PEAKNETFP: 基于峰值的神经音频指纹识别对极端时间拉伸具有鲁棒性

圭尔梅 · 科尔特斯-塞巴斯蒂亚 [13] 本杰明 · 马丁 [2] 埃米利奥 · 莫利纳 [1]
谢瓦利耶 塞拉 [3] 罗曼 亨宁昆 [2]
[1] BMAT Licensing S.L., 西班牙巴塞罗那
[2] Deezer Research, 法国巴黎
[3] 音乐技术小组，庞培法布拉大学，西班牙巴塞罗那
研究@bmat.com, 研究@deezer.com

ABSTRACT

这项工作介绍了 *PeakNetFP*，首个专门围绕频谱峰值设计的神经音频指纹（AFP）系统。该新颖系统旨在利用传统基于峰值的 AFP 方法通常计算出的稀疏频谱坐标。*PeakNetFP* 执行与计算机视觉模型 *PointNet++*类似的分层点特征提取技术，并使用对比学习进行训练，类似于最先进的深度学习 AFP，神经指纹算法。这种结合使得峰值网络浮点精度能够超越传统的 AFP 系统，并在处理具有挑战性的时域拉伸音频数据时达到与神经网络指纹相当的性能。在广泛的评估中，对于从 50% 到 200%的拉伸因子，*PeakNetFP* 保持超过 90%的第一命中率。此外，*PeakNetFP* 提供了显著的效率优势：与神经网络指纹相比，它的参数减少了 100 倍，并且使用了小 11 倍的输入数据。这些特性使得 *PeakNetFP* 成为一个轻量级且高效的解决方案，在涉及时间拉伸的 AFP 任务中表现出色。总体而言，该系统代表了未来 AFP 技术的一个有希望的发展方向，因为它成功地将基于峰值的 AFP 的轻量化特点与基于神经网络方法的适应性和模式识别能力结合起来，为领域内更可扩展和高效的解决方案铺平了道路。

1. 介绍

音频指纹识别（AFP）是在参考曲目数据库中识别音频录音的 MIR 任务。早期的 AFP 系统可以追溯到二十年前，如 Shazam [1] 和 Philips [2] 系统。自那时以来，AFP 已被广泛研究用于各种用例，例如示例查询 [1]、完整性验证 [3]、基于内容的复制检测 [4]、DJ 集合监控 [5] 或高特定音频检索 [6]。基于峰值的 AFP 系统在该领域有着悠久的历史，多个研究工作利用这种方法来增强其对音高变换和时间拉伸的鲁棒性 [7]、背景音乐识别 [8] 或创建可以在嵌入式系统中运行的轻量级 AFP [9]。这些算法基于从时频表示计算出的显著谱峰的提取和链接。这些都是成熟且可投入生产使用的系统，不需要训练，并可以扩展到包含数百万参考数据库的工业级别 [10, 11]。因此，拥有大规模数据目录的公司依赖它们进行内容识别 [12]。

表示学习系统，如现在播放 [13] 或神经 *FP* [6]，最近作为新颖的方法出现，利用对比学习（CL）和卷积神经网络（CNNs）来学习扭曲的音频片段与其相应参考轨道之间的相似性。它们被设计用于执行高度敏感的音频检索，能够匹配短片段，并在具有挑战性的条件下显著优于传统的峰值基础方法。这归因于它们能够从数据中捕捉更复杂和细微的特征，使其对传统方法难以处理的各种类型的扭曲和噪声更具鲁棒性 [6, 14]。

这需要大量的计算资源、密集的输入数据、模型训练和 GPU 计算，这些可能不适合某些应用。在工业解决方案中，这些要求可能很难克服，基于峰值的特征仍被视为可行的替代方案 [11, 12, 15, 16]。实际上，常见的是音频特征需要在客户端设备上进行计算，然后上传到服务器以与参考数据库进行比对识别。在这种情况下，需要密集的频谱图作为音频特征显著增加了要上传的数据量，相比只发送稀疏的频谱峰值。当考虑指纹生成运行于客户端时，在训练好的模型上进行推理可能比简单的基于规则的峰值提取算法更复杂且耗电，特别是在示例查询应用 [15] 中，客户端设备一般规格各异（例如智能手机）。另一种设置是完全在设备内的音频识别 [13]，尽管这通常意味着对设备计算

要求更加严格，在内存占用方面，并限制了数据库大小。此外，音乐版权所有者往往不愿意分享任何可能被逆向或用于除指纹外其他任务的密集表示，而更倾向于计算携带较少信息且很难用于设计以外用途（少数例外 [17]）的稀疏目标特征。最后，由于基于峰值的 AFP 已广泛应用于工业系统中，对于私营公司来说利用大量预计算的频谱峰值数据集来进行神经音频指纹识别方法 [1,12] 是有意义的。因此，在这项工作中我们建议保留传统的基于峰值的特征作为输入，并在现代神经网络方法中使用它们。

