526 | 10.70 | 2.384 |
| Sine | 正弦($\omega x$) | $[\omega]$ | 恒等函数 | 13.36 | 0.838 | 7.96 | 1.660 | **7.47** | **1.722** | **9.59** | **1.407** |
| | | | RFF | **39.02** | **0.582** | **13.06** | 1.412 | **16.57** | 1.156 | **22.88** | **1.050** |
| | | | NeFF | **42.39** | **0.537** | **33.58** | **0.914** | **22.02** | **0.696** | **32.66** | **0.716** |
| Incode-Sine | $a$正弦($b\omega x + c$) $+ d$ | $[\omega], a, b, c, d$ | 恒等函数 | **15.98** | **0.778** | **8.16** | 1.611 | 0.01 | 3.865 | 8.05 | 2.085 |
| | | | RFF | 38.10 | 0.595 | 12.86 | 1.559 | 15.13 | 1.241 | 22.03 | 1.132 |
| | | | NeFF | 41.40 | 0.556 | 32.24 | 1.038 | 21.33 | 0.763 | 31.99 | 0.786 |

表 2：坐标多层感知机的基准测试排行榜。对于不同的位置编码（恒等函数、RFF、NeFF），最佳结果以粗体标出，次佳结果加下划线。请注意，“$a$”表示可学习参数，而“$[a]$”表示超参数。其余 10 种激活函数的基准测试结果见附录（附录D）。

倒金字塔频率设置。给定 Fourier-KAN 的深度$L$和宽度，超参数$[\Omega_0, \cdots, \Omega_l, \Omega_L]$决定了每层傅里叶基函数的数量，以及网络学习频率分量的倾向。在网络容量相同的情况下，较大的 Ω 能提高频率分辨率，增强网络捕捉音频信号周期性和波动的能力。与位置编码在坐标多层感知机中的作用类似，倒金字塔频率设置有助于 Fourier-KAN 捕捉高频信息，从而加快收敛。例如，Ω 设置为 $[64, 5, 3]$ 的 3 层 Fourier-KAN 优于 $[8, 8, 8]$,，后者可能导致收敛问题。

# 实验

## 实验设置

**数据集**。**GTZAN**音乐数据集（Tzanetakis 和 Cook，2002年）包含 1000 段时长为 30 秒、涵盖十种流派的音乐片段。**CSTR VCTK** 语音语料库（Yamagishi、Veaux 和 MacDonald，2019）包含 110 位具有不同口音的说话者的语音录音，每人朗读约 400 个句子。在基准测试中，我们使用了 SIREN（Sitzmann 等人 2020）提供的两段 7 秒音频片段（“巴赫”和“计数”），以及一段来自 GTZAN 数据集的 30 秒音频片段（“布鲁斯”）。为了全面评估有效方法的性能，我们从 GTZAN 数据集中选取了十段不同流派的音频片段，并从 CSTR VCTK 数据集中选取了十段带有不同口音的音频片段。

**网络**。 我们确保网络参数具有可比性，范围在 250K 至 270K 之间。对于坐标多层感知机，每个网络的深度为 6，宽度为 256。相比之下，傅里叶ASR网络的深度为 6，宽度为 64，其中输入层、隐藏层和输出层的最大频率阈值分别设为 1024、5 和 3。

**评估指标**。 我们分别使用信噪比（SNR）（Roux 等，2019）和对数谱距离（LSD）（Gray 和 Markel，1976）来评估重构音频的时域误差和频谱误差。由于 LSD 只能间接衡量频域表现，我们主要关注信噪比指标。

## 基准测试排行榜

我们首先考察非线性映射的影响。人们通常认为隐式音频表示中存在非线性映射，但此前尚未对此进行分析。根据公式 3，表 2 展示了使用 16 种不同激活函数和 3 种位置编码（恒等函数、NeFF 和 RFF）进行音频信号表示的评估结果。基于这项全面的基准测试，可以得出以下结论：

- 大多数激活函数（Sigmoid、ReLU、Tanh 等），除具有强非线性（高斯型）和周期性（正弦型）的激活函数外，都无法捕捉音频信号的高频成分和局部周期性。
- 位置编码显著增强了坐标多层感知机表示音频信号的能力，这得益于其高维映射能力，从而提升了模型捕捉高频信息的能力。对于高斯（11.02dB ↑ 信噪比）和正弦（18.96dB ↑信噪比）激活函数而言，这种增强尤为明显。
- 在位置编码的背景下，RFF 引入随机高斯噪声，使其更适用于

| GTZAN 数据集 | | 指标 | blu. | cla. | cou. | dis. | hip. | jaz. | met. | pop. | reg. | roc. | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 基线 | 高斯（多层感知机） | 信噪比 ↑ | 0.68 | 0.25 | 2.64 | 2.40 | 1.02 | 0.15 | 1.66 | 4.61 | 1.33 | 1.06 | 1.58 |
| | | 对数谱距离 ↓ | 3.059 | 3.175 | 3.028 | 3.379 | 3.761 | 3.494 | 3.634 | 3.030 | 3.047 | 3.219 | 3.383 |
| | 正弦（MLP） | 信噪比 ↑ | 7.47 | 2.86 | 5.92 | 7.34 | 3.04 | 6.08 | 4.32 | 8.776 | 4.87 | 6.84 | 5.76 |
| | | 对数谱距离 ↓ | 1.722 | 1.755 | 2.338 | 2.204 | 2.754 | 1.771 | 2.830 | 2.595 | 1.969 | 1.900 | 2.184 |
| | B样条（KAN） | 信噪比 ↑ | 0.00 | 0.01 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 2.07 | 0.00 | 0.01 | 0.23 |
| | | 对数谱距离 ↓ | 4.643 | 4.278 | 6.685 | 5.799 | 4.359 | 4.899 | 6.000 | 3.864 | 7.533 | 4.899 | 5.30 |
| Ours | RFF+高斯（MLP） | 信噪比 ↑ | 11.80 | 10.76 | 11.98 | 12.00 | 11.30 | 12.75 | 11.25 | 12.07 | 11.57 | 11.84 | 11.73 |
| | | 对数谱距离 ↓ | 1.346 | 1.721 | 1.474 | 1.299 | 1.148 | 1.936 | 1.731 | 1.439 | 1.362 | 1.481 | 1.494 |
| | NeFF+正弦（MLP） | 信噪比 ↑ | **22.02** | **25.95** | **16.35** | **17.70** | **13.92** | **19.22** | **13.05** | **15.27** | **17.79** | **19.16** | **18.04** |
| | | 对数谱距离 ↓ | **0.696** | **0.585** | **1.064** | **1.036** | **0.741** | **0.983** | **0.902** | **1.245** | **0.883** | **0.714** | **0.885** |
| | Fourier-ASR（KAN） | 信噪比 ↑ | 13.80 | 15.05 | 12.54 | 12.87 | 12.22 | 13.67 | 12.21 | 13.27 | 12.42 | 12.65 | 12.76 |
| | | 对数谱距离 ↓ | 1.245 | 0.913 | 1.249 | 1.158 | 1.059 | 1.244 | 1.203 | 1.302 | 1.110 | 1.399 | 1.110 |

| CSTR VCTK 数据集 | | 指标 p225 | p234 | p238 | p245 | p248 | p253 | p335 | p345 | p363 | p374 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 基线 | 高斯（MLP） | 信噪比 ↑ | 1.88 | 2.09 | 1.16 | 4.06 | 0.23 | 2.39 | 1.37 | 3.06 | 5.56 | 2.25 | 2.41 |
| | | 对数谱距离 ↓ | 2.126 | 2.034 | 2.557 | 1.831 | 2.884 | 2.065 | 2.277 | 1.896 | 1.791 | 1.827 | 2.129 |
| | 正弦（MLP） | 信噪比 ↑ | 14.86 | 10.88 | 12.38 | 14.41 | 10.32 | 13.85 | 9.61 | 15.89 | 12.78 | 12.53 | 12.75 |
| | | 对数谱距离 ↓ | 1.743 | 1.588 | 1.748 | 1.665 | 1.630 | 1.672 | 1.619 | 1.556 | 1.716 | 1.500 | 1.644 |
| | B样条（KAN） | 信噪比 ↑ | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 | 0.02 | 0.02 | 0.05 | 0.11 | 0.03 |
| | | 对数谱距离 ↓ | 3.312 | 3.113 | 3.317 | 3.151 | 3.506 | 3.160 | 3.000 | 2.957 | 2.705 | 2.631 | 3.085 |
| Ours | RFF+高斯（MLP） | 信噪比 ↑ | 11.67 | 12.93 | 16.19 | 11.99 | 15.52 | 12.21 | 13.32 | 15.95 | 12.79 | 12.28 | 12.81 |
| | | 对数谱距离 ↓ | 2.401 | 2.128 | 1.789 | 2.218 | 2.183 | 2.258 | 2.076 | 1.704 | 2.012 | 2.059 | 1.983 |
| | NeFF+正弦（MLP） | 信噪比 ↑ | **25.20** | **31.63** | **19.56** | **32.03** | **27.00** | **27.11** | **16.87** | **28.38** | **29.25** | **30.83** | **26.79** |
| | | 对数谱距离 ↓ | **1.015** | **0.734** | **1.235** | **0.866** | **1.134** | **0.917** | **1.207** | **1.032** | **0.753** | **0.877** | **0.977** |
| | Fourier-ASR（KAN） | 信噪比 ↑ | 18.30 | 20.68 | 17.12 | 18.26 | 21.34 | 17.40 | 15.79 | 17.34 | 17.86 | 20.20 | 18.43 |
| | | 对数谱距离 ↓ | 1.495 | 1.228 | 1.615 | 1.310 | 1.464 | 1.397 | 1.456 | 1.417 | 1.321 | 1.267 | 1.397 |

表 3：在 GTZAN 和 CSTR VCTK 数据集上对 Fourier-ASR 和新型非线性映射设计进行评估。

高斯型激活函数（≈ 3dB ↑信噪比）。相反，NeFF 采用傅里叶映射，与正弦型激活函数（≈9dB ↑信噪比）更兼容。

## Fourier-ASR 与新设计的评估

根据表 2 所示的基准测试排行榜，我们选取了有效的非线性映射，与 Fourier-ASR 在 GTZAN（Tzanetakis 和 Cook，2002）（Yamagishi、Veaux 和 MacDonald，2019）和 CSTR VCTK（Yamagishi、Veaux 和 MacDonald，2019）数据集上进行比较。值得注意的是，尽管高斯（Ramasinghe 和 Lucey 2022）和正弦（Sitzmann 等人 2020）激活函数旨在缓解位置编码相关的复杂参数调整问题和频谱偏差，但我们发现，由于音频信号具有高频特性和局部周期性，位置编码仍然不可或缺。因此，我们设计了**新型非线性映射**，即 RFF+高斯和 NeFF+正弦，以应对这些挑战。

如表 3 所示，RFF+高斯和 NeFF+正弦这两种设计显著增强了坐标多层感知机表示音频信号的能力。在 GTZAN 数据集上，这些方法分别将信噪比提高了 10.15dB ↑和 12.28dB ↑。在 CSTR VCTK 数据集上，信噪比分别提高了 10.40dB ↑和 14.04dB ↑。得益于傅里叶基函数的周期性以及频率自适应学习策略（FaLS），我们提出的 Fourier-ASR（KAN）显著优于正弦（多层感知机）（≈ 6dB ↑）和 B样条（KAN）（≈ 18dB ↑）。然而，由于现有优化策略并未完全适配 KAN网络（Liu 等人，2024），Fourier-ASR（KAN）的表现略逊于具有局部周期性的 NeFF+正弦（多层感知机）。尽管如此，Fourier-ASR（KAN）无需位置编码，因此避免了繁琐的超参数调优。

