arXiv:2603.04032v1 [cs.SD] 4 Mar 2026

采用 BandSplit-RoFormer 分离与 HiFi++ GAN 的多阶段音乐源修复

Tobias Morocutti⋆、*Emmanouil Karystinaios*⋆、*Jonathan Greif*、*Gerhard Widmer*

计算感知研究所（CP-JKU），林茨约翰开普勒大学，奥地利 firstname.lastname@jku.at

摘要

音乐源修复（MSR）旨在从完全混音并完成母带处理的音频中恢复原始、未经处理的乐器音轨，而制作效果和分发伪影会违背常见的线性混合假设。本技术报告介绍了 CP-JKU 团队参加 MSR ICASSP 2025 挑战赛的系统。我们将 MSR 分解为分离和修复两个步骤。首先，使用单个 BandSplit-RoFormer 分离器预测八条音轨以及一条辅助其他声部音轨，并采用三阶段课程进行训练：从四轨热启动微调（使用 LoRA）逐步过渡到通过输出头扩展实现八轨扩展。其次，我们采用 HiFi++ GAN 波形修复器，先将其训练为通用模型，再专门化为八个乐器专用专家。

索引词— 音乐源修复、音源分离、生成式音频修复、BS-RoFormer。

1. 引言

专业音乐制作违背了传统音乐源分离方法通常采用的线性混合假设。发行流程通常包括均衡、动态范围压缩、混响、饱和失真和失真、立体声拓宽、限幅以及编解码器伪影，而且往往还叠加了其他劣化。因此，目标音源不仅经过混合，还受到系统性变换，使得在干净音轨假设下直接进行分离并不足够。因此，MSR 挑战赛旨在从此类混合音频中恢复八类乐器的原始未处理音源（人声、吉他、键盘、合成器、贝斯、鼓、打击乐、管弦乐）。

我们采用模块化学习框架来解决 MSR 问题，明确将去混音问题与去母带处理和伪影去除问题分开。具体而言，我们采用先分离、后修复的策略：由一个多声部音源分离器估计退化音轨，再由针对各音轨的修复网络将这些估计结果映射为干净目标音轨。这种设计使我们能够在大规模增强混合音频上训练分离器，同时训练修复模型，使其能够在符合实际情况的分离器误差分布下逆转制作效果并抑制伪影。在本次提交中，我们提出两项实用贡献：(i) 一套将 BS-RoFormer 从 4 个音轨适配至 8 个音轨的课程式方案，采用参数高效 LoRA 微调和输出头扩展；(ii) 使用分离器生成的输入训练的乐器专属修复专家，以改善训练与测试的一致性。

⋆：作者贡献均等

2. 系统概述

我们的系统通过两阶段流水线处理混合波形 x。在第一阶段，单个多声部音源分离器 $\mathcal{S}$会估计八个目标音轨以及一个辅助其他 音轨：

$$\hat{s}_k = \mathcal{S}(x)_k, \quad k \in \{1, \ldots, 8\}, \qquad \hat{s}_{\text{other}} = \mathcal{S}(x)_{\text{other}}. \tag{1}$$

在第二阶段，乐器专用修复专家$\mathcal{R}_k$ 将每个分离后的估计映射为修复后的音轨：

$$\tilde{s}_k = \mathcal{R}_k(\hat{s}_k), \quad k \in \{1, \ldots, 8\}. \tag{2}$$

为改善修复任务中的训练与测试对齐，我们将训练好的分离器应用于合成训练混音，生成修复输入；随后训练每个专家，将得到的分离器输出映射为对应的干净声部，使修复器能够接触到真实的分离误差。[1]

2.1. 声源分离

我们使用 BandSplit-RoFormer (BS-RoFormer) 分离器 [1]。该模型采用频带拆分前端，分别处理不同频率区域，并使用 RoFormer 模块对时间依赖关系和跨频带依赖关系进行建模，以估计掩码。单个模型通过 9 个掩码估计头部预测所有音轨（8 种乐器类别加上辅助的其他）。