在这一首个关于神经稀疏峰值模型的出版物中，我们选择将研究重点放在极端条件下的时间拉伸上，这是文献中尚未充分探索的一个领域。时间拉伸是一种音频处理技术，可以改变音轨的速度而不改变其音调。这种方法通常被 DJ 用来同步混音中不同歌曲的节奏或创建放慢或加快速度的混音 [18]。在复杂的场景下，比如混音、混合或规避授权尝试，时间拉伸发生在复杂的身份识别情境中，在这种情况下对短片段进行严重的节奏修改，使得它们非常难以被自动识别 [19]。

这项工作的主要贡献是引入了一种新型的 AFP 系统，该系统以轻量级频谱峰值作为输入，但基于表示学习方法，并在时间拉伸的背景下进行了评估。具体来说，我们的模型 *PeakNetFP* 应用对比学习从稀疏频谱峰值输入中学习指纹，利用分层点集学习算法 *PointNet++* [20]。该设计旨在展示神经网络最新技术的良好性能，同时由于稀疏输入保持了较低的内存占用。据我们所知，这是首次尝试将传统峰值与表示学习结合用于音频指纹识别，并且是第一次使用点云网络进行 AFP。作为后续贡献，我们在一个新的场景中评估了 *PeakNetFP* 与时间拉伸方面的最先进算法四重浮点精度 [21]（这是一种基于峰值的方法）以及最先进神经音频指纹识别技术神经网络指纹 [6]。我们最终展示了峰值网络指纹/*PeakNetFP* 尽管参数和输入数据分别比后者少 100 倍和 11 倍，但仍能达到接近最先进方法神经网络指纹的性能。

在第 2节中，我们总结了与本出版物相关的工作。在第 3节中，我们描述了分层峰值集特征提取以及位于核心的对比表示学习框架峰值网络浮点精度。最后，在第 4节中，我们在极端时间拉伸的背景下展示了其评估，并说明它与基于峰值的时间拉伸基线四倍浮点精度和基于频谱图的最先进模型神经网络指纹相比如何。*PeakNetFP* 代码、数据集和模型是开放且可获取的 [1]。

[1] https://github.com/guillemcortes/peaknetfp

2. 相关工作

在过去二十年里，研究社区致力于推进音频指纹系统以应对多种应用场景。其中一些创新包括用于抗噪声的小波 [22]，用于抗音调变换的恒 Q 变换 [23] 或基频图 [24]，以及用于广播监测的余弦滤波器 [25] 等。

我们可以将 AFP 方法分为三大类：基于局部描述符的 [4,22,24–29]、基于峰值的 [1,7,19,21,23,28,30–32] 和神经音频指纹 [6, 13, 33–37]。基于峰值的指纹始于 Shazam 算法 [1]，该算法奠定了将谱峰对链接以形成抗噪声散列的基础。然后，Six & Leman 提出了链接三元组以获得对抗时间和频率修改的鲁棒性在巴纳科 [7] 中，尽管它不适合短查询或极端时间拉伸，因为它旨在用于通过重放数字化的老录音音频集合的内容去重复。类似地，Sonnleitner & Wilder 提出了四倍浮点数 [21]，该方法将盲天体测量研究 [38] 适应以构建峰值四元组并生成对显著时间和频率修改具有鲁棒性的哈希值 [19]。其他基于峰值的 AFP 工作 [23,28,30–32] 也研究了如何提高对时间和频率修改的鲁棒性。基于峰值的传统方法即使在存在噪声、压缩或混响等改动的情况下也能表现出色。它们通常生成轻量级哈希值，可以高效地索引到查找表中，这使得它们能够扩展到数十万甚至数千万首音乐作品。此外，这类方法不需要训练或加速计算硬件。

2.1 四浮点精度

然而，在存在极端挑战性场景的情况下，如强时间拉伸 [21]，这些传统方法表现显著不佳。四倍浮点精度在这方面作为最先进的峰值基础 AFP 之一脱颖而出。它设计得能够抵御时间和频率上的修改，其核心创新在于使用四重峰描述符，不仅捕捉每个峰值的位置，还捕捉其与邻近峰值的关系。每一个四重描述了时间-频率域中的四个峰值（局部极大值）的星座图，有效地编码了局部模式和峰值之间的关系。这种方法相较于其他指纹识别方法 [21]，如帕纳科 [7]，在抗噪声和音频内容变化方面更为稳健。一旦四重特征被提取，四倍浮点精度使用哈希机制将这些描述符映射到数据库中。在这篇出版物中，我们使用四倍浮点精度作为最先进的时间拉伸鲁棒性的基于峰值的 AFP 基线。这也符合本研究的应用场景，该应用场景限制输入数据为频谱峰值。我们的目标是展示神经网络可以如何改进最佳的传统时间拉伸系统。