## 结论与未来工作

我们提出了首个用于评估基于坐标多层感知机的隐式神经音频信号表示的开源基准测试，填补了标准化性能评估方面的关键空白。我们证明了在连续音频表示领域中，结合位置编码与激活函数的非线性映射设计是有效的。此外，我们提出了一种新颖的音频信号表示框架 Fourier-ASR，它融合了傅里叶级数定理与柯尔莫哥洛夫-阿诺德表示定理，具有更强的可解释性和更稳定的表示能力。我们的工作不仅为基于坐标多层感知机的音频信号表示的组件选择提供指导，也推动了音频表示应用的发展。得益于隐式神经表示的优异特性，例如连续可微性以及与空间分辨率的解耦，我们的工作可有效应用于音频超分辨率、去噪、压缩和生成等下游任务。

## 致谢

本研究部分得到了中国国家自然科学基金项目（批准号：62272343）的资助；部分得到了上海市教育发展基金会和上海市教育委员会曙光计划项目（批准号：21SG23）的资助；部分得到了中央高校基本科研业务费的资助。

## 参考文献

Arnold, V. I. 1957。《关于三元函数》。苏联科学院院报，114：679–681。

Barron, J. T.；Mildenhall, B.；Verbin, D.；Srinivasan, P. P.；以及 Hedman, P. 2022。Mip-NeRF 360：无界抗混叠神经辐射场。载于*IEEE*计算机视觉与模式识别会议论文集，5460–5469。美国路易斯安那州新奥尔良：IEEE。

Chen, A.；Xu, Z.；Geiger, A.；Yu, J.；以及 Su, H. 2022。Ten-soRF：张量辐射场。载于欧洲计算机视觉会议论文集，333–350。柏林、海德堡：施普林格出版社。

Chng, S.-F.；Ramasinghe, S.；Sherrah, J.；以及 Lucey, S. 2022。用于高保真重建与位姿估计的高斯激活神经辐射场。载于欧洲计算机视觉会议论文集，264–280。柏林、海德堡：施普林格出版社。

Davidson, K. R.；和 Donsig, A. P. 2010年。傅里叶级数与逼近，360–405。纽约州纽约：施普林格纽约。

de Boor, C. 1978年。《样条实用指南》。载于应用数学科学。

Glorot, X.；和 Bengio, Y. 2010年。理解训练深度前馈神经网络的困难。载于第十三届国际人工智能与统计会议论文集，第9卷，249–256。撒丁岛基亚拉古纳度假村，意大利：PMLR。

Gray, A.；和 Markel, J. 1976年。语音处理的距离度量。*IEEE*语音与音频处理汇刊，24(5)：380–391。

He, K.；Zhang, X.；Ren, S.；和 Sun, J. 2015年。深入研究整流器：在 ImageNet 分类上超越人类水平的性能。载于*IEEE/CVF*国际计算机视觉会议论文集，1026–1034。智利圣地亚哥：IEEE。

Hertz, A.; Perel, O.; Giryes, R.; Sorkine-Hornung, O.; and Cohen-Or, D. 2021年。SAPE：用于神经优化的空间自适应渐进编码。载于第*35*届神经信息处理系统国际会议论文集，第34卷，8820–8832页。美国纽约州雷德胡克：Curran Associates Inc.

Hornik, K.; Stinchcombe, M.; and White, H. 1989。多层前馈网络是通用逼近器。神经网络，2(5)：359–366。

Kazerouni, A.; Azad, R.; Hosseini, A.; Merhof, D.; and Bagci,U.2024。INCODE：结合先验知识嵌入的隐式神经条件化。载于论文集

*IEEE*冬季计算机视觉应用会议，1287–1296。美国：电气与电子工程师协会公司。Klocek, S.; Maziarka, L.; Wołczyk, M.; Tabor, J.; Nowak, J.; and undefinedmieja, M. 2019. 超网络函数图像表示。载于人工神经网络与机器学习论文集，496–510。柏林，海德堡：施普林格出版社。Kolmogorov, A. N. 1956. 关于用较少变量的连续函数叠加表示多个变量的连续函数。苏联科学院论文集，108: 179–182. Lapedes, A.; and Farber, R. 1987. 神经网络如何工作。载于*1987*年神经信息处理系统国际会议论文集，NIPS'87，442–456。美国马萨诸塞州剑桥：麻省理工学院出版社。Lindell, D. B.; Van Veen, D.; Park, J. J.; and Wetzstein, G. 2022. Bacon：用于多尺度场景表示的带限坐标网络。载于*IEEE*计算机视觉与模式识别会议论文集，16231–16241。美国路易斯安那州新奥尔良：IEEE。Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Solja˘ci´c, M.; Hou, T. Y.; and Tegmark, M. 2024. KAN：Kolmogorov-Arnold网络。arXiv:2404.19756. Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. 占据网络：在函数空间中学习三维重建。载于*IEEE*计算机视觉与模式识别会议论文集，4455–4465。美国加利福尼亚州长滩：IEEE。Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF：将场景表示为用于视图合成的神经辐射场。载于欧洲计算机视觉会议论文集，405–421。Cham：Springer国际出版公司。M¨uller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. 采用多分辨率哈希编码的即时神经图形基元。《*ACM*图形学汇刊》，41(4): 102:1–102:15. Parascandolo, G.; Huttunen, H.; and Virtanen, T. 2017. 驯服波形：正弦函数作为深度神经网络的激活函数。Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF：为形状表示学习连续符号距离函数。载于*IEEE*计算机视觉与模式识别会议论文集，165–174。美国加利福尼亚州长滩：IEEE。Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.; Bengio, Y.; and Courville, A. 2019. 关于神经网络的频谱偏差。载于第*36*届国际机器学习会议论文集，第97卷机器学习研究论文集，5301–5310. PMLR。Ramasinghe, S.; and Lucey, S. 2022. 超越周期性：迈向坐标多层感知机激活函数的统一框架。载于欧洲计算机视觉会议论文集，142–158。Cham：Springer Nature瑞士。

Ramasinghe, S.; MacDonald, L.; and Lucey, S. 2022. 关于正则化坐标多层感知机。arXiv:2202.00790. Roux, J. L.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR——半生不熟还是恰到好处？载于*IEEE*国际声学、语音与信号处理会议论文集, 626–630. 英国布莱顿：IEEE。Rudelson, M.; and Vershynin, R. 2007. 从大型矩阵中采样：一种基于几何泛函分析的方法。《*ACM*期刊》, 54(4): 21–es. Saragadam, V.; LeJeune, D.; Tan, J.; Balakrishnan, G.; Veer-araghavan, A.; and Baraniuk, R. G. 2023. WIRE：小波隐式神经表示。载于*IEEE*计算机视觉与模式识别会议论文集, 18507–18516. 加拿大不列颠哥伦比亚省温哥华：IEEE。Sitzmann, V.; Martel, J. N. P.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020. 具有周期性激活函数的隐式神经表示。载于第*34*届神经信息处理系统国际会议论文集, 第33卷, 7462–7473. 美国纽约州雷德胡克：Curran Associates Inc. Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. 傅里叶特征使网络能够学习低维域中的高频函数。载于第*34*届神经信息处理系统国际会议论文集, 第33卷, 7537–7547. 美国纽约州雷德胡克：Curran Associates, Inc. Tzanetakis, G.; and Cook, P. 2002年. 音频信号的音乐流派分类。*IEEE*语音与音频处理汇刊, 10(5): 293–302. Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021年. NeuS：通过体渲染学习用于多视图重建的神经隐式表面。载于第*35*届神经信息处理系统国际会议论文集, 第34卷, 27171–27183. 美国纽约州雷德胡克：Curran Associates Inc. Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK语料库：供CSTR语音克隆工具包使用的英语多说话人语料库（版本0.92）。https://doi.org/10. 7488/ds/2645. Yariv, L.; Hedman, P.; Reiser, C.; Verbin, D.; Srinivasan, P. P.; Szeliski, R.; Barron, J. T.; and Mildenhall, B. 2023. BakedSDF：为实时视图合成构建神经符号距离函数网格。载于*ACM SIGGRAPH 2023*会议论文集。美国纽约州纽约：计算机协会。Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixel-NeRF：由一张或少数几张图像生成神经辐射场。载于*IEEE*计算机视觉与模式识别会议论文集, 4576–4585. 美国田纳西州纳什维尔：IEEE。Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF：探索用于神经隐式表面重建的单目几何线索。载于第*36*届神经信息处理系统国际会议论文集, 第35卷, 25018–25032. 美国纽约州雷德胡克：Curran Associates Inc.

## 目录

A. 坐标多层感知机的性质 A.1 位置编码对参数敏感 A.2 激活函数对参数敏感 A.3 周期性激活函数对初始化方案敏感 B. 记忆 B.1 坐标多层感知机的记忆化 B.2 Fourier-KAN的记忆化 B.3 参数利用率比较 C. 可解释性 C.1 坐标多层感知机的可解释性 C.2 Fourier-KAN的可解释性 D. 进一步实验 D.1 实验设置 D.2 基准排行榜补充 D.3 定性实验 参考文献

# A. 坐标多层感知机的性质

## A.1 位置编码对参数敏感

**分析**。 关于隐式神经表示的研究表明，采用 ReLU 等激活函数的网络会表现出频谱偏差（Rahaman等，2019）。这些网络往往首先学习低频函数，因为这类函数表现出全局行为，并且优化时更稳定。然而，频谱偏差也会妨碍网络学习高频信息的能力。为解决这一问题，NeRF（Mildenhall等，2020）提出将输入坐标映射到高维空间，以便在辐射场中学习高频信息。具体而言，给定输入坐标 $t \in \mathbb{R}$，NeRF将其映射到 $\mathbb{R}^{2L}$ 空间，如下所示，

$$\gamma_{\text{NeFF}}(t) = \left[\sin(2^0\pi t), \cos(2^0\pi t), \ldots, \sin(2^{L-1}\pi t), \cos(2^{L-1}\pi t)\right] \tag{10}$$

其中， $L$ 表示高频空间的维数。我们将这种位置编码方案称为 NeFF（NeRF 傅里叶特征）。

根据神经切线核（NTK）理论，FFN（Tancik等，2020）在 NeRF 位置编码中引入高斯随机噪声（称为 RFF，即随机傅里叶特征），如下所示:

$$\gamma_{\text{RFF}}(t) = \left[\cos(2\pi\mathbf{b}_1^{\mathrm{T}}t), \sin(2\pi\mathbf{b}_1^{\mathrm{T}}t), \ldots, \cos(2\pi\mathbf{b}_L^{\mathrm{T}}t), \sin(2\pi\mathbf{b}_L^{\mathrm{T}}t)\right]^{\mathrm{T}}, \quad \mathbf{b}_i \sim \mathcal{N}(0, \sigma^2), \tag{11}$$

其中，频率向量 $\mathbf{b}_i$ 从标准差为 $\sigma$ （超参数）的高斯分布中随机采样。该映射将 NTK 转换为平稳（平移不变）核，并通过修改频率向量 $\mathbf{b}_i$来调节 NTK 的频谱，从而控制相应多层感知机能够学习的频率范围。

这些编码方案有助于学习复杂函数，但也引入了第二个现象：网络频谱偏差的减弱（Ramasinghe、MacDonald和Lucey，2022年）。这种情况具有两方面的影响。位置编码可能导致神经优化过程不稳定，从而可能收敛到次优的局部极小值。此外，在拟合具有不同细节程度或频率的函数时，负责预测平滑低频区域的神经元仍会接触到高频编码维度。这种接触会使学习过程变得复杂，因为网络必须学习何时以及如何忽略这些嵌入维度。事实上，FFN（Tancik等，2020年）认为，应谨慎选择 RFF位置编码中频率分布的标准差 $\sigma$ ，并指出，较低的取值可能导致网络输出缺乏细节，而过高的取值则可能引入噪声伪影。因此，对于这种静态编码， $\sigma$ 的取值需要针对每个样本进行微调。

**实验**。表 4 展示了 NeFF（公式 10）位置编码的维度参数 $L$ 的敏感性实验。显然， $L$ 的最优值会因所使用的激活函数而异。此外，值得注意的是，最优参数与次优参数的音频重建结果存在显著差异。例如，使用 ReLU激活函数时，最优参数 $L = 16$ 的信噪比为 24.76 dB，而次优参数 $L = 32$ 的信噪比仅为 5.37 dB。这种显著差异使参数调优过程更加复杂。