我们分三个阶段训练分离器，起始点是一个公开提供的四音轨 BS-RoFormer 检查点：

[1]代码和模型详情见：https://github.com/CPJKU/music-source-restoration。

• **阶段1**（四音轨，干净混音）。在干净混音上进行微调（不添加额外的数据增强或母带处理），以分离人声、鼓、贝斯和其他音轨。

• **阶段2**（四音轨、母带处理混音）。继续在混音上微调，这些混音由在线的逐音轨降质流程生成，随后再进行混音/母带处理。目标音轨为降质后的音轨。

• **阶段3**（8音轨）。将模型扩展为8音轨。共享层和原有的四个头部从阶段2初始化。新头部随机初始化，并且仅训练新的掩码头，同时冻结骨干网络。

在阶段1和阶段2中，我们使用LoRA微调Transformer层。我们采用以下损失的加权组合：i) 掩码SI-SNR损失，ii) 多分辨率STFT损失，iii) L1损失，以及iv) 低振幅惩罚损失。

对于四音轨阶段，我们结合使用MUSDB18-HQ [2], DSD100 [3], 和MoisesDB；对于八音轨阶段，我们结合使用MoisesDB [4], Slakh2100 [5], MedleyDB v2 [6], Raw-Stems [7], 以及MUSDB25风格扩展 [8]。我们对每个音源分别应用在线增强和降质策略，然后采用混音和母带处理策略，将降质后的音源合成为最终混音；训练分离器以预测这些降质后的音轨。模型使用4块NVIDIA H100 GPU进行训练。

2.2. 修复

修复模块是一个 HiFi++GAN 组合，包含 SpectralUNet 前端、上采样阶段、Wave-UNet 细化网络，以及用于残差频谱校正的 SpectralMaskNet。我们首先训练一个通用修复模型，然后将其微调为八个乐器专用专家（每个目标音轨对应一个），从而构建一个混合专家系统，其路由由已知的音轨身份决定。

我们按照 [9] 为阶段 1–3 提出的策略，将修复系统分五个阶段进行训练：

• **阶段 1–3：** 还原原始音乐内容（通用模型预热），随后引入 GAN训练 和特征匹配，再引入我们的音乐感知指标。

• **阶段 4：** 通过额外的数据增强（包括留声机噪声），重点抑制噪声伪影。

• **阶段 5：** 使用由分离模型在分离训练混合音频上生成的乐器筛选训练对，对八个乐器专用专家进行微调。

通用修复模型在 Sonic-Master 数据集 [10], 上训练，随后我们引入一个以噪声为重点的训练阶段，使用 Gramophone RecordNoise Dataset [11]对数据进行增强。为实现专家特化，我们在约 8 万个摘录对样本上进行微调，这些样本提取自我们的八源分离数据，其中修复输入通过运行分离推理生成，以匹配测试时的误差特征。修复训练在单块 NVIDIA RTX PRO 6000 Blackwell GPU 上进行。

表 1。MSR Challenge 2025 中*CP-JKU* 各音轨的客观评测结果。最后一行报告未加权平均值。

Stem	MMSNR	Zimt	FAD
Bass	1.5486	0.0119	0.4490
鼓	0.9552	0.0182	0.4912
吉他	1.2919	0.0194	0.5375
键盘	0.7831	0.0174	0.7202
管弦乐元素	0.5125	0.0208	0.7077
打击乐器	0.0994	0.0169	0.8479
合成器	0.6376	0.0271	0.9153
人声	0.8351	0.0193	0.3092
平均值	**0.8329**	**0.0189**	**0.6223**

3. 结果与局限性

表 1 总结了我们在 MSR Challenge 2025 测试集上的官方结果。系统取得了具有竞争力的客观性能（MMSNR 0.8329、Zimtohrli 0.0189、FAD 0.3814），系统级平均意见分为 3.5510。除了官方评估外，我们还报告了后续模型迭代在 MSRBench 数据集 [12]上的平均得分，达到。638 FAD 和 2。338 多梅尔信噪比（MMSNR）。