2.2 神经网络指纹

在过去十年中，神经网络已被成功用于自动指纹识别。2017 年，谷歌提出了第一个神经 AFP 系统现在

播放 [13] 。它被设计为在有限数据集的移动设备上运行，并具有对噪声的高度鲁棒性。一个非常近期的方法 GraFPrint [39] 利用图神经网络（GNNs）的结构学习能力，从时频表示中生成稳健的指纹。与我们的方法相反，它不使用稀疏谱峰，而是从频谱图中提取位置感知的低维特征，采用卷积编码器。其他神经 AFP 系统 [6, 33–37] 则在对比学习框架中使用有针对性的数据增强来实现对噪声、混响、回声和其他失真的鲁棒性。其中，神经网络指纹 [6] 是唯一一个完全可复现的神经 AFP。该实现是开源的，并且有一个公共数据集和模型权重。神经网络指纹也比其他提出的模型（如基于 transformer 的 AFP [34, 36]）更轻量级。出于这些原因，我们使用神经 *FP* 作为开发 *PeakNetFP* 的基础。神经网络指纹利用对比学习实现高灵敏度的音频检索，采用卷积编码器从梅尔频谱图中提取有意义的特征。在本文中，我们建议重用神经网络指纹的对比学习框架，同时将输入数据从密集谱图更改为稀疏峰值特征。原始的神经网络指纹用作参考模型，展示了当考虑完整的谱图为输入时可以实现的效果。

某些指纹识别方法已被设计用于有效处理时间拉伸，这是本研究的重点。

2.3 时间拉伸的 AFP 方法

四倍浮点精度 [21] 在这个主题上被视为一个里程碑。通过结合基于四元组的频谱峰值分组（参见第 2.1 节）和最大化查询中生成的四元组数量的非对称查询参考指纹配置，他们报告了针对 20 秒查询的多种节奏修改下的高精确度和准确率。他们的参考数据库由来自 Jamendo 的 10 万条音轨组成，并测试了 300 个使用在 70%到 130%之间的 13 种拉伸因子进行时间拉伸的查询音轨。在这个实验中，他们报告对于 10 秒的查询平均准确率为 92.9%，但对于 2.5 秒的查询平均准确率仅为 28.1%。这显示了随着查询长度减小性能是如何崩溃的。如果我们使用更少每秒四元组（在工业环境中更有可能的情况），我们可以预期这种性能会更低。姚等人。[40] 使用与 [21] 相同的数据集。实验是在 70%到 130%之间的 13 种拉伸因子和 20 秒的查询长度下进行的。他们报告了类似四倍浮点精度的性能，但召回率下降了 20%。SAMAF [41] 报告对于从 1 到 6 秒的不同查询长度，他们实现了超过 80%的轻度拉伸（0.9 和 1.1）准确率，但对于重度拉伸（0.5, 1.5），准确率下降到低于 13%。Panako [7] 报告了在包含 30,000 首歌曲的数据库上，对于查询时间分别为 20 秒、40 秒和 60 秒的结果。尽管经过 8%的时间拉伸修改后，不到三分之一的查询得到了正确解答。Son 等人 [24] 在节奏修改范围为 70%到 130%的情况下实现了完美的精度。然而，他们的数据集仅包含 100 个音频文件，并且使用完整的音频长度进行查询。George & Jhunjhunwala [42] 提出使用仅基于频率信息编码特征的方法，从而使其独立于时间，与 [1] 相比，后者是基于时间编码的。他们测试了节奏在 ±50%范围内的修改。他们达到了超过 97%的准确率，但在一个包含 300 个样本的数据集上进行测试，每个样本持续 20 秒。他们的算法也不适用于短查询。

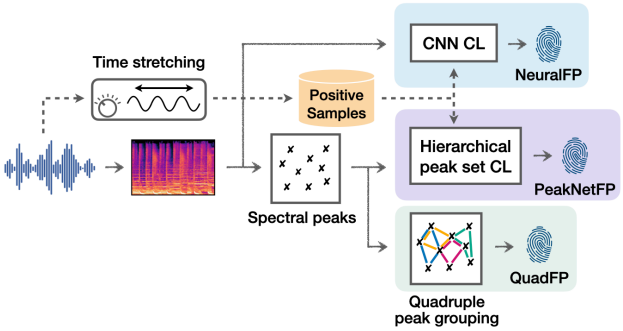

Figure 1. 考虑 AFP 系统概述。从上到下：神经 *FP*，峰值网路浮点精度（我们的），四重浮点精度。虚线表示用于训练的额外数据。我们的模型峰值网 *FP* 在与神经 *FP* 相同的对比学习框架中，从与四倍浮点精度相同输入中学得特征。

3. PEAKNETFP

在本节中，我们描述了从稀疏输入数据到对比学习框架的整个过程中所提出的模型 *PeakNetFP*，重点介绍了分层峰值集特征提取过程。此外，我们还介绍了用于评估的数据集。图 1提供了所有考虑的 AFP 系统的概述，包括我们的 *PeakNetFP*、基线四倍浮点精度 [21] 以及 SOTA AFP 模型神经 *FP* [6]。

3.1 稀疏输入数据

正如我们在介绍中所述，我们的模型设计用于处理以三维频谱峰值形式出现的稀疏数据，这些特征是从第三方传统的 AFP 系统中提取出来的。通常，这样的峰值代表了从声谱图中选择的一组局部最大值，基于识别最显著的那些的标准 [1, 7, 30]。在我们的系统中，我们使用 3x3 内核和步幅为 1 来提取梅尔频谱图中的局部最大值作为传统峰值指纹的一个简化代理，并避免偏向其他系统的标准。这使得神经网络能够学习哪些峰值对于匹配或分类最重要。尽管更精细的峰值选择可以帮助降低输入维度，从而提高计算效率。在实践中，我们每秒段选取 256 个最高振幅的局部最