表 5 和表 6 分别展示了 RFF 位置编码对维数参数 $L$ 和方差参数 $\sigma$的敏感性实验。结果表明，从最优、次优和第三优配置的分布可以看出，RFF 对维数参数 $L$ 更为敏感。对于高斯激活函数，最优信噪比在 $L = 256$ 处取得，尽管这也会导致参数量更高。对于 ReLU 激活函数，最优值为 $L = 32$。相比之下，对于周期性激活函数（例如，正弦、Incode-Sine），最优值为 $L = 2$，但最优参数与次优参数之间的重建质量存在显著差异。

根据表 6，尽管能获得最高信噪比的方差参数 $\sigma$ 通常在 1000 左右，但需要注意的是，较高的 $\sigma$ 并不一定会带来更好的结果。事实上， $\sigma$ 的次优取值往往小于 1000。这种变化使参数调优过程更加复杂，难以确定最优设置。

## A.2 激活函数对参数敏感性

**分析**。 我们的基准测试排行榜表明，大多数激活函数无法捕捉音频信号的高频和局部周期性特征。然而，有效的激活函数会采用一个缩放因子超参数，以表示不同频率的信号。如表7所示，高斯型激活函数使用方差因子 $a$，而正弦型激活

| 数据集 | 参数 $L$ / 参数 | | 2<br>264K | 4<br>265K | 8<br>268K | 16<br>272K | 32<br>280K | 64<br>296K |
|---|---|---|---|---|---|---|---|---|
| 巴赫（7秒） | 高斯 | 信噪比 ↑ | 16.12 | **29.08** | 25.16 | 19.86 | 21.06 | *23.48* |
| | | 对数谱距离 ↓ | 0.810 | **0.737** | *1.785* | 2.115 | 2.035 | 1.880 |
| | ReLU | 信噪比 ↑ | 0.00 | 0.02 | 4.60 | **24.76** | 5.37 | *4.89* |
| | | 对数谱距离 ↓ | 3.997 | 3.590 | **1.151** | 1.590 | *2.812* | 3.000 |
| | Sine | 信噪比 ↑ | 12.12 | 14.75 | *34.63* | **46.40** | 34.36 | 37.62 |
| | | 对数谱距离 ↓ | 0.914 | *0.864* | 0.663 | **0.530** | 1.328 | 1.188 |
| | Incode-Sine | 信噪比 ↑ | 14.25 | 15.70 | 33.38 | 36.94 | *33.46* | **37.53** |
| | | 对数谱距离 ↓ | 0.814 | *0.837* | **0.663** | 1.120 | 1.332 | 1.145 |
| 计数（7秒） | 高斯 | 信噪比 ↑ | 7.50 | 9.89 | **19.31** | 9.36 | 10.71 | *10.66* |
| | | 对数谱距离 ↓ | 1.760 | **1.567** | *2.589* | 3.429 | 3.323 | 3.221 |
| | ReLU | 信噪比 ↑ | 0.00 | 0.01 | 0.59 | **13.27** | 6.64 | *6.47* |
| | | 对数谱距离 ↓ | 3.548 | 3.256 | 1.938 | **1.326** | *2.618* | 2.812 |
| | Sine | 信噪比 ↑ | 7.04 | 10.36 | 12.73 | **35.46** | *24.16* | 27.08 |
| | | 对数谱距离 ↓ | *1.723* | 1.803 | 1.713 | **1.377** | 2.216 | 1.979 |
| | Incode-Sine | 信噪比 ↑ | 8.65 | 11.03 | 12.55 | **28.40** | *13.06* | 14.14 |
| | | 对数谱距离 ↓ | **1.649** | *1.784* | 1.752 | 1.855 | 3.145 | 3.056 |
| 布鲁斯（30秒） | 高斯 | 信噪比 ↑ | 3.12 | 8.84 | **19.82** | 7.81 | 8.50 | *8.78* |
| | | 对数谱距离 ↓ | 2.057 | 1.580 | **0.849** | 1.594 | *1.553* | 1.536 |
| | ReLU | 信噪比 ↑ | 0.00 | 0.01 | 0.67 | **14.74** | *2.58* | 2.81 |
| | | 对数谱距离 ↓ | 5.140 | 4.708 | 2.843 | **0.890** | 1.730 | *1.760* |
| | Sine | 信噪比 ↑ | 7.39 | 9.76 | 12.96 | **24.08** | 10.80 | *11.68* |
| | | 对数谱距离 ↓ | 1.697 | 1.578 | 1.360 | **0.606** | 1.422 | *1.370* |
| | Incode-Sine | 信噪比 ↑ | 6.81 | 8.22 | 11.90 | **22.86** | 10.26 | *11.09* |
| | | 对数谱距离 ↓ | 1.751 | 1.667 | *1.449* | **0.670** | 1.454 | 1.405 |

表 4：NeFF 位置编码（公式 10）中的参数 $L$ 对音频表示质量的影响。每行中，最佳、次佳和第三佳结果分别以**第一**、第二和第三标示。

| 数据集 | 参数 $L$ / 参数量 | | 2<br>264K | 4<br>265K | 8<br>267K | 16<br>271K | 32<br>280K | 64<br>296K | 128<br>329K | 256<br>394K |
|---|---|---|---|---|---|---|---|---|---|---|
| 巴赫（7秒） | 高斯 | 信噪比 ↑ | 18.61 | 19.41 | 19.67 | 20.21 | 20.84 | *23.89* | 27.06 | **30.18** |
| | | 对数谱距离 ↓ | 2.199 | 2.143 | 2.125 | 2.090 | 2.046 | *1.850* | 1.680 | **1.527** |
| | ReLU | 信噪比 ↑ | 0.82 | 13.44 | 14.57 | 14.87 | **15.62** | 14.55 | *14.94* | 15.08 |
| | | 对数谱距离 ↓ | 1.580 | 1.088 | 1.019 | 0.988 | 0.978 | *0.951* | 0.932 | **0.909** |
| | Sine | 信噪比 ↑ | **43.24** | 41.97 | *41.60* | 40.04 | 39.02 | 38.27 | 38.37 | 38.82 |
| | | 对数谱距离 ↓ | 0.778 | 0.548 | **0.547** | *0.561* | 0.582 | 0.595 | 0.599 | 0.580 |
| | Incode-Sine | 信噪比 ↑ | **41.82** | 40.01 | *39.84* | 39.02 | 38.10 | 37.17 | 36.71 | 36.94 |
| | | 对数谱距离 ↓ | 0.877 | **0.567** | 0.568 | *0.575* | 0.595 | 0.614 | 0.627 | 0.610 |
| 计数（7秒） | 高斯 | 信噪比 ↑ | 8.73 | 9.37 | 9.69 | 10.33 | 12.14 | **20.09** | 16.78 | *16.34* |
| | | 对数谱距离 ↓ | 3.474 | 3.424 | 3.400 | 3.347 | 3.195 | 2.335 | **2.119** | *2.445* |
| | ReLU | 信噪比 ↑ | 0.74 | 4.09 | 4.52 | *4.48* | **4.93** | 4.22 | 4.28 | 4.42 |
| | | 对数谱距离 ↓ | 1.690 | 1.489 | 1.430 | 1.412 | 1.400 | *1.379* | 1.361 | **1.354** |
| | Sine | 信噪比 ↑ | **28.64** | 25.76 | *17.62* | 13.50 | 13.05 | 12.97 | 13.02 | 12.92 |
| | | 对数谱距离 ↓ | 1.396 | **1.077** | 1.184 | *1.253* | 1.412 | 1.439 | 1.434 | 1.571 |
| | Incode-Sine | 信噪比 ↑ | **27.59** | 19.26 | *13.34* | 12.95 | 12.86 | 12.83 | 12.81 | 12.80 |
| | | 对数谱距离 ↓ | *1.332* | **1.203** | 1.269 | 1.472 | 1.558 | 1.649 | 1.648 | 1.682 |
| 布鲁斯（30秒） | 高斯 | 信噪比 ↑ | 7.74 | 8.57 | 8.89 | 9.66 | 11.74 | *18.06* | 21.98 | **23.16** |
| | | 对数谱距离 ↓ | 1.593 | 1.541 | 1.522 | 1.475 | 1.348 | *0.975* | 0.763 | **0.672** |
| | ReLU | 信噪比 ↑ | 0.56 | 2.57 | 2.88 | 2.91 | **3.23** | 2.83 | 2.92 | *2.90* |
| | | 对数谱距离 ↓ | 2.343 | 1.964 | 1.913 | 1.920 | **1.855** | 1.950 | 1.920 | *1.919* |
| | Sine | 信噪比 ↑ | **21.13** | *19.91* | 19.01 | 17.68 | 17.08 | 17.29 | 18.35 | 20.49 |
| | | 对数谱距离 ↓ | **0.742** | 0.908 | *0.986* | 1.078 | 1.130 | 1.103 | 1.035 | 0.874 |
| | Incode-Sine | 信噪比 ↑ | **21.09** | 19.48 | *18.37* | 16.77 | 15.60 | 15.11 | 14.14 | 16.71 |
| | | 对数谱距离 ↓ | **0.765** | 0.953 | *1.047* | 1.445 | 1.215 | 1.123 | 1.172 | 1.137 |

表5：RFF位置编码（公式11）中的参数 $L$ 对音频表示质量的影响。

| 数据集 | 参数 $\sqrt{\sigma}$ | | 1 | 20 | 40 | 60 | 80 | 100 | 1000 | 10000 |
|---|---|---|---|---|---|---|---|---|---|---|
| 巴赫（7秒） | 高斯 | 信噪比 ↑ | 19.02 | **27.72** | 23.21 | *21.68* | 21.12 | 20.84 | 20.58 | 20.63 |
| | | 对数谱距离 ↓ | **0.801** | 1.628 | *1.899* | 1.993 | 2.028 | 2.046 | 2.064 | 2.063 |
| | ReLU | 信噪比 ↑ | 0.01 | 4.42 | 9.69 | 11.91 | *13.95* | 15.62 | **30.30** | 5.99 |
| | | 对数谱距离 ↓ | 3.968 | 1.165 | 0.988 | 0.977 | **0.970** | *0.978* | 1.374 | 3.011 |
| | Sine | 信噪比 ↑ | 12.00 | 28.38 | 34.43 | 36.92 | *38.16* | 39.02 | **45.27** | 33.23 |
| | | 对数谱距离 ↓ | 0.920 | 0.661 | 0.636 | 0.612 | *0.592* | 0.582 | **0.517** | 1.344 |
| | Incode-Sine | 信噪比 ↑ | 13.71 | 24.91 | 33.03 | 36.04 | *37.30* | 38.10 | **46.35** | 33.49 |
| | | 对数谱距离 ↓ | 0.859 | 0.694 | 0.631 | 0.624 | *0.607* | 0.595 | **0.516** | 1.330 |
| 计数（7秒） | 高斯 | 信噪比 ↑ | 8.12 | **20.95** | 16.45 | *14.11* | 13.24 | 12.14 | 10.31 | 10.37 |
| | | 对数谱距离 ↓ | **1.651** | 2.327 | *2.849* | 3.041 | 3.108 | 3.195 | 3.353 | 3.349 |
| | ReLU | 信噪比 ↑ | 0.00 | 0.589 | 1.40 | 2.61 | 3.82 | *4.93* | **13.21** | 8.58 |
| | | 对数谱距离 ↓ | 3.417 | 1.964 | 1.630 | *1.466* | 1.402 | **1.400** | 1.939 | 3.211 |
| | Sine | 信噪比 ↑ | 6.29 | 12.63 | 12.75 | 12.85 | 12.91 | *13.06* | **35.13** | 22.59 |
| | | 对数谱距离 ↓ | 1.666 | 1.706 | 1.685 | 1.640 | *1.532* | 1.412 | **1.113** | 2.364 |
| | Incode-Sine | 信噪比 ↑ | 6.23 | 12.13 | 12.56 | 12.78 | 12.84 | *12.86* | **34.65** | 16.09 |
| | | 对数谱距离 ↓ | 1.598 | 1.762 | 1.702 | 1.652 | *1.585* | 1.559 | **1.114** | 2.890 |
| 布鲁斯（30秒） | 高斯 | 信噪比 ↑ | 4.85 | **21.37** | 17.33 | *14.19* | 12.59 | 11.74 | 8.56 | 8.57 |
| | | 对数谱距离 ↓ | 1.876 | **0.920** | 1.026 | 1.211 | *1.302* | 1.348 | 1.549 | 1.549 |
| | ReLU | 信噪比 ↑ | 0.00 | 0.66 | 1.39 | 2.08 | 2.68 | *3.23* | **13.58** | 9.05 |
| | | 对数谱距离 ↓ | 4.893 | 2.826 | 2.386 | 2.136 | 1.986 | *1.855* | **0.949** | 1.374 |
| | Sine | 信噪比 ↑ | 4.91 | 15.21 | 14.29 | 15.40 | 16.22 | 17.09 | **23.53** | *16.78* |
| | | 对数谱距离 ↓ | 1.895 | 1.235 | 1.267 | 1.215 | 1.171 | *1.130* | **0.583** | 1.084 |
| | Incode-Sine | 信噪比 ↑ | 3.866 | 10.75 | 12.08 | 13.14 | *14.32* | 15.60 | **22.86** | 13.88 |
| | | 对数谱距离 ↓ | 1.977 | 1.542 | 1.426 | 1.366 | 1.277 | 1.215 | **0.619** | *1.250* |

