arXiv:2601.04343v1 [cs.SD] 7 Jan 2026

摘要 首届音乐源恢复挑战赛总结 E

Yongyi Zang[1] *Jiarui Hai*[2] *Wanying Ge*[1] *Qiuqiang Kong*[3]
Zheqi Dai[3] *Helin Wang*[2] *Yuki Mitsufuji*[4] *Mark D. Plumbley*[5]

[1]独立研究者 [2]约翰斯·霍普金斯大学
[3]香港中文大学 [4]Sony AI [5]伦敦国王学院

摘要

音乐源恢复（MSR）旨在从经过专业混音和劣化的音频中恢复未经处理的原始乐器分轨，这需要逆转制作效果以及现实世界中的劣化。我们推出了首届 MSR 挑战赛，采用多梅尔信噪比、Zimtohrli 和 FAD-CLAP 对录音室制作的混音进行客观评估，并对现实世界中劣化的录音进行主观评估。共有五支团队参赛。获胜系统的多梅尔信噪比达到 4.46 分贝，MOS-总体达到 3.47，相较于第二名系统，分别提升了 91% 和 18%。分轨分析表明，不同乐器的恢复难度差异显著：所有团队的低音平均为 4.59 分贝，而打击乐平均仅为 0.29 分贝。数据集、评估协议和基线系统可在 https://msrchallenge.com/ 获取。

索引词— 音乐源恢复、音频信号处理、深度学习

1. 引言

音乐源分离（MSS）传统上假设混音是各个源信号 [1, 2] 的线性组合。然而，真实录音并不符合这一假设。在制作过程中，音频工程师会使用均衡、动态范围压缩、混响和母带处理效果，改变每种乐器的频谱和时间特性。在传输和存储过程中，录音还会因有损音频编解码器 [3, 4] 以及噪声、失真等声学伪影而进一步劣化。音乐源恢复（MSR） [5] 将 MSS 扩展至应对这些复杂情况，从劣化的混音中恢复未经处理的原始源信号。

MSS 社区通过音乐解混挑战赛 [6] 和声音解混挑战赛 [7, 8]，等竞赛，建立了严格的评估框架，并采用 MUSDB18-HQ [9] 和 MoisesDB [10] 等数据集。然而，这些基准无法评估恢复保真度，因为其中的真实音轨已经包含混音时施加的制作效果。MSR 挑战赛填补了这一空白，提供首个采用未经处理的参考音轨的基准，从而能够同时评估分离准确度和恢复质量。

本文在第 2 节介绍挑战赛设置，在第 3 节展示结果，在第 4 节总结参赛系统，并在第 5 节讨论主要发现。

2. 挑战赛设置

2.1. 任务与数据

该挑战任务要求系统从混音音频中分离出八个乐器分轨：人声、吉他、键盘乐器、低音、合成器、鼓、打击乐和管弦乐元素。系统以劣化的混音作为输入，必须输出每个分轨在应用任何制作效果之前的原始、未处理版本。

名为 MSRBench [11], 的验证集包含 2,000 段由专业人员混音的 10 秒片段，采样率为 48 kHz、立体声；每段片段均配有对应的未处理音轨和处理后音轨。混音在 13 种条件下进行评估：原始母带音频以及 12 种退化类型。这些退化类型涵盖模拟伪影（无线电传输、盒式磁带、黑胶唱片、现场房间声学）、传统有损编解码器（AAC 和 MP3，码率为 64 和 128 kbps） [12], , 以及神经音频编解码器（DAC [4] 和 Encodec [3]）。

本挑战赛包含两个测试集。非盲测试集包含 1,000 段片段，并提供真实音轨，可用于计算客观指标。盲测集包含 500 段代表真实世界退化场景的片段，且不提供真实音轨：20 世纪初的历史圆筒录音、现场音乐会录音、调频广播，以及低码率流媒体音频。该盲测集用于检验模型对训练期间未见过的退化类型的泛化能力。

2.2. 评估指标

对于非盲测试集上的客观评估，我们采用三种互补指标。多梅尔信噪比（Multi-Mel-SNR）在多种时频分辨率下衡量频谱重建准确性，旨在避免传统波形域指标（如信号失真比 SDR）对相位过度敏感的问题 [2]。Zimtohrli [13] 基于心理声学原理（包括伽马音调滤波器组分析和时间掩蔽）对感知相似度进行建模。FAD-CLAP [14]通过计算 CLAP嵌入上的 Fr´echet 距离来捕捉语义相似度，从而衡量恢复后的音频听起来是否对应正确的乐器。