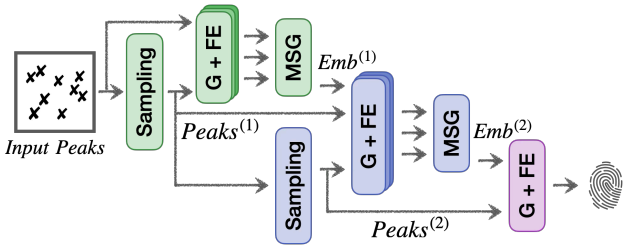

Figure 2. 峰值网络 *FP* 概述包括采样、分组加特征提取（G+FE）和多尺度特征分组层（MSG）。

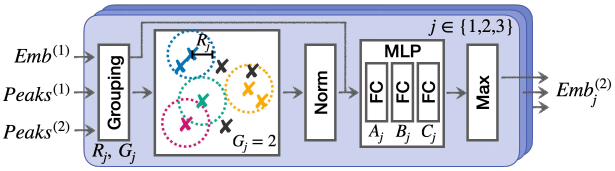

Figure 3. 第二层的分组 + 特征提取块 ($i = 2$)。对于每个特定的子层 j、R_j 和 G_j 分别是查询球的半径和组大小，而 (A_j, B_j, C_j) 是 MLP 层的维度。

大值，确保我们在每个窗口内捕捉到最突出的特征。

当处理峰值而非连续数据点时，将局部性注入卷积核变得具有挑战性，这些卷积核通常依赖于密集的、网格状的输入结构。峰值创建了数据的稀疏表示，就像计算机视觉任务中使用的点云一样。这种稀疏性复杂化了传统卷积方法的直接应用，因为没有内在的邻域结构。然而，像 *PointNet* [43] 和从点云文献中衍生的方法提供了一种潜在解决方案，通过将局部峰值分组并以类似于卷积在频谱图上操作的方式处理它们。通过利用峰值之间的局部关系，我们可以在不需要密集连续输入的情况下捕获有意义的模式。

3.2 分层峰值集特征提取

分层的点网，或 *PointNet++* [20]，引入了一种多层次的方法来捕捉稀疏数据中的局部和全局特征，类似于传统 CNN 中找到的嵌套卷积。*PointNet++*将点（在我们的上下文中是峰值）组织成分层组合，其中局部邻域逐步被采样和处理，类似于卷积跨密集数据扫描的方式。这种分层结构使 *PointNet++*能够有效地从稀疏数据中学习细粒度和高层次特征。在 *PeakNetFP* 中，我们将来自 *PointNet++*的分层峰值集特征提取整合到来自神经 *FP* 的对比学习框架中。图 2描述了 *PeakNetFP* 的架构，而图 3则描述了第二层的分组和特征提取（G+FE）块，在图 2中以蓝色表示。稀疏峰值编码从两个集合抽象（SA）层开始，分别用绿色和蓝色表示在图 2上，负责通过多尺度特征分组（MSG）以层次结构对相邻的峰值进行分组。每一层 i 都经过三个关键步骤操作：

（一）抽样: 具有最大振幅的 $N^{(i)}$ 峰被选为锚峰，这些将是峰值组的中心。这一步有助于控制网络加深时的计算复杂度。

(二) 分组+特征提取块 (G+FE)：每个块由 3 个平行层组成，每层 j 包括：

（i）分组: 对于每个锚点峰，我们选择半径 $R_j^{(i)}$ 内最接近的 $G_j^{(i)}$ 个峰值形成局部邻域。这些邻域充当局部感受野，类似于 CNN 中的卷积块。查询球体是一个关键元素，因为它允许对点云中分层搜索的距离和半径进行精确控制。与依赖固定网格结构的传统卷积不同，查询球体通过根据实际空间接近程度将峰值分组来适应不规则分布。

(ii) 特征提取：在每个邻域内，应用一个具有 $A_j^{(i)}, B_j^{(i)}, C_j^{(i)}$ 维的三层 MLP 来学习局部特征。该 MLP 聚合每个点的特征，并使用最大池化将它们汇总成一个表示局部区域的维度为 $N^{(i)} \times C_j^{(i)}$ 的单向量。

（三）多尺度特征分组：所有并行抽取层的特征被拼接成一个维度为 $N^{(i)} \times (C_1^{(i)} + C_2^{(i)} + C_3^{(i)})$ 的单一嵌入。这一步允许模型在分组阶段使用不同的邻域半径来同时捕获多尺度的特征，有助于考虑细粒度和粗粒度的特征。

每个 SA 层之后，输出是一组维度更高但数量更少的峰值。这些向量被传递到下一个 SA 层，在那里重复这一过程（如图 2中的 $Peaks^{(2)}$ 所示），进一步抽象数据。随着我们深入网络，感受野变大，使网络能够捕捉更广泛的情境信息同时保持局部细节。最后一个 SA 层（在图 2中用紫色表示）类似于一个分组+特征提取模块，但是其中所有个点被组合在一起，形成一个单一的 128 维特征向量。因此，这个最终的向量编码了关于峰值的局部和全局信息，并可以作为指纹使用。