表 6：RFF位置编码（公式 11）中的参数 $\sigma$ 对音频表示质量的影响。每行中，最佳结果以**第**一、第二和第三名突出显示。

函数使用频率因子 $\omega$。这两种类型都需要谨慎进行参数调优，才能有效表示音频信号。

| 激活函数 | 方程 | 参数 |
|---|---|---|
| 高斯 | $e^{\frac{-x^2}{2a^2}}$ | $[a]$ |
| 拉普拉斯 | $e^{\frac{-\lvert x\rvert}{a}}$ | $[a]$ |
| 超高斯 | $[e^{\frac{-x^2}{2a^2}}]^b$ | $[a, b]$ |
| Sine | $\sin(\omega x)$ | $[\omega]$ |
| Incode-Sine | $a\sin(b\omega x + c) + d$ | $[\omega], a, b, c, d$ |

表7：带超参数的激活函数。

**实验**。表8展示了高斯激活函数对方差参数 $a$ 的敏感性实验。结果表明，高斯函数的方差会显著影响音频信号的表示，因此不同音频信号对应的最优参数 $a$ 值各不相同。这种差异使参数调优变得复杂，增加了调整过程的难度。

表9展示了针对不同位置编码设置，评估周期性激活函数 Sine 和 Incode-Sine 对频率超参数 $\omega$ 敏感性的实验结果。结果清楚地表明，位置编码降低了周期性激活函数对频率的敏感性，最优频率参数趋近于 30。这在一定程度上可归因于周期性激活函数采用的频率相关初始化策略（见 A.3），

| 数据集 | 参数 $a$ | 0.01 | 0.1 | 0.2 | 0.3 | 1.0 |
|---|---|---|---|---|---|---|
| 巴赫（7秒） | 信噪比 ↑ | 0.01 | **6.35** | 0.04 | 0.00 | -0.01 |
| | 对数谱距离 ↓ | 1.799 | **1.130** | 3.851 | 4.120 | 6.266 |
| 计数（7秒） | 信噪比 ↑ | **0.74** | 0.34 | 0.01 | 0.00 | -0.01 |
| | 对数谱距离 ↓ | **1.889** | 2.165 | 3.334 | 4.107 | 5.169 |

表8：方差参数 $a$ 对高斯激活函数的影响。每行中，最佳结果以**第**一项突出显示。

该策略使最优频率与初始化策略所用频率接近。此外，Incode-Sine 通过将 Sine 函数中的参数设为可学习参数，降低了 Sine 激活函数对频率的敏感性。

## A.3 周期性激活函数对初始化方案敏感

**分析**。 自1987年以来，周期性激活函数得到了广泛研究，最初源于 Lapedes 和 Farber 的工作（Lapedes 和 Farber，1987年）。该研究表明，使用此类激活函数的网络通常难以训练。Parascandolo 等人（Parascandolo、Huttunen 和 Virtanen，2017年）进一步探讨了使用周期性激活函数训练网络所面临的挑战。他们

| 数据集 | 参数 $\omega$ | | 3 | 30 | 300 | 3,000 | 30,000 | 300,000 |
|---|---|---|---|---|---|---|---|---|
| 巴赫（7秒） | Incode-Sine | 信噪比 ↑ | -1.0 | **15.98** | 0.00 | -0.01 | -0.02 | -0.01 |
| | | 对数谱距离 ↓ | 6.341 | **0.778** | 2.441 | 2.442 | 2.432 | 2.466 |
| | RFF+Incode-Sine | 信噪比 ↑ | 10.97 | **38.10** | 0.05 | -0.02 | -0.01 | 0.01 |
| | | 对数谱距离 ↓ | 0.997 | **0.595** | 2.465 | 2.389 | 2.491 | 2.483 |
| | NeFF+Incode-Sine | 信噪比 ↑ | 0.00 | **15.70** | 0.00 | -0.01 | 0.00 | -0.02 |
| | | 对数谱距离 ↓ | 5.685 | **0.837** | 2.442 | 2.446 | 2.464 | 2.419 |
| | Sine | 信噪比 ↑ | 0.30 | 12.27 | 23.96 | 39.59 | **40.36** | 0.01 |
| | | 对数谱距离 ↓ | 2.478 | 0.885 | 0.669 | **0.575** | 0.683 | 2.526 |
| | NeFF+正弦 | 信噪比 ↑ | 4.13 | **42.39** | 0.80 | 0.70 | 1.12 | 0.00 |
| | | 对数谱距离 ↓ | 1.348 | **0.537** | 2.537 | 2.589 | 2.523 | 2.678 |
| 计数（7秒） | Incode-Sine | 信噪比 ↑ | -0.05 | **8.16** | -0.02 | -0.02 | -0.06 | -0.01 |
| | | 对数谱距离 ↓ | 5.184 | **1.611** | 2.116 | 2.100 | 2.126 | 2.139 |
| | RFF+Incode-Sine | 信噪比 ↑ | 4.04 | **12.86** | -0.09 | -0.07 | -0.08 | -0.03 |
| | | 对数谱距离 ↓ | 1.730 | **1.559** | 2.138 | 2.114 | 2.106 | 2.125 |
| | NeFF+Incode-Sine | 信噪比 ↑ | -0.03 | **11.03** | -0.09 | -0.07 | -0.06 | -0.01 |
| | | 对数谱距离 ↓ | 5.167 | **1.784** | 2.104 | 2.126 | 2.121 | 2.114 |
| | Sine | 信噪比 ↑ | 0.26 | 7.98 | 11.79 | **12.94** | 0.40 | -0.01 |
| | | 对数谱距离 ↓ | 2.153 | 1.716 | 1.710 | **1.533** | 2.120 | 2.129 |
| | NeFF+正弦 | 信噪比 ↑ | 4.81 | **33.57** | 0.82 | 0.89 | 0.27 | 0.06 |
| | | 对数谱距离 ↓ | 1.849 | **0.914** | 2.135 | 2.124 | 2.145 | 2.123 |

表9：频率参数 $\omega$ 对正弦和 Incode-Sine 激活函数的影响。每行中的最佳结果以**首位**标出。

指出，只有当网络不依赖给定函数的周期性时，训练才能成功，并建议使用截断正弦函数。受离散余弦变换的启发，Klocek 等人（Klocek 等人，2019年）提出使用余弦激活函数，以构建权重由超网络决定的网络。

近期，SIREN（Sitzmann 等人 2020）表明，采用正弦激活函数的多层感知机天然适合编码高频信号，因此无需位置编码层。尽管坐标多层感知机颇具潜力，但相关研究往往更关注位置编码，而非正弦激活函数。这一趋势可能有两个主要原因。首先，Sitzmann 等人主要将正弦激活函数的成功归因于其周期性，尽管目前缺乏充分证据支持这种关系。对这一点缺乏理解，掩盖了其有效性的一些基本原理，阻碍了它们在更广泛场景中的可靠应用（Ramasinghe 和 Lucey 2022）。其次，采用正弦激活函数的多层感知机性能对初始化高度敏感。如果多层感知机的初始化未严格遵循 Sitzmann 等人提出的准则，性能可能会显著下降。

具体而言，SIREN（Sitzmann 等人 2020）和 INCODE（Kazerouni 等人 2024）等采用正弦激活函数（$\sigma(x) = \sin(\omega x)$）的模型，必须采用**Sitzmann 初始化**方案。对于权重为$\mathbf{w} \in \mathbb{R}^n$的 $n$ 输入$\mathbf{x} \in \mathbb{R}^n$ 的线性组合，其输出为 $y = \sin\left(\omega \mathbf{w}^T \mathbf{x} + \mathbf{b}\right)$。如果神经元位于输入层，$\mathbf{w}$ 的每个分量都必须服从定义在 $[-1, 1]$ 上的均匀分布，

$$w_i \sim \mathcal{U}(-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}).$$

对于不在输入层的神经元，$\mathbf{w}$的每个分量都必须服从与频率 $\omega$ 相关的分布，

$$w_i \sim \mathcal{U}(-\frac{c}{\omega\sqrt{n}}, \frac{c}{\omega\sqrt{n}}),$$

其中，$c$ 是一个额外参数，SIREN 将其设为 6。

由于该初始化方案依赖频率超参数 $\omega$，且频率参数缺乏泛化能力，正弦型激活函数对初始化方案非常敏感。

| 初始化 | Bach | | 计数 | | 布鲁斯 | |
|---|---|---|---|---|---|---|
| | 信噪比 ↑ | 对数谱距离 ↓ | 信噪比 ↑ | 对数谱距离 ↓ | 信噪比 ↑ | 对数谱距离 ↓ |
| Kaiming | 2.5-e5 | 2.471 | -6.9-e5 | 2.144 | -3.4-e5 | 3.451 |
| Xvavier | -3.7-e5 | 4.42 | 3.3-e5 | 2.171 | 1.3-e5 | 3.167 |
| Sitzmann | **13.36** | **0.838** | **7.96** | **1.660** | **7.47** | **1.722** |

表10：周期性激活函数对初始化方案较为敏感。

**实验**。如表10所示，我们使用 Kaiming 初始化（He 等，2015）、Xavier 初始化（Glorot 和 Bengio，2010）以及 Sitzmann 初始化（Sitz mann 等人 2020），评估了 Sine（表7）激活函数表示音频信号的性能。结果清楚地表明，如果没有采用 Sitzmann 提出的依赖频率的初始化方法，正弦型周期性激活函数将无法收敛，也无法表示音频信号。

## B. 记忆

由于隐式神经表示将离散信号编码为神经网络权重，网络的有效性

在很大程度上取决于其记忆训练数据的能力。换言之，这种有效性由网络表示信号复杂性的能力决定。因此，本节主要分析坐标多层感知机和 Kolmogorov–Arnold 网络的表示能力，包括我们研究中提出的 Fourier-KAN。

## B.1 坐标多层感知机的记忆

Beyond Periodicity（Ramasinghe 和 Lucey 2022）表明，坐标多层感知机的有效性与其 Lipschitz 光滑性以及隐藏层表示的特征值分布密切相关。这些指标的最优值取决于待编码信号的特性。下面将概述主要的证明过程。

**推导**。具有 $k-1$ 个非线性隐藏层的多层感知机（MLP） $f$ 可描述如下，

$$f : \mathbf{x} \to g^k \circ \psi^{k-1} \circ g^{k-1} \circ \cdots \circ \psi^1 \circ g^1(\mathbf{x}),$$