对于盲测集上的主观评估，专业音频工程师使用 5 分制平均意见分（MOS），从三个维度对每个恢复样本进行评分。MOS分离度衡量输出从混合音频中分离目标乐器的效果。MOS恢复度评估制作效果和音频劣化是否得到有效去除。MOS总体评分反映恢复音轨的综合感知质量。

3. 结果

五支团队提交了结果：xlancelab、CUPAudioGroup、ACDC、Hachimi 和 cp-jku。表1和表2展示了总体客观与主观结果，表3给出了各分轨的多梅尔信噪比得分。

xlancelab团队在所有指标中均排名第一，相较于排名第二的CUPAudioGroup，在多梅尔信噪比方面相对提升了91%，在MOS-总体方面提升了18%。客观和主观排名高度一致（斯皮尔曼

表1。非盲测试集上的客观评估结果。MM-SNR：多梅尔信噪比（分贝，↑）；Zimtohrli（↓）；FAD-CLAP（↓）。

Team	MMSNR	Zimt	FAD
xlancelab	4.46	0.014	0.199
CUPAudioGroup	2.34	0.016	0.225
AC D̲C	1.45	0.018	0.291
Hachimi	2.00	0.018	0.294
cp-jku	0.83	0.019	0.381

表2。盲测集上的主观评估结果（MOS，采用 1–5 分制，↑）。

Team	Sep	Rest	总体
xlancelab	4.24	3.39	3.47
CUPAudioGroup	3.84	2.92	2.93
Hachimi	3.58	2.63	2.72
AC D̲C	3.54	2.48	2.54
cp-jku	3.55	2.08	2.14

$\rho = 0.9$），只有AC DC和Hachimi在两种评估路径中的排名互换。

表3中的分轨结果表明，不同乐器类型的修复难度存在显著差异。在所有团队的平均结果中，低音得分最高，为4.59分贝，其次是鼓（3.42分贝）和键盘乐器（3.08分贝）。打击乐始终极具挑战性，平均得分仅为0.29分贝，且五个团队中有四个得分低于0.2分贝。值得注意的是，尽管人声是传统音乐源分离研究的主要关注对象之一 [15, 16]，其平均得分也仅为1.17分贝。xlancelab系统在复音乐源上的优势最为明显：其在管弦乐元素和键盘乐器上的得分分别比第二名团队高出4.72分贝和3.37分贝，而在人声上的优势则缩小至仅0.20分贝。

4. 参赛系统

所有参赛团队均采用了基于变换器的架构，这类架构已被证明对音乐源分离十分有效。下面将对各个系统进行简要介绍。

xlancelab团队采用了串联式BSRoformers [17, 18],频带拆分变换器架构，依次应用三个预训练模块：分离、去混响和降噪。他们使用L1损失，并结合短时傅里叶变换多分辨率损失，在MoisesDB [10] 和经人工清理的Raw-Stems数据集 [5]上进行训练。

CUPAudioGroup构建了一个集成系统，融合了三种互补架构：BSRNN [19]（频带拆分循环神经网络）、BSRoformer [17], 和MDX23。所有模型均以预训练权重初始化，并在RawStems、MUSDB18-HQ [9], 和MoisesDB上进行训练。

AC DC团队提出了DTT-BSR，这是一种新颖的架构，结合了DTTNet（一种双路径TFC-TDF U-Net） [20] 、频带序列建模 [19] 和RoPE Transformer瓶颈层。他们采用多频率判别器进行对抗训练。

Hachimi采用了梅尔频带分离骨干网络 [18] ，结合重建损失和GAN损失。他们使用了最多样化的训练数据，结合了六个数据集：MUSDB18-HQ、MoisesDB、MedleyDB、 [21], RawStems、URMP、 [22], 以及MAE-STRO [23]。

cp-jku团队使用 BSRoformer [17] 进行分离，并使用 HiFi++ GAN 套件（SpectralUNet、Upsampler、WaveUNet、SpectralMaskNet）进行修复，并通过 LoRA 适配器训练了八个针对特定声源的专家模型。

表3。各分离音轨的多梅尔信噪比（单位：分贝）（↑）。

Team	人声	吉他	键盘	合成器	低音	鼓	打击乐	管弦乐
xlancelab	1.56	3.95	6.71	2.26	8.22	5.65	1.17	6.17
CUPAudio	1.36	1.95	2.76	0.98	5.29	4.92	0.16	1.32
ACDC	1.05	1.14	1.82	0.95	2.86	2.73	0.02	1.05
Hachimi	1.05	1.03	3.34	1.24	5.05	2.85	0.00	1.45
cp-jku	0.84	1.29	0.78	0.64	1.55	0.96	0.10	0.51
平均值	*1.17*	*1.87*	*3.08*	*1.21*	*4.59*	*3.42*	*0.29*	*2.10*