3.3 对比学习框架

峰值网 *FP* 依赖于神经 *FP* 对比学习框架 [6] 来学习指纹，我们将在下文中描述这一点。它操作的是具有 50%重叠的 1 秒窗口。通过应用时间拉伸到短音频片段来创建数据对。每个小批量 MB 由 N 个样本和同一样本的 N 个增强副本组成，以生成正对 x_i 和 x_j，使得 $MB = \{x_i, x_j, ..., x_{Ni}, x_{Nj}\}$ 和 $|MB| = 2N$。NT-Xent 损失 [44] 被选中以最大化小批量中正样本对的一致性 MB。不进行显式的负采样，因此，给定一个正样本对时，其他 $2(N-1)$ 数据点应被视为负样本。给定嵌入

$\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 的 NT-Xent 损失定义为：

$$l(i,j) = -\log \frac{\exp(a_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(a_{i,k}/\tau)} \qquad (1)$$

其中 $a_{i,j} = \boldsymbol{z}_i^T \boldsymbol{z}_j$ 对于 $i,j \in \{1,\ldots,2N\}$ 成立。τ 是 softmax 中的温度缩放因子。在 softmax 函数中计算 Top-1 等同于最大内积搜索（MIPS）。$\mathbb{1}_{[k\neq i]}$ 确保求和排除了锚点-正样本对。然后，损失 $\mathcal{L}$ 在所有正样本对中平均值为 l，包括 (i,j) 和 (j,i)：

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} l(2k-1,2k) + l(2k,2k-1) \qquad (2)$$

在检索过程中，如同 [6] 中所示，我们从使用 *Faiss* [45] 构建的倒排文件产品量化（IVFPQ）索引中检索 20 个候选片段。然后，我们执行一个序列匹配，其中每个片段的嵌入通过内积与候选嵌入进行比较，并根据此分数对结果进行排序。

3.4 数据集

为了开发 *PeakNetFP*，我们使用与神经网络指纹 [6] 中相同的数据库，但将增强方法改为时间拉伸。该数据集由来自中等_fma 数据库 [46] 的多个音频文件组成，并且带有定义好的子集，我们也用这些子集来训练和测试我们的模型。训练子集包含 10,000 个 30 秒的音频片段，而测试-查询/数据库包含 500 个 30 秒的音频片段。为了增加参考集，我们使用测试-占位符-数据库，它包括了每条平均长度为 278 秒的 100,000 首完整曲目。这有助于测试系统的可扩展性。在评估步骤中，我们使用从 500 个片段中随机选择的同样 2,000 个片段进行神经网络指纹评估。

在训练过程中，拉伸增强是在频谱图级别进行的，我们仅使用双线性插值沿时间轴调整大小。与基于波形的拉伸方法从 SOX[2] 不同，这种简单的方法可以轻松集成到训练流程中，并确保我们的模型不会过度拟合特定模型的任何特殊之处，而是学会处理拉伸。在测试时，我们使用 SOX 从测试查询/数据库集的 DB 轨道生成真实的查询。我们为增加歌曲速度的拉伸因子 1.05、1.1、1.2、1.4、1.6、1.8 和 2 以及减慢速度的对应值 0.975、0.95、0.9、0.8、0.7、0.6 和 0.5 生成查询。请注意，拉伸因子为 2 时速度翻倍而 0.5 时则减半。此测试集在 Zenodo[3] 中公开可用。

[2] https://sourceforge.net/projects/sox/

[3] https://zenodo.org/records/15646861

4. 评估

在本节中，我们介绍了评估框架的特异性以及所使用的度量和结果。最后，我们也传播了基准系统的计算成本。

4.1 评估框架

峰值网 *FP* 评估严格基于神经网络指纹，以实现公平比较，并可在随附此出版物的存储库中进行检查。我们训练了 *PeakNetFP* 和神经 *FP*，使用了一个大小为 240 的批量，在 100 个周期内采用了 Adam 优化器 [47]，遵循作者的建议 [6]。表 1概述了 *PeakNetFP* 层的参数，包括锚点峰值的数量 N、查询球半径 R 以及每个查询球分组中的峰值数量 G。

由于没有公开的四倍浮点精度实现，我们根据原始论文创建了自己的版本 [21]。虽然并未提供所有实现细节以供精确复现，但我们严格遵循了模型的关键方面，例如为查询计算更多四元组而非参考，并在比较过程中应用级联启发式方法高效过滤无关的四元组。我们在第 4.2节通过将我们的结果与原始出版物中的结果进行对比来验证我们的实现 [21]。

我们评估了在时间拉伸从极端值 0.5 倍到 2 倍原始速度的情况下 *PeakNetFP*、四倍浮点精度以及神经网络指纹。此外，我们使用长度为 2 秒、3 秒、5 秒、6 秒和 10 秒的查询对每个模型进行了测试，以确保其适用于示例查询应用。未考虑 1 秒的查询，因为它们在时间因素超过 1 时会导致大小小于神经 *FP* 窗口大小。

为了公平地比较 AFP 系统，我们按照文献 [6] 中使用的方法，采用 Top-1 命中率 HR@1，定义为 Top-1 命中的数量除以查询的数量。请注意，神经网络指纹和峰值网 *FP* 总是返回一个匹配结果，因此在这种情况