其中 $g^i : \mathbf{x} \to \mathbf{A}^i \cdot \mathbf{x} + \mathbf{b}^i$ 表示具有可训练权重的仿射变换，$\mathbf{A}^i \in \mathbb{R}^{\dim(\mathbf{x}^i)\times\dim(\mathbf{x}^{i-1})}$ 是权重矩阵，$\mathbf{b}^i \in \mathbb{R}^{\dim(\mathbf{x}^i)}$ 是偏置项，$\psi^i$是非线性激活函数。最后一层是线性变换，得到 $f : \mathbf{x} \to g^k \circ \phi(\mathbf{x})$，其中 $\phi$是由多层感知机的前 $k-1$ 层构成的复合函数，不包括最后的线性层。给定 $N$ 个训练样本，总（训练）嵌入矩阵可定义如下，

$$\mathbf{X} \in \mathbb{R}^{D\times N} := \left[\phi(\mathbf{x}_1)^T\ \phi(\mathbf{x}_2)^T\ \ldots\ \phi(\mathbf{x}_N)^T\right],$$

其中 $\{\mathbf{x}_n\}_{n=1}^N$ 表示原始训练输入。

需要注意的是，多层感知机的最后一层通常是没有非线性的仿射投影。为简化记法，可以省略偏置项，从而得到:

$$\tilde{\mathbf{Y}} = \mathbf{A}^k\mathbf{X},$$

其中 $\tilde{\mathbf{Y}} \in \mathbb{R}^{q\times N}$ 表示多层感知机的输出。

假设$\mathbf{Y} \in \mathbb{R}^{q\times N}$ 表示多层感知机旨在学习的真实训练输出。如果多层感知机完美记忆了训练集，即 $\mathbf{Y} = \mathbf{Y}$，那么$\mathbf{Y}$的每一行都可以表示为$\mathbf{X}$各行的线性组合。假设对$\mathbf{Y}$没有先验知识，则$\mathbf{Y}$ 的各行可以是 $\mathbb{R}^N$中的任意向量。如果$\mathbf{X}$ 的各行线性无关，那么它们构成 $\mathbb{R}^N$ 的一组基（假设 $D \geq N$）。因此，如果 rank(**X**) $= N$，则可以保证（假设完美收敛）多层感知机能够找到权重矩阵$\mathbf{A}^k$，以确保$\mathbf{Y}$得到完美重构。

需要注意的是，尽管条件 rank(**X**) $=N$ 足以确保对任意信号实现完美记忆，但它并非总是必要条件。这是因为自然信号通常具有冗余——也就是说，它们的带宽有限。信号类别的带宽可以定义为表示该类别所需的线性无关（归一化）基向量的数量。因此，对于许多信号类别，rank(**X**) 可以小于 $N$，而多层感知机仍能实现完美的信号重构。此外，稳定秩为秩提供了一个下界（Rudelson and Vershynin 2007年）。较高的稳定秩通常会带来更好的重建性能，不过这一指标受网络宽度（$D$）的限制，而网络宽度小于数据点的数量（$N$）。反之，对冗余有限或没有冗余的含噪信号进行编码，需要更大的网络宽度；当 $D \ll N$ 时，通常会观察到较差的结果。

**分析**。基于上述推导，我们可以从光滑性和特征表示两个方面，简要分析 Gaussian-MLPs 和 ReLU 多层感知机的音频表示能力。一方面，高斯激活函数光滑且无穷可微，因此网络能够以更细致的方式拟合数据。此外，高斯函数的形状使每个激活函数都能在输入空间中以不同的尺度和位置生成局部特征，从而使网络能够在特定输入区域内进行更精细的调整。相比之下，ReLU激活函数并不光滑（在零点处不连续），且其导数在非零区域内为常数。因此，采用 ReLU 激活函数的多层感知机的频率响应能力相对较低，因为其激活特性有限，无法像高斯函数那样细致地表示函数空间。因此，在深度和宽度相同的情况下，Gaussian-MLPs 通常具有更大的网络带宽，这得益于其更高的光滑性和更复杂的特征表示能力。因此，对于给定的音频输入（其中 $N$ 为常数），Gaussian-MLPs 的带宽 $D$ 大于 ReLU 多层感知机，这与我们的基准测试结果一致，表明 Gaussian-MLPs 更适合音频信号表示。

此外，由于位置编码会将低维输入坐标映射到高维空间，因此能够扩展网络的带宽 $D$。因此，嵌入位置编码可以有效提升坐标多层感知机表示音频信号的能力。

## B.2 Fourier-KAN 的记忆

Liu 等人（Liu 等，2024）证明，采用 B 样条的 Kolmogorov–Arnold 网络能够以极低的误差范围逼近任意函数。下面我们将概述主要的证明过程。

**推导**。具有 $L$ 层的 KAN 网络可表示为:

$$f(\mathbf{x}) = (\mathbf{\Phi}_{L-1} \circ \mathbf{\Phi}_{L-2} \cdots \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0)(\mathbf{x}) \tag{12}$$

其中，$\Phi_l$ 表示第 $(l-1)$ 层与第 $l$ 层之间的传递矩阵，

$$\mathbf{\Phi}_l = \begin{pmatrix} \phi_{l,1,1}(\cdot) & \phi_{l,1,2}(\cdot) & \cdots & \phi_{l,1,n_l}(\cdot) \\ \phi_{l,2,1}(\cdot) & \phi_{l,2,2}(\cdot) & \cdots & \phi_{l,2,n_l}(\cdot) \\ \vdots & \vdots & & \vdots \\ \phi_{l,n_{l+1},1}(\cdot) & \phi_{l,n_{l+1},2}(\cdot) & \cdots & \phi_{l,n_{l+1},n_l}(\cdot) \end{pmatrix}.$$

根据一维B样条的古典音乐理论（de Boor，1978），并且已知 $\Phi_{l,\ i,\ j}$ 是连续的

函数在有界域上可以一致有界，因此存在有限网格B样条函数 $\Phi_{l,\,i,\,j}^{G}$，使得对于任意 $0 \leq m \leq k$,

$$\| \left(\Phi_{l,i,j} \circ \mathbf{\Phi}_{l-1} \circ \mathbf{\Phi}_{l-2} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0\right) \mathbf{x} - \left(\Phi_{l,i,j}^{G} \circ \mathbf{\Phi}_{l-1} \circ \mathbf{\Phi}_{l-2} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0\right) \mathbf{x} \|_{C^m} \leq C^{-k-1+m},$$

其中常数 $C$ 与 $G$无关，并使用 $C^m$-范数来衡量最高至$m$阶导数的大小,

$$\|g\|_{C^m} = \max_{|\beta| \leq m} \sup_{x \in [0,1]^n} \left|D^\beta g(x)\right|.$$

若固定这些 B样条近似，则残差 $R_l$ 可定义如下,

$$R_l := \left(\mathbf{\Phi}_{L-1}^{G} \circ \cdots \circ \mathbf{\Phi}_{l+1}^{G} \circ \mathbf{\Phi}_l \circ \mathbf{\Phi}_{l-1} \circ \cdots \circ \mathbf{\Phi}_0\right) \mathbf{x} - \left(\mathbf{\Phi}_{L-1}^{G} \circ \cdots \circ \mathbf{\Phi}_{l+1}^{G} \circ \mathbf{\Phi}_l^{G} \circ \mathbf{\Phi}_{l-1} \circ \cdots \circ \mathbf{\Phi}_0\right) \mathbf{x},$$

其满足,

$$\|R_l\|_{C^m} \leq C G^{-k-1+m},$$

其中常数与 $G$无关。最后注意到

$$f - \left(\mathbf{\Phi}_{L-1}^{G} \circ \mathbf{\Phi}_{L-2}^{G} \circ \cdots \circ \mathbf{\Phi}_1^{G} \circ \mathbf{\Phi}_0^{G}\right) \mathbf{x} = R_{L-1} + R_{L-2} + \cdots + R_1 + R_0.$$

根据上述推导，我们可以得出以下结论。对于式 12 中的函数 $f$(x)，若每个 $\Phi_{l,i,j}$ 均为 $(k+1)$ 次连续可微，则存在一个常数 $C$，它取决于 $f$ 及其表示形式，使得关于网格尺寸 $G$的以下近似误差界成立：存在 $k$ 阶 B样条函数 $\Phi_{l^{'},i,j}^{G}$，使得对于任意 $0 \leq m \leq k$，以下不等式成立：

$$\left\| f - \left(\mathbf{\Phi}_{L-1}^{G} \circ \mathbf{\Phi}_{L-2}^{G} \circ \cdots \circ \mathbf{\Phi}_1^{G} \circ \mathbf{\Phi}_0^{G}\right) \mathbf{x} \right\|_{C^m} \leq C G^{-k-1+m}.$$

**分析**。 Liu 等人（Liu 等，2024）已证明，基于 B样条基函数的 KAN网络能够在指定的误差范围内逼近任意函数。因此，如果能够证明 B样条基函数可以由傅里叶基函数逼近，那么 Fourier-KAN 网络也能够在给定的误差容限内逼近任意信号函数。下面将概述主要的证明过程。

给定定义在区间$[a,b]$上的 B样条函数 $S(x)$,

$$S(x) = \sum_{i=0}^{n} c_i B_{i,k}(x),$$

其中， $c_i$ 表示相应的控制点，$B_{i,\,k}(x)$表示 $k$阶B样条基函数，其定义如下递归关系式所示,

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1}, \\ 0 & \text{otherwise}, \end{cases}$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x),$$

其中， $t_i$ 表示节点序列， $k$ 表示样条的阶数。该递归公式由低阶基函数$B_{i,\,k}(x)$和$B_{i,\,k-1}(x)$以及$B_{i+1,\,k-1}(x)$构造高阶B样条基函数。

为将 B样条函数转换为周期函数，我们定义周期延拓函数 $S(x)$,

$$\tilde{S}(x) = \begin{cases} S(x) & \text{if } x \in [a,b], \\ S(x-T) & \text{if } x \in [a+T, b+T], \end{cases}$$

其中， $T = b - a$ 表示周期。此延拓过程将非周期 B样条函数转换为周期函数。

考虑到根据傅里叶级数的性质，任意周期函数 $f(x)$ 都可以表示为:

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left[ a_n \cos\left(\frac{2\pi n x}{T}\right) + b_n \sin\left(\frac{2\pi n x}{T}\right) \right],$$

其中，傅里叶系数 $a_n$ 和 $b_n$ 的计算方式如下:

$$a_0 = \frac{1}{T} \int_0^T f(x)\, dx,$$

$$a_n = \frac{2}{T} \int_0^T f(x) \cos\left(\frac{2\pi n x}{T}\right) dx,$$

$$b_n = \frac{2}{T} \int_0^T f(x) \sin\left(\frac{2\pi n x}{T}\right) dx.$$

根据傅里叶级数逼近定理（Davidson 和 Donsig, 2010年），任意平方可积周期˜函数 $f(x)$（包括 $S(x)$）都可以在 $L^2$ 意义下由其傅里叶级数逼近，并且随着项数 $N$ 增加，逼近误差会逐渐减小,

$$\|\tilde{S}(x) - S_N(x)\|_{L^2} \to 0 \text{ as } N \to \infty,$$

其中， $S_N(x)$表示含有 $N$ 项的傅里叶级数部分和:

$$S_N(x) = a_0 + \sum_{n=1}^{N} \left[ a_n \cos\left(\frac{2\pi n x}{T}\right) + b_n \sin\left(\frac{2\pi n x}{T}\right) \right].$$

总之，由于 B样条函数 $S(x)$ 是定义在有限区间上的分段多项式，其周期延˜拓 $S(x)$ 是一个周期函数。此外， $S(x)$ 可以在均方意义下由其傅里叶级数在 $L^2$ 空间中逼近。因此，任意 B样条函数 $S(x)$ 都可以通过其˜周期延拓 $S(x)$，由傅里叶级数 $S_N(x)$ 逼近；随着傅里叶级数项数 $N$ 增加，逼近误差趋近于零。