我们流程的一项关键局限是对含噪混音较为敏感，尤其是现场与历史录音：在此类情况下，分离器可能无法生成足够忠实的音轨估计，进而限制下游修复。我们还观察到，数据集不匹配和错位，尤其是在 RawStems 这类大规模数据集中，可能会使修复训练产生偏差并导致残余噪声伪影。最后，训练用真值音轨中包含的时变效果（例如混响、合唱效果或延迟）使得始终一致的“干声”目标变得模糊，因为模型实际上被要求移除参考音轨中可能存在的效果。因此，我们认为当前的大多数失效模式都源于数据质量和领域覆盖，并计划在未来工作中优先推进数据整理、对齐验证和效果感知条件化。

4. 结论与未来工作

我们提出了一种两阶段音乐源修复（MSR）系统，将经过课程学习训练的 BS-RoFormer 分离器与针对不同乐器的 HiFi++GAN 修复专家相结合。该分离器通过 LoRA 进行适配，并经分阶段训练从四音轨扩展至八音轨预测；修复模型则先作为通用模型进行训练，再使用分离器生成的输入针对不同乐器进行专门化，以匹配测试时误差。

致谢

本研究得到欧洲研究委员会（ERC）地平线2020项目资助，项目编号为 #101019375，项目名称为“WhitherMu- sic?”。

5. 参考文献

[1] Wei-Tsung Lu、Ju-Chiang Wang、Qiuqiang Kong 和 Yun-Ning Hung，《基于频带分割 RoPE Transformer 的音乐源分离》，收录于*ICASSP 2024-2024 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024 年，第 481–485 页。

[2]Zafar Rafii、Antoine Liutkus、Fabian-Robert St¨oter、Stylianos Ioannis Mimilakis和Rachel Bittner，“Musdb18-hq——musdb18的未压缩版本”，2019年8月。[3]Antoine Liutkus、Fabian-Robert St¨oter、Zafar Rafii、Daichi Kitamura、Bertrand Rivet、Nobutaka Ito、Nobutaka Ono和Julie Fontecave，“2016年信号分离评估活动”，载于潜变量分析与信号分离——第*12*届国际会议，*LVA/ICA 2015*，捷克共和国利贝雷茨，*2015*年*8*月*25*日至*28*日，会议论文集，Petr Tichavsk´y、Massoud Babaie-Zadeh、Olivier J.J. Michel和Nad`ege Thirion-Moreau编，尚，2017年，第323–332页，施普林格国际出版。[4] Igor Pereira、Felipe Ara´ujo、Filip Korzeniowski和Richard Vogl，“Moisesdb：面向四音轨以外声源分离的数据集”，2023年。[5] Ethan Manilow、Gordon Wichern、Prem Seetharaman和Jonathan Le Roux，“给音乐源分离一点Slakh：用于研究训练数据质量与数量影响的数据集”，载于*IEEE*音频与声学信号处理应用研讨会论文集（*WASPAA*）。IEEE，2019年。[6] Rachel M Bittner、Julia Wilkins、Hanna Yip和Juan P Bello，“MedleyDB 2.0：新数据与可持续数据采集系统”，*ISMIR*最新进展与演示论文，第36卷，2016年。[7] Yongyi Zang、Zheqi Dai、Mark D Plumbley和Qiuqiang Kong，“音乐源修复”，*arXiv*预印本 *arXiv:2505.21827*，2025年。[8] Karn Whatsarachupat及其合作者，“Musdb25：面向音乐源分离的完整多轨数据集”，数据集预印本，2025年。[9] Nicholas Babaev、Kirill Tamogashev、Azat Saginbaev、Ivan Shchekotov、Hanbin Bae、Hosang Sung、WonJun Lee、Hoon-Young Cho和Pavel Andreev，“终于：快速且通用的录音室级质量语音增强”，2024年。[10] Jan Melechovsky、Ambuj Mehrish和Dorien Herremans，“Sonicmaster：迈向可控的一体化音乐修复与母带处理”，2025年。[11]Eloi Moliner和Vesa V¨alim¨aki，“用于历史录音高保真降噪的两阶段U-Net”，载于*ICASSP 2022——2022*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第841–845页。[12]Yongyi Zang、Jiarui Hai、Wanying Ge、Qiuqiang Kong、Zheqi Dai、Helin Wang、Yuki Mitsufuji和Mark D Plumbley，“MSRBench：音乐源修复基准数据集”，*arXiv*预印本 *arXiv:2510.10995*，2025年。