					多层感知器		
Layer	N	j	G	R	A	B	C
SA + 消息 1	200	1	4	0.1	16	16	32
		2	8	0.2	32	32	64
		3	16	0.3	32	48	64
SA + MSG 2	100	1	4	0.2	32	32	64
		2	8	0.3	64	64	128
		3	16	0.4	64	64	128
SA					128	256	128

Table 1. 峰值网络 *FP* 层及其参数：锚点数量 N，层索引 j，分组峰值数量 G，分组半径 R，以及 3 个 MLP 层的维度 A、B 和 C。

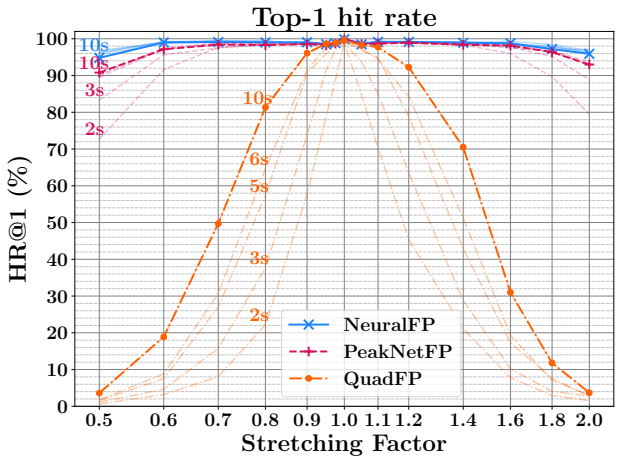

Figure 4. Top-1 击中率（HR@1）作为拉伸因子的函数，针对神经 *FP*、*PeakNetFP*（我们的模型）和四倍浮点精度。每条曲线代表一个（模型，查询长度）配对。

下，Top-1 命中率等同于精确率和召回率。未来的工作可以包括训练一个分类器来适应超出词汇表的查询。

4.2 结果

为了验证我们自定义实现的有效性，我们将我们的四浮点精度结果与表 I 中报告的 [21] 进行了比较。我们的四倍浮点精度实现在短查询（≤5 秒）中显示出了比原实现更好的结果，对于 2 秒的查询，平均值为 48%HR@1，而 [21] 报告的 2.5 秒查询的结果为 28%，而对于较大的查询结果则略微差一些，我们的实现为 94%，而 [21] 的 20 秒查询结果为 98%。我们承认两个研究的数据集有所不同（FMA 与 Jamendo），但我们假设由于它们的规模和性质相似，AFP 系统中的比较仍然是有效的。总之，我们得出结论，我们的实现结果与 [21] 相当，剩余的差异来自于评估数据集或实现差异。

图 4展示了所有 3 个模型的 Top-1 命中率 HR@1 作为拉伸因子的函数。对于每个模型，我们报告了 5 条不同的曲线，这些曲线对应于测试的 5 种查询长度，在蓝色实线中表示神经网络指纹，橙色虚点线中表示四重浮点精度，以及在品红色虚线中表示我们的模型 *PeakNetFP*。我们突出显示了对应 10 秒查询的曲线，以与我们的基准四重浮点精度最佳设置进行比较。

PeakNetFP 全球范围内优于四倍浮点数，有效处理时间拉伸。它还表现出卓越的性能，在通常报告的 0.7 到 1.4 倍的时间拉伸因子范围内，对于 10 秒的查询实现了超过 98%的 HR@1 成绩 [21, 24, 40]。对于极端拉伸因子（小于 0.7 和大于 1.4），*PeakNetFP* 的性能略有下降，但仍保持在 90%以上 HR@1。作为参考，四倍浮点精度在 0.5 倍的因子下仅实现了 3.6%的 HR@1。事实上，我们观察到四倍浮点精度在轻微拉伸（0.9 到 1.1）时表现出色如之前报告的 [21]，但随着拉伸偏离 1 越来越远，其性能迅速下降，在极端拉伸因子为 0.5 和 2 时几乎达到零 HR@1。这种效应可能是由于在如此强烈的时间拉伸因子下保留的四边形不足所致。就查询长度而言，四倍浮点精度的性能随着查询变小而迅速下降。这里应该提醒的是，四倍浮点精度是一种基于规则的算法，不需要训练，这与 *PeakNetFP* 或神经网络指纹不同。

SOTA 模型神经 *FP* 在时间拉伸方面表现出强大的鲁棒性，即使在极端情况下也是如此。作为提醒，神经网络指纹处理整个频谱图而 *PeakNetFP* 仅处理稀疏峰值。然而，对于通常报告的时间拉伸因子（0.7 到 1.4），*PeakNetFP* 和神经网络指纹获得几乎相同的表现，最大差异为 ±0.7%HR@1。对于更极端的因素，两个系统的性能都下降了，其中 *PeakNetFP* 受到的影响比神经网络指纹更大，系统之间的最大差异在因素为 0.5 时达到 1.85%HR@1。关于查询长度，*PeakNetFP* 至少需要 5 秒的查询才能使性能在极端时间因素下系统性地保持在 0.9 以上。