**结论**。基于上述分析，我们间接证明了 Fourier-KAN 网络能够逼近任意复杂度的信号。随着信号复杂度的变化，我们可以动态调整频率网格大小 Ω（见论文中的“频率自适应学习策略”一节）。这种调整体现了网络捕获信号不同频率分量的能力，并增强了

Fourier-KAN 网络的表示能力。此外，只有足够大的频率网格大小才能逼近复杂音频信号，这为采用倒金字塔式频率设置提供了理论依据。

### B.3 参数利用率比较

如 B.1 和 B.2 节所示，坐标多层感知机和 Fourier-KAN 在理论上都具备记忆任意复杂度信号的能力。然而，从非线性的角度来看，显然 Fourier-KAN 具有更强的表征能力。尽管如此，评估这些模型的记忆容量仍颇具挑战。因此，我们根据参数的利用情况，对坐标多层感知机和 Fourier-KAN 的存储容量进行比较分析。

**分析**。为简单起见，考虑一个深度为$L$的网络，其中每层的宽度均为 $n_0 = n_1 = \cdots = n_L = N$，且各层的最大频率阈值一致，$\Omega_1 = \Omega_2 = \cdots = \Omega_L = \Omega$。此类 Fourier-KAN 的参数总数为：

$$P_{\text{Fourier-KAN}} = O(N^2 L\Omega).$$

相比之下，具有相同深度 $L$ 和宽度 $N$ 的多层感知机（MLP）的参数数量为：

$$P_{\text{MLP}} = O(N^2 L)$$

。这意味着，在表示容量相当的离散信号时，多层感知机通常需要更多层，这可能会导致梯度爆炸或梯度消失等问题。因此，Fourier-KAN 的参数效率显著优于多层感知机。

| 模型 | 网络 | 参数 | 信噪比↑ | 对数谱距离↓ |
|---|---|---|---|---|
| ReLU | [1,256,256,256,256,1] | 263K | 0.00 | 4.623 |
| RFF+ReLU | [1,256,256,256,256,1] | 280K | 15.62 | 0.978 |
| NeFF+ReLU | [1,256,256,256,256,1] | 270K | 22.29 | 1.129 |
| Fourier-KAN | [1,64,64,64,64,1] | **254K** | **33.14** | **0.961** |

表11：Fourier-KAN 与 ReLU多层感知机在“巴赫”音频片段上的参数利用率比较。

**实验**。我们比较 ReLU多层感知机与 Fourier-KAN 网络的参数数量以及表示音频信号的能力。所考察的 ReLU多层感知机网络深度为 6，宽度为 256。相比之下，Fourier-KAN 网络的深度为 6，宽度为 64，输入层、隐藏层和输出层的最大频率阈值 Ω 分别设为 1024、5 和 3。

如表11所示，在深度保持不变时，宽度仅为64的 Fourier-KAN，其参数数量与宽度为256的 ReLU 多层感知机相近。这表明，Fourier-KAN 无需复杂的网络，就能表示容量相同的音频信号。此外，由于 Fourier-KAN 所使用的傅里叶基函数具有周期性，它能够以比 ReLU 多层感知机更少的参数，更有效地表示音频信号，展现出更高的参数效率。

## C. 可解释性

要通过符号回归彻底解释神经网络既困难又往往徒劳。不过，从局部视角对网络的表示形式进行可解释性分析是可行的。我们的 Fourier-KAN 基于局部周期性和傅里叶级数定理这一假设构建。因此，我们将通过这两个概念，简要分析坐标多层感知机和 Fourier-KAN 的可解释性。

**假设 2** 局部周期性假设。对于复数非平稳信号 $f(t)$，存在一个足够小的时间区间 $\epsilon > 0$，使得 $f(t)$ 可以用周期函数 $p(t)$近似：

$$\exists \epsilon > 0, \quad \forall t, \quad |t| < \epsilon : \quad f(t) \approx p(t)$$

其中 $p(t)$ 是周期为 $T$ 的周期函数，满足$p(t+T) = p(t)$。

**定理 2** 傅里叶级数定理。任何周期为 $T$ 的周期信号$p(t)$ 都可以表示为正弦函数和余弦函数的无穷级数，

$$p(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos\left(\frac{2\pi nt}{T}\right) + b_n \sin\left(\frac{2\pi nt}{T}\right) \right)$$

其中 $a_0$、$a_n$ 和 $b_n$ 是傅里叶系数。

根据局部周期性的假设（假设 2）和傅里叶级数定理（定理 2），任何复音频信号 $f(t)$ 都可以被视为在一个**局部**区域内具有周期性。因此，它可以表示为正弦函数和余弦函数的级数，

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos\left(\frac{2\pi nt}{T}\right) + b_n \sin\left(\frac{2\pi nt}{T}\right) \right). \tag{13}$$

### C.1 坐标多层感知机的可解释性

**分析**。对于大多数激活函数，例如 ReLU、Tanh 和高斯激活函数，网络的学习能力主要可以从逼近函数 $f(t)$（公式 13）的角度进行解释。因此，这些激活函数的平滑性、连续性和可微性可能会影响网络表示音频信号的能力。例如，高斯激活函数比 ReLU 更平滑且连续性更好，这解释了为什么 Gaussian-MLPs 相较于 ReLU多层感知机具有更强的拟合音频信号能力。这一发现与我们在 B.1 节中得出的结论一致。

对于正弦等激活函数，可以从傅里叶级数定理的角度探讨其可解释性。

考虑正弦多层感知机的输入层：设输入时间坐标记为 $t \in \mathbb{R}$，权重矩阵为$\mathbf{W} \in \mathbb{R}^{3\times 1}$，偏置向量为$\mathbf{b} \in \mathbb{R}^{3\times 1}$。那么，输入层的输出$\mathbf{y} \in \mathbb{R}^{3\times 1}$可表示为：

$$\mathbf{y} = \sin(\omega \mathbf{W} t + \mathbf{b})$$

将该表达式展开可得：

$$\mathbf{y} = \sin(\omega \mathbf{W} t + \mathbf{b}) = \begin{bmatrix} \sin(\omega w_1 t + b_1) \\ \sin(\omega w_2 t + b_2) \\ \sin(\omega w_3 t + b_3) \end{bmatrix},$$

然后，

$$\begin{aligned} y_1 &= \sin(\omega w_1 t + b_1) \\ y_2 &= \sin(\omega w_2 t + b_2) \\ y_3 &= \sin(\omega w_3 t + b_3). \end{aligned} \tag{14}$$

这意味着，输入层的主要功能是将时间坐标映射到高维空间，这与傅里叶特征位置编码（例如 NeFF、RFF）的作用类似。因此，在辐射场和图像表示等应用中，正弦型激活函数可以有效消除对位置编码的需求。

考虑正弦多层感知机的输出层：

- 输入隐变量：$\mathbf{z} \in \mathbb{R}^3$，其中 $\mathbf{z} = [z_1, z_2, z_3]^T$。
- 偏置向量：$\mathbf{c} \in \mathbb{R}^3$，其中 $\mathbf{c} = [c_1, c_2, c_3]^T$。
- 权重向量：$\mathbf{v} \in \mathbb{R}^3$，其中 $\mathbf{v} = [v_1, v_2, v_3]^T$。
- 频率标量：$\omega \in \mathbb{R}$。
- 输出振幅：$A \in \mathbb{R}$。

那么，Sine-MLP 输出层的输出可以用矩阵形式表示为：

$$A = \sin(\omega \cdot (\mathbf{v}^T \mathbf{z} + \mathbf{c}^T \mathbf{1}))$$

其中，$\mathbf{1}$ 是一个全 1 向量 $[1, 1, 1]^T$，用于确保偏置被正确地加到线性变换中。展开该表达式可得：

$$A = \sin(\omega \cdot (v_1 z_1 + v_2 z_2 + v_3 z_3 + c_1 + c_2 + c_3)). \tag{15}$$

**结论**。对于大多数激活函数（ReLU、高斯等），评估主要基于它们近似音频信号的能力，并考虑光滑性、连续性和可微性等性质。相比之下，对于周期性的正弦型激活函数，如式 14 所示，输入层实际上起到傅里叶特征位置编码的作用。这解释了为什么这类函数即使没有显式的位置编码，在辐射场和图像表示等应用中也具有强大的表征能力。另一方面，如式 15 所示，输出层使用单一频率 $\omega$ 的正弦函数来表示音频振幅。这种表示本质上涉及特征融合，尽管其周期性特征局限于网络内部。鉴于不同音频信号具有不同的频率尺度，这也解释了为什么正弦型激活函数对频率超参数较为敏感，并且依赖于频率相关的初始化方案。

## C.2 Fourier-KAN 的可解释性

**分析**。给定一个维度为$[n_0, n_1, \cdots, n_L]$的 Fourier-KAN，可以使用以下方式定义

递归关系，

$$\begin{aligned} \mathbf{z}^{(0)} &= t \\ \mathbf{z}^{(l+1)} &= \sum_{\omega=1}^{\Omega_l} \left[ \mathbf{a}^{(l,\omega)} \cos(\omega \mathbf{z}^{(l)}) + \mathbf{b}^{(l,\omega)} \sin(\omega \mathbf{z}^{(l)}) \right] + \mathbf{c}^{(l)} \\ f(t) &= \mathbf{z}^{(n_L)} \left( \mathbf{z}^{(n_L-1)} \left( \mathbf{z}^{(l)} \left( ...(\mathbf{z}^{(0)}) \right) \right) \right), l = n_0, \ldots, n_L, \end{aligned}$$

其中，$\mathbf{a}^{(l,\omega)}$ 和 $\mathbf{b}^{(l,\omega)} \in \mathbb{R}^{d_{l+1} \times d_l}$ 分别表示频率为 $\omega$ 时第 $l$ 层的傅里叶系数权重，$\mathbf{c}^{(l)} \in \mathbb{R}^{d_{l+1}}$ 是第 $l$ 层的偏置项，$z^{(l)} \in \mathbb{R}^{d_l}$表示第 $l$ 层的隐藏单元，而 $\Omega_l$ 是一个超参数，表示第 $l$ 层的最大频率阈值。

考虑 Fourier-KAN 的输入层：

- 输入时间坐标：$t \in \mathbb{R}$。
- 输出维度：$d_1 = 3$。
- 第一层的频率 $\Omega_0$ 为 3。
- $\mathbf{a}^{(0,\omega)} = [a_{(0,\omega)\ 1}, a_{(0,\omega)\ 2}, a_{(0,3}{}^{\omega}{}]^T)$ 表示与频率$\omega$对应的余弦项权重系数。
- $\mathbf{b}^{(0,\omega)} = [b_{(0,\omega)\ 1}, b_{(0,\omega)\ 2}, b_{(0,3}{}^{\omega}{}]^T)$ 表示与频率 $\omega$对应的正弦项权重系数。
- $\mathbf{c}^{(0)} = [c_{(0)\ 1}, c_{(0)\ 2}, c_{(0)\ 3}]^T$ 是第一层的偏置项向量。

根据这些定义，第一层的输出记为$\mathbf{z}^{(1)}$，可表示为：

$$\begin{aligned} \mathbf{z}^{(1)} = [\cos(t) \quad \sin(t) \quad \cos(2t) \quad \sin(2t) \quad \cos(3t) \quad \sin(3t)] \\ \times \begin{bmatrix} a_1^{(0,1)} \\ b_1^{(0,1)} \\ a_1^{(0,2)} \\ b_1^{(0,2)} \\ a_1^{(0,3)} \\ b_1^{(0,3)} \end{bmatrix} + \begin{bmatrix} c_1^{(0)} \\ c_2^{(0)} \\ c_3^{(0)} \end{bmatrix}. \end{aligned}$$