5. 讨论

多阶段处理有利于顶尖系统。 排名前两位的系统都采用了多阶段处理方法，而不是试图通过单一模型解决恢复问题。xlancelab 系统依次串联了分离、去混响和降噪模块，而 CUPAudioGroup 则集成了三个互补的分离模型。这种模块化设计能够利用预训练的 MSS 检查点 [8] ，并降低每个处理阶段需要处理的复杂度。然而，单靠多阶段处理并不能保证成功：cp-jku 同样采用了两阶段流程，但排名第五，这表明每个阶段中的架构选择仍然至关重要。

数据质量比数量更重要。 所有参赛团队都使用 RawStems 数据集 [5] 进行训练，但只有 xlancelab 投入精力对其进行人工清理，以解决原始数据中已知的对齐和源泄漏问题。尽管 Hachimi 将六个不同的数据集结合使用，而 xlancelab 只使用了两个数据集，但 xlancelab 的多梅尔信噪比高出 91%，这表明对于这项任务而言，数据质量比数据多样性更为重要。

简单的重建损失优于对抗训练。排名前两位的团队仅采用 L1 和短时傅里叶变换重建损失，未使用对抗训练。相比之下，采用基于生成对抗网络的训练的团队（AC DC、Hachimi 和 cp-jku）排名第三至第五。这表明，对抗训练可能无法为 MSR 带来明显优势，或者需要格外谨慎地调优，而这些系统未能做到这一点。

多音源和瞬态声源带来了不同的挑战。xlancelab 系统的性能优势主要体现在管弦乐元素（比第二名高 +4.72 分贝）和键盘乐器（高 +3.37 分贝）等多音源上，因为这些声源需要保留复杂的谐波关系。然而，在人声（高 +0.20 分贝）等单音乐源上，优势则显著缩小。与此同时，打击乐对所有系统而言仍然难以处理（平均为 0.29 分贝，而低音为 4.59 分贝），这是由于其具有脉冲式的宽带特性，给相位重建带来了根本性挑战。

6. 结论

MSR 挑战赛建立了首个标准化的音乐源恢复基准。结果表明，采用预训练 MSS 模型并结合简单重建损失的序列架构和集成架构能够取得最佳性能。在所有团队中取平均值后，低音（4.59 分贝）与打击乐（0.29 分贝）之间的 16× 性能差距表明，在实际部署中可能需要针对不同声源采用专门的方法。数据集和基线实现已公开发布于 https://msrchallenge.com/。

7. 致谢

本研究得到英国工程与自然科学研究理事会（EPSRC）资助，资助编号为 EP/T019751/1、EP/Y028805/1 和 UKRI397。为实现开放获取，作者已对由此产生的任何作者接受稿版本申请知识共享署名许可协议（CC BY）许可。

8. 参考文献

[1] Estefania Cano、Derry FitzGerald、Antoine Liutkus、Mark D Plumbley 和 Fabian-Robert St¨oter，《音乐源分离：简介》，*IEEE* 信号处理杂志，第36卷第1期，第31–40页，2018年。[2]Emmanuel Vincent、R´emi Gribonval 和 C´edric F´evotte，《盲音频源分离的性能测量》，*IEEE*音频、语音与语言处理汇刊，第14卷第4期，第1462–1469页，2006年。[3] Alexandre D´efossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi，《高保真神经音频压缩》，机器学习研究汇刊，2023年。[4] Rithesh Kumar、Prem Seetharaman、Alejandro Luebs、Ishaan Kumar 和 Kundan Kumar，《采用改进RVQGAN的高保真音频压缩》，神经信息处理系统进展，第36卷，第27980–27993页，2023年。[5]Yongyi Zang、Zheqi Dai、Mark D Plumbley 和 Qi-uqiang Kong，《音乐源恢复》，*arXiv*预印本 *arXiv:2505.21827*，2025年。[6] Yuki Mitsufuji、Giorgio Fabbro、Stefan Uhlich、Fabian-Robert St¨oter、Alexandre D´efossez、Minseok Kim、Woosung Choi、Chin-Yun Yu 和 Kin-Wai Cheuk，《2021年音乐解混挑战赛，808395，2022 1，》，信号处理前沿，第卷，第页。[7]Stefan Uhlich、Giorgio