我们得出结论，*PeakNetFP* 的性能与神经网络指纹相当，在极端拉伸因子下略有下降。然而，*PeakNetFP* 比神经 *FP* 显著更轻。它使用了 256 个 3D 峰值作为输入，大约是神经网络指纹的 256 × 32 频谱图的 1/11。具有 169k 个可训练参数，*PeakNetFP* 的模型大小比神经 *FP* 的 16.9M 小 100 倍。这还需要大幅减少推理内存：对于 *PeakNetFP* 为 800MiB，而对于神经网络指纹法为 2338 MiB（单个 RTX 3090 上的批次大小为 125）。这提高了我们模型的可扩展性，并减少了目录嵌入生成所需的内存使用。

5. 结论

在这项工作中，我们介绍了一种新的音频指纹系统，*PeakNetFP*，它被设计为一种结合了传统峰值基础指纹系统的优点和现代基于神经网络的表示学习方法优势的混合方法。我们在对比学习方法中使用了一个受计算机视觉启发的点云网络来处理稀疏峰值，这种方法类似于现代 AFP 方法。

在极端时间拉伸背景下的评估显示，*PeakNetFP* 在时间拉伸数据上始终优于 SOTA 系统，四倍浮点精度。此外，我们还展示了基于频谱图的 AFP SOTA 系统神经网络指纹在这种任务中表现非常好，而我们的模型 *PeakNetFP* 可以达到可比较的表现，同时只使用峰值工作，并且输入数据大小仅为前者的 1/11，参数量也

少了 100 倍。

总之，*PeakNetFP* 为涉及显著节奏变化的音频识别任务提供了一种可扩展且高效的解决方案，结合了基于峰值方法的紧凑性与神经网络的强大性和灵活性。它在严重到极端拉伸因素方面优于传统方法，并作为完全基于神经网络的方法的一种替代方案出现，特别是在内存和计算效率至关重要的情况下。未来的工作将集中在改进模型以应用于时间拉伸之外的应用，如音调变化。

6. 致谢

本研究是 *resCUE – 智能系统用于音频视觉制作中音乐作品的自动使用报告 (SAV-20221147)* 项目的一部分，该项目由西班牙工业、贸易和旅游部与欧盟下一代计划共同资助，并得到西班牙科学、创新和大学部以及数字转型和公共职能部的支持。此外，该研究还得到了加泰罗尼亚自治区企业与知识部、大学与研究秘书处的工业博士计划的资金支持，资助协议编号为 DI46-2020。

7. REFERENCES

[1] A. Wang, “An industrial strength audio search algorithm,” in *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR 2003)*, Baltimore, Maryland, USA, 2003, pp. 7–13.

[2] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR 2002)*, Paris, France, 2002, pp. 107–115.

[3] E. Gomez, P. Cano, L. Gomes, E. Batlle, and M. Bonnet, “Mixed watermarking-fingerprinting approach for integrity verification of audio recordings,” in *Proceedings of the International Telecommunications Symposium*, Natal, Brazil, 2002.

[4] C. Ouali, P. Dumouchel, and V. Gupta, “A robust audio fingerprinting method for content-based copy detection,” in *12th International Workshop on Content-Based Multimedia Indexing (CBMI 2014)*, Klagenfurt, Austria, 2014, pp. 1–6.

[5] R. Sonnleitner, A. Arzt, and G. Widmer, “Landmark-based audio fingerprinting for DJ mix monitoring,” in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York City, New York, USA, 2016, pp. 185–191.

[6] S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, “Neural audio fingerprint for high-specific audio retrieval based on contrastive learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, Toronto, Ontario, Canada, 2021, pp. 3025–3029.

[7] J. Six and M. Leman, “Panako: A scalable acoustic fingerprinting system handling time-scale and pitch modification,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, 2014, pp. 259–264.

[8] H. Kim, J. Kim, J. Park, S. Kim, C. Park, and W. Yoo, “Background music monitoring framework and dataset for tv broadcast audio,” *ETRI Journal*, 2024.

[9] J. Six, “Olaf: Overly lightweight acoustic fingerprinting,” in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*, Montréal, Canada, 2020.

[10] A. L.-C. Wang and D. Culbert, “Robust and invariant audio pattern matching,” United States Patent US007 627 477B2, 2009, Shazam Investments Ltd. and Apple Inc.

[11] A. Master, B. Mont-Reynaud, K. Mohajer, and T. Stonehocker, “Systems and methods for providing identification information in response to an audio segment,” United States Patent US10 657 174B2, 2020, SoundHound, Inc.

[12] K. Akesbi, “Audio denoising for robust audio fingerprinting,” Master’s thesis, Ecole normale supérieure Paris-Saclay, Paris, France, 2022.

[13] B. Gfeller, B. Aguera-Arcas, D. Roblek, J. D. Lyon, J. J. Odell, K. Kilgour, M. Ritter, M. Sharifi, M. Velimirovi, R. Guo, and S. Kumar, “Now playing: Continuous low-power music recognition,” in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017) Workshop: Machine Learning on the Phone*, Long Beach, CA, USA, 2017.