展开后，第一层的输出 $\mathbf{z}^{(1)}$ 为：

$$\begin{aligned} \mathbf{z}_1^{(1)} &= a_1^{(0,1)} \cos(t) + b_1^{(0,1)} \sin(t) \\ &\quad + a_1^{(0,2)} \cos(2t) + b_1^{(0,2)} \sin(2t) \\ &\quad + a_1^{(0,3)} \cos(3t) + b_1^{(0,3)} \sin(3t) \\ &\quad + c_1^{(0)} \\ \mathbf{z}_2^{(1)} &= a_2^{(0,1)} \cos(t) + b_2^{(0,1)} \sin(t) \\ &\quad + a_2^{(0,2)} \cos(2t) + b_2^{(0,2)} \sin(2t) \\ &\quad + a_2^{(0,3)} \cos(3t) + b_2^{(0,3)} \sin(3t) \\ &\quad + c_2^{(0)} \\ \mathbf{z}_3^{(1)} &= a_3^{(0,1)} \cos(t) + b_3^{(0,1)} \sin(t) \\ &\quad + a_3^{(0,2)} \cos(2t) + b_3^{(0,2)} \sin(2t) \\ &\quad + a_3^{(0,3)} \cos(3t) + b_3^{(0,3)} \sin(3t) \\ &\quad + c_3^{(0)} \end{aligned} \tag{16}$$

比较式 13 和式 16 可知，Fourier-KAN 的输入层将时间坐标 $t$ 映射到一个高维傅里叶空间，其中包含各种频率。这使网络能够学习不同频率域中的频率分量。

现在，我们将注意力转向 Fourier-KAN 的输出层:

- 输出层的输入：$\mathbf{z}_{(L)} \in \mathbb{R}^3$。
- 输出维度： $d_{L+1} = 1$。
- 输出振幅： $A \in \mathbb{R}$。
- 输出层的频率 $\Omega_L$ 为 3。
- $\mathbf{a}_{(L,\omega)} = [a_{(L,\omega)\ 1}, a_{(L,\omega)\ 2}, a_{(L,3}{}^{\omega}{}]^T)$ 表示输出层中频率 $\omega$ 对应的余弦项的权重系数。
- $\mathbf{b}_{(L,\omega)} = [b_{(L,\omega)\ 1}, b_{(L,\omega)\ 2}, b_{(L,3}{}^{\omega}{}]^T)$ 表示频率 $\omega$ 对应的输出层正弦项的权重系数。
- $\mathbf{c}_{(L)} = [c_{(L)}]$是输出层的偏置项，它是一个标量。

随后，输出层的输出振幅为:

$$A = \begin{bmatrix}\cos(\mathbf{z}^{(L)}) & \sin(\mathbf{z}^{(L)}) & \cdots & \cos(3\mathbf{z}^{(L)}) & \sin(3\mathbf{z}^{(L)})\end{bmatrix} \times \begin{bmatrix} a_1^{(L,1)} \\ b_1^{(L,1)} \\ a_1^{(L,2)} \\ b_1^{(L,2)} \\ a_1^{(L,3)} \\ b_1^{(L,3)} \end{bmatrix} + \begin{bmatrix} c^{(L)} \end{bmatrix}.$$

展开上述表达式可得:

$$\begin{aligned} A = {} & a_1^{(L,1)} \cos(\mathbf{z}^{(L)}) + b_1^{(L,1)} \sin(\mathbf{z}^{(L)}) \\ & + a_1^{(L,2)} \cos(2\mathbf{z}^{(L)}) + b_1^{(L,2)} \sin(2\mathbf{z}^{(L)}) \\ & + a_1^{(L,3)} \cos(3\mathbf{z}^{(L)}) + b_1^{(L,3)} \sin(3\mathbf{z}^{(L)}) \\ & + c^{(L)}. \end{aligned} \tag{17}$$

将公式 13 与公式 17 进行比较可以发现，Fourier-KAN 的输出层天然采用了傅里叶级数的形式。这意味着 Fourier-KAN 通过傅里叶级数表示，隐式地预测坐标 $t$ 的局部邻域，这与局部周期性的假设一致。

**结论**。 如上所述，根据傅里叶级数定理，Fourier-KAN 将时间坐标 $t$ 映射到具有不同频率的高维空间，从而能够学习具有各种频率的邻域特征。在输出层，Fourier-KAN 使用傅里叶级数将邻域特征表示为相应的振幅值。这表明，Fourier-KAN 以傅里叶级数的形式刻画离散坐标 $t$ 的邻域，从而生成连续的音频信号。

与坐标多层感知机相比，Fourier-KAN具有更强的可解释性。我们认为，Fourier-KAN在音频领域的进一步应用具有巨大潜力。

# D. 进一步实验

## D.1 实验设置

**评估指标**。给定参考信号 $y$ 及其近似值 $\hat{y}$，信噪比（SNR）定义如下（Roux 等，2019），

$$\mathrm{SNR}(\hat{y}, y) = 20 \log_{10} \left( \frac{\|y\|_2^2}{\|\hat{y} - y\|_2^2} \right).$$

SNR 是信号处理文献中用于量化信号近似质量的标准指标。另一方面，对数谱距离（LSD）评估各个频率上的重建质量，其定义如下（Gray 和 Markel，1976）：

$$\mathrm{LSD}(\hat{y}, y) = \frac{1}{L} \sum_{\ell=1}^{L} \sqrt{\frac{1}{K} \sum_{k=1}^{K} (X(\ell, k) - \hat{X}(\ell, k))^2},$$

其中， $X$ 和 $\hat{X}$ 分别是 $y$和$\hat{y}$的对数谱功率幅值，定义为 $X = \log |S|^2$，其中 $S$ 是信号的短时傅里叶变换（STFT）。这里，$\ell$ 和 $k$ 分别表示帧和频率的索引。在我们的实验中，我们使用了长度为 2048 的帧。由于 LSD 是频域中的间接评估指标，因此我们主要关注 SNR 指标。

**实现细节**。 我们的 NeAF 模型使用 Python 和 PyTorch 框架实现，并在配备 AMD EPYC 7302 16 核处理器和 NVIDIA GeForceRTX3090 GPU 的工作站上进行训练。训练过程持续 1000 个 epoch，批量大小为 16,384，采用 1e-4 的学习率，并使用基于 CosineAnnealingLR 方法的学习率调度策略。对于 RFF位置编码，方差 $\sigma$ 设为 10，频率维度 $L$设为 32。对于 NeFF 位置编码，频率维度定义为 $log(1/(2 * (2 * 1/B)))$，其中 $B$ 表示批量大小。正弦和 Incode-Sine 激活函数的频率参数$\omega$ 设为 30。此外，对于高斯型激活函数，方差因子 $a$ 固定为 0.1。

**数据集详情**。**GTZAN** 音乐数据集（Tzanetakis 和 Cook，2002年）包含 1000 段音乐片段，涵盖十种不同流派，每段时长为 30 秒。**CSTR VCTK** 语料库（Yamagishi、Veaux 和 MacDonald，2019）包含 110 位带有不同口音的英语使用者所说的语音数据。每位使用者朗读了约 400 个句子，这些句子选自报纸、彩虹篇章，以及用于语音口音档案的引导段落。

对于 GTZAN 数据集（Tzanetakis 和 Cook，2002年），我们从十种音乐流派（布鲁斯、古典音乐、乡村音乐、迪斯科、嘻哈、爵士乐、金属乐、流行音乐、雷鬼和摇滚）中各选取编号为 00000 的音频文件，作为实验音频数据。

对于 CSTR VCTK 数据集（Yamagishi、Veaux 和 MacDonald，2019），我们从十种不同的口音（英语、苏格兰英语、北爱尔兰英语、爱尔兰英语、印度英语、威尔士英语、新西兰英语、美国英语、加拿大英语和澳大利亚英语）中选取了音频数据，这些口音对应数据集中录音时长最长的样本。

对应的说话人 ID 为 225、234、238、245、248、253、335、345、363 和 374，序列号为 023，麦克风编号为 1。

### D.2 基准排行榜补充

如表 12 所示，我们对坐标多层感知机在音频信号表示任务中的表现进行了全面的基准测试，并额外纳入了两种 Kolmogorov–Arnold 网络——B样条 KAN 和我们提出的 Fourier-ASR——的评估结果。

根据这项全面的基准测试，我们得出以下详细结论：

- 大多数激活函数（如 Gabor-Wavelet、二次、多二次函数、指数正弦、Sigmoid、SoftPlus、Tanh、ELU 函数、SiLU、PReLU、ReLU）在不使用位置编码的情况下，都不足以表示音频信号。
- 只有高度非线性的函数（如高斯、拉普拉斯、超高斯）或周期函数（如正弦、Incode-Sine）才能部分捕捉音频信号的高频和局部周期性特征。

- 位置编码显著增强了坐标多层感知机表示音频信号的能力。这一提升主要归因于位置编码的高维映射，它提高了模型的非线性和带宽（见附录 B.1）。

- 得益于 Fourier-KAN 的周期性，我们提出的 Fourier-ASR 优于基于 B 样条的 KAN 网络（Liu 等人，2024）。

- 在不使用位置编码的情况下，我们提出的 Fourier-ASR 达到了最先进的性能，以 10.42 dB 的信噪比提升显著优于次佳方法（正弦）。

### D.3 定性实验

如图 3 所示，我们开展了一项定性实验，以提供全面的基准测试排名。显然，嵌入位置编码后，周期性的正弦函数和 Incode-Sine 函数能够有效捕捉音频信号中的高频细节。同样，我们提出的 Fourier-ASR 也具有周期性，即使无需位置编码，也能准确表示音频信号。对于高度非线性的高斯激活函数，尽管其表示音频信号的能力有一定局限，但其重构误差在各个时间坐标上仍均匀分布，不受音频振幅影响。相比之下，周期性激活函数正弦和 Incode-Sine 的重构误差会随振幅变化，并且在变化显著的区域更为明显，这表明它们难以捕捉音频信号中的瞬时变化。然而，Fourier-ASR 将周期性与强非线性相结合，能够更有效地捕捉音频中的瞬时变化，生成更符合人类听觉感知的重建结果。