[14] G. Cortès, A. Ciurana, E. Molina, M. Miron, O. Meyers, J. Six, and X. Serra, "BAF: An audio fingerprinting dataset for broadcast monitoring," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, Bengaluru, India, 2022, pp. 908–916.

[15] A. Wang, "The shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[16] S. Bilobrov, "Indexing based on time-variant transforms of an audio signal's spectrogram," United States Patent US10 418 051B2, 2019, Facebook, Inc.

[17] M. Pfister, R. Michael, M. Boll, C. Körfer, K. Rieck, and D. Arp, "Listening between the bits: Privacy leaks in audio fingerprints," in *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, Cham, 2024, pp. 184–204.

[18] D. Schwarz and D. Fourer, "Unmixdb: A dataset for dj-mix information retrieval," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, Paris, France, 2018.

[19] R. Sonnleitner, "Audio identification via fingerprinting. achieving robustness to severe signal modifications," PhD thesis, Johannes Kepler University Linz, Linz, Österrreich, 2017.

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: deep hierarchical feature learning on point sets in a metric space," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, p. 5105 – 5114.

[21] R. Sonnleitner and G. Widmer, "Robust quad-based audio fingerprinting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 409–421, 2016.

[22] S. Baluja and M. Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern Recognition*, vol. 41, no. 11, pp. 3467–3480, 2008.

[23] S. Fenet, G. Richard, and Y. Grenier, "A scalable audio fingerprint method with robustness to pitch-shifting," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, Florida, USA, 2011, pp. 121–126.

[24] H. Son, S. Byun, and S. Lee, "A robust audio fingerprinting using a new hashing method," *IEEE Access*, vol. 8, pp. 172 343–172 351, 2020.

[25] M. Ramona and G. Peeters, "Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*. Vancouver, BC, Canada: IEEE, 2013, pp. 818–822.

[26] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," *Content Based Multimedia Indexing, Brescia, Italy*, 2001.

[27] X. Anguera, A. Garzon, and T. Adamek, "MASK: robust local features for audio fingerprinting," in *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME*. Melbourne, Australia: IEEE Computer Society, 7 2012, pp. 455–460. [Online]. Available: https://doi.org/10.1109/ICME.2012.137

[28] E. Dupraz and G. Richard, "Robust frequency-based audio fingerprinting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, Texas, USA, 2010, pp. 281–284.

[29] A. Agarwaal, P. Kanaujia, S. S. Roy, and S. Ghose, "Robust and lightweight audio fingerprint for automatic content recognition," 2023. [Online]. Available: https://arxiv.org/abs/2305.09559

[30] R. Sonnleitner and G. Widmer, "Quad-based audio fingerprinting robust to time and frequency scaling," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, 9 2014, pp. 173–180.

[31] M. Malekesmaeili and R. K. Ward, "A local fingerprinting approach for audio copy detection," *Signal Process.*, vol. 98, pp. 308–321, 2014.

[32] J.-Y. Lee and H.-G. Kim, "Audio fingerprinting using a robust hash function based on the MCLT peak-pair,"

The Journal of the Acoustical Society of Korea, vol. 34, no. 2, pp. 157–162, 2015.

[33] Z. Yu, X. Du, B. Zhu, and Z. Ma, "Contrastive unsupervised learning for audio fingerprinting," *Computing Research Repository (CoRR)*, 2020.

[34] A. Singh, K. Demuynck, and V. Arora, "Attention-based audio embeddings for query-by-example," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, Bengaluru, India, 2022, pp. 52–58.

[35] X. Wu and H. Wang, "Asymmetric contrastive learning for audio fingerprinting," *IEEE Signal Process. Lett.*, vol. 29, pp. 1873–1877, 2022.

[36] A. Singh, K. Demuynck, and V. Arora, "Simultaneously learning robust audio embeddings and balanced hash codes for query-by-example," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, 2023, pp. 1–5.

[37] Y. Fujita and T. Komatsu, "Audio fingerprinting with holographic reduced representations," in *25th Annual Conference of the International Speech Communication Association (Interspeech 2024)*, Kos, Greece, 2024.

[38] D. Lang, D. W. Hogg, K. Mierle, M. Blanton, and S. Roweis, "Astrometry. net: Blind astrometric calibration of arbitrary astronomical images," *The astronomical journal*, vol. 139, no. 5, p. 1782, 2010.

[39] A. Bhattacharjee, S. Singh, and E. Benetos, "Grafprint: A gnn-based approach for audio identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, Hyderabad, India, 2025.

[40] S. Yao, B. Niu, and J. Liu, "Enhancing sampling and counting method for audio retrieval with time-stretch resistance," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 2018.

[41] A. Báez-Suárez, N. Shah, J. A. Nolazco-Flores, S.-H. S. Huang, O. Gnawali, and W. Shi, "SAMAF: Sequence-to-sequence autoencoder model for audio fingerprinting," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, 2020.

[42] J. George and A. Jhunjhunwala, "Scalable and robust audio fingerprinting method tolerable to time-stretching," in *2015 IEEE International conference on digital signal processing (DSP)*, 2015, pp. 436–440.

[43] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML), PmLR*, 2020, pp. 1597–1607.

[45] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[46] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 316–323.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, 2015.