| 激活函数 $\sigma(\cdot)$ | 方程 | 参数 | 位置编码 $\gamma(\cdot)$ | 巴赫 (7s) | | 计数 (7s) | | 布鲁斯 (30s) | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ |
| Gabor-小波 | $e^{jax} \cdot e^{-\lvert bx\rvert^2}$ | $a, b$ | 恒等函数 | - | - | - | - | - | - | - | - |
| | | | RFF | - | - | - | - | - | - | - | - |
| | | | NeFF | - | - | - | - | - | - | - | - |
| 二次 | $\frac{1}{1+(ax)^2}$ | $[a]$ | 恒等函数 | - | - | -3.37 | 5.182 | -1.553 | 7.573 | - | - |
| | | | RFF | -1.55 | 6.346 | -8.28 | 5.182 | -1.97 | 7.572 | -3.93 | 6.367 |
| | | | NeFF | -5.69 | 6.344 | -1.97 | 5.182 | -2.59 | 7.572 | -3.42 | 6.366 |
| 多二次函数 | $\frac{\sqrt{1+1(ax)}}{2}$ | $[a]$ | 恒等函数 | - | -2.48 | 5.182 | - | -1.55 | 7.573 | - | - |
| | | | RFF | - | - | -1.19 | 5.182 | -3.11 | 7.573 | - | - |
| | | | NeFF | -5.17 | 6.344 | -2.17 | 5.182 | -1.55 | 7.573 | -2.96 | 6.366 |
| 指数正弦 | $e^{-\sin(ax)}$ | $[a]$ | 恒等函数 | 0.00 | 6.339 | 0.00 | 5.180 | -8.28 | 7.571 | -2.76 | 6.363 |
| | | | RFF | 7.97 | 5.297 | 1.45 | 4.171 | 0.00 | 6.178 | 3.14 | 5.215 |
| | | | NeFF | 0.00 | 3.326 | 1.04 | 4.550 | 0.00 | 5.520 | 0.35 | 4.465 |
| Sigmoid | $\frac{1}{1+e^{-x}}$ | | 恒等函数 | 0.00 | 6.339 | 0.00 | 5.181 | -6.73 | 7.572 | -2.24 | 6.364 |
| | | | RFF | 0.03 | 3.186 | 7.04 | 3.945 | 0.00 | 5.921 | 2.36 | 4.351 |
| | | | NeFF | 0.51 | 1.905 | 0.00 | 4.259 | 0.00 | 5.212 | 0.17 | 3.792 |
| SoftPlus | $\frac{1}{a}\log(1+e^{ax})$ | $[a]$ | 恒等函数 | -1.92 | 6.311 | 0.00 | 5.165 | -3.11 | 7.554 | -1.68 | 6.343 |
| | | | RFF | 0.55 | 1.647 | 0.03 | 2.709 | 0.06 | 3.773 | 0.21 | 2.710 |
| | | | NeFF | 3.04 | 1.469 | 0.26 | 1.933 | 0.51 | 2.724 | 1.27 | 2.042 |
| Tanh | $\frac{e^x-e^{-x}}{e^x+e^{-x}}$ | | 恒等函数 | -7.25 | 6.334 | 0.00 | 5.181 | -3.11 | 7.574 | -3.45 | 6.363 |
| | | | RFF | 7.95 | 1.082 | 0.97 | 1.786 | 1.32 | 2.420 | 3.41 | 1.763 |
| | | | NeFF | 11.09 | 1.046 | 6.87 | 1.593 | 3.60 | 1.927 | 7.19 | 1.522 |
| ELU | $\begin{cases} ax, & \text{如果 } x>0 \\ ae^x - a, & \text{否则} \end{cases}$ | $[a]$ | 恒等函数 | 0.00 | 5.979 | 0.00 | 5.104 | -4.40 | 7.535 | -1.47 | 6.206 |
| | | | RFF | 7.77 | 1.111 | 0.67 | 1.956 | 0.75 | 2.723 | 3.06 | 1.930 |
| | | | NeFF | 12.26 | 0.997 | 4.89 | 1.557 | 2.52 | 2.057 | 6.56 | 1.537 |
| SiLU | $\frac{x}{1+e^{-x}}$ | | 恒等函数 | -5.18 | 6.265 | -6.21 | 5.174 | -2.07 | 7.560 | -4.49 | 6.333 |
| | | | RFF | 7.51 | 1.104 | 0.51 | 1.952 | 0.74 | 2.643 | 2.92 | 1.900 |
| | | | NeFF | 12.71 | 0.950 | 7.50 | 1.618 | 4.23 | 1.881 | 8.15 | 1.483 |
| PReLU | $\begin{cases} x, & \text{如果 } x>0 \\ ax, & \text{否则} \end{cases}$ | $[a]$ | 恒等函数 | 0.00 | 4.724 | 0.00 | 4.630 | 0.00 | 7.031 | 0.00 | 5.462 |
| | | | RFF | 13.42 | 1.010 | 3.38 | 1.437 | 2.50 | 2.035 | 6.43 | 1.494 |
| | | | NeFF | 17.50 | 1.133 | 7.88 | 1.575 | 5.20 | 1.539 | 10.19 | 1.416 |
| ReLU | $\max(0, x)$ | | 恒等函数 | 0.00 | 4.623 | -7.66 | 4.546 | 0.00 | 6.774 | -2.55 | 5.314 |
| | | | RFF | 15.62 | 0.978 | 4.93 | **1.400** | 3.23 | 1.862 | 7.93 | 1.413 |
| | | | NeFF | 22.29 | 1.129 | 9.57 | 1.538 | 7.64 | 1.324 | 13.17 | 1.330 |
| 高斯 | $e^{\frac{-x^2}{2a^2}}$ | $[a]$ | 恒等函数 | 6.35 | 1.130 | 0.74 | 2.165 | 0.68 | 3.059 | 2.59 | 2.118 |
| | | | RFF | 20.85 | 2.046 | 12.14 | 3.195 | 11.80 | 1.346 | 14.93 | 2.196 |
| | | | NeFF | 19.68 | 2.127 | 9.20 | 3.438 | 7.74 | 1.597 | 12.21 | 2.387 |
| 拉普拉斯 | $e^{\frac{-\lvert x\rvert}{a}}$ | $[a]$ | 恒等函数 | 12.04 | 0.932 | 1.34 | **1.561** | 1.37 | 2.434 | 4.92 | 1.642 |
| | | | RFF | 15.57 | 2.386 | 10.97 | 2.632 | 14.74 | **1.112** | 13.76 | 2.043 |
| | | | NeFF | 15.26 | 2.434 | 8.67 | 3.191 | 8.16 | 1.526 | 10.70 | 2.384 |
| 超高斯 | $[e^{\frac{-x^2}{2a^2}}]^b$ | $[a, b]$ | 恒等函数 | 6.38 | 1.101 | 0.75 | 2.166 | 0.68 | 3.059 | 2.60 | 2.109 |
| | | | RFF | 20.86 | 2.044 | 12.44 | 3.169 | 11.80 | 1.346 | 15.03 | 2.186 |
| | | | NeFF | 19.69 | 2.127 | 9.20 | 3.437 | 7.74 | 1.597 | 12.21 | 2.387 |
| Sine | $\sin(\omega x)$ | $[\omega]$ | 恒等函数 | 13.36 | 0.838 | 7.96 | 1.660 | **7.47** | **1.722** | **9.59** | **1.407** |
| | | | RFF | **39.02** | **0.582** | **13.06** | 1.412 | **16.57** | 1.156 | **22.88** | **1.050** |
| | | | NeFF | **42.39** | **0.537** | **33.58** | **0.914** | **22.02** | **0.696** | **32.66** | **0.716** |
| Incode-Sine | $a$正弦$(b\omega x + c) + d$ | $[\omega], a, b, c, d$ | 恒等函数 | **15.98** | **0.778** | **8.16** | 1.611 | 0.01 | 3.865 | 8.05 | 2.085 |
| | | | RFF | 38.10 | 0.595 | 12.86 | 1.559 | 15.13 | 1.241 | 22.03 | 1.132 |
| | | | NeFF | 41.40 | 0.556 | 32.24 | 1.038 | 21.33 | 0.763 | 31.99 | 0.786 |
| 基函数 $\phi(\cdot)$ | 方程 | 参数 | 位置编码 $\gamma(\cdot)$ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ | 信噪比↑ | 对数谱距离↓ |
| B样条 KAN | $\frac{ax}{1+e^{-x}} + spline(x)$ | $a,$ 样条系数 | No | 1.54 | 1.531 | 0.00 | 3.293 | 0.00 | 4.643 | 0.51 | 3.156 |
| | | | RFF | - | - | - | - | - | - | - | - |
| | | | NeFF | - | - | - | - | - | - | - | - |
| Fourier-ASR | $a\sin(\omega x) + b\cos(\omega x)$ | $a, b$ | No | **33.14** | 0.961 | **13.10** | 1.377 | **13.80** | 1.240 | **20.01** | 1.193 |
| | | | RFF | 16.37 | 2.352 | **18.36** | 2.625 | 12.59 | 1.305 | 15.77 | 2.094 |
| | | | NeFF | 15.99 | 2.380 | 6.96 | 3.603 | 8.32 | 1.564 | 10.42 | 2.516 |

表 12：基准测试排行榜。对于不同类型的位置编码（恒等函数、RFF、NeFF），最佳结果以**第**一名和第二名标出。请注意，“$a$”表示可学习参数，而“$[a]$”表示超参数，“-”表示模型无法收敛。

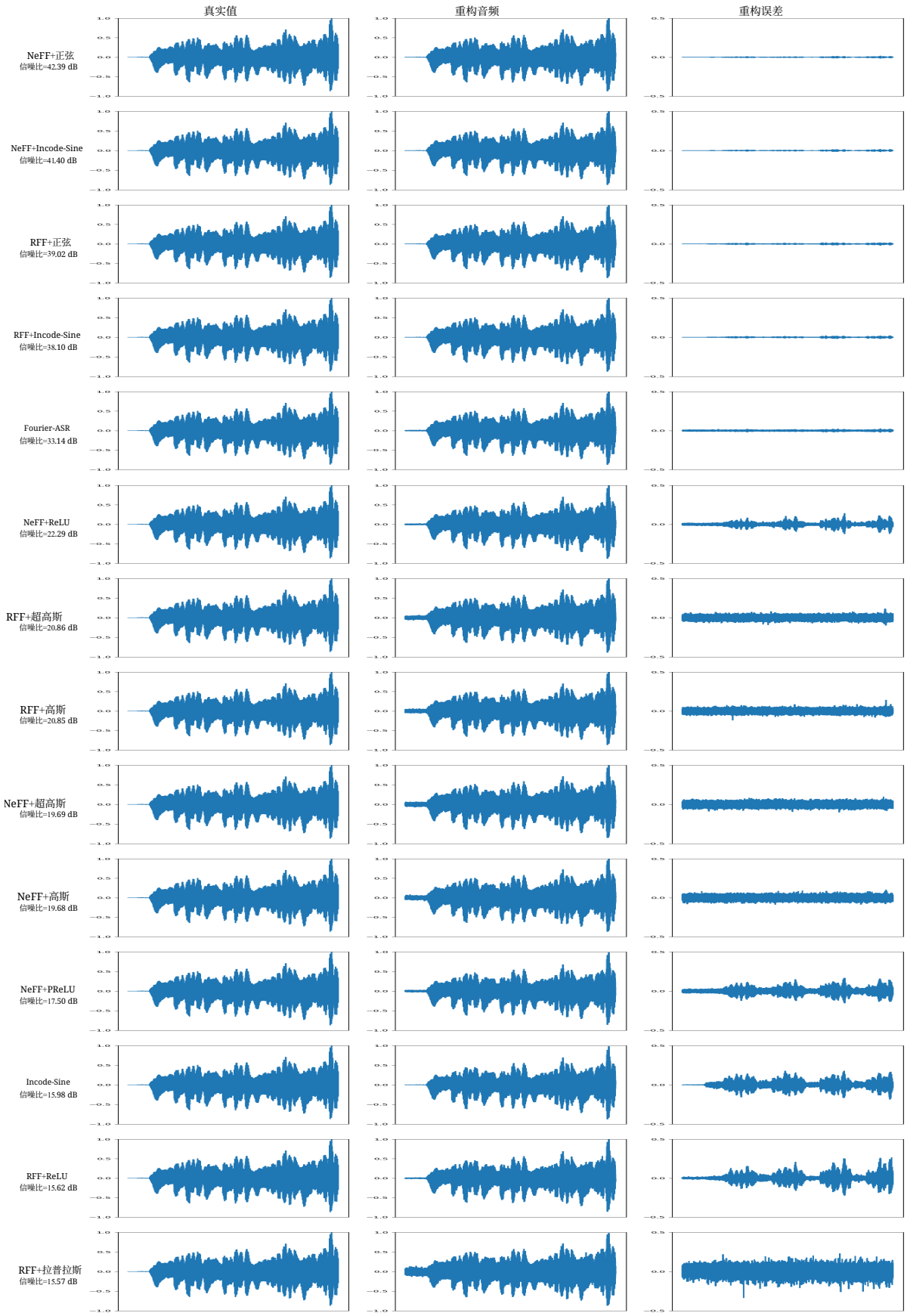

真实值
重构音频
重构误差
NeFF+正弦
信噪比=42.39 dB
NeFF+Incode-Sine
信噪比=41.40 dB
RFF+正弦
信噪比=39.02 dB
RFF+Incode-Sine
信噪比=38.10 dB
Fourier-ASR
信噪比=33.14 dB
NeFF+ReLU
信噪比=22.29 dB
RFF+超高斯
信噪比=20.86 dB
RFF+高斯
信噪比=20.85 dB
NeFF+超高斯
信噪比=19.69 dB
NeFF+高斯
信噪比=19.68 dB
NeFF+PReLU
信噪比=17.50 dB
Incode-Sine
信噪比=15.98 dB
RFF+ReLU
信噪比=15.62 dB
RFF+拉普拉斯
信噪比=15.57 dB

图3：“巴赫”上的定性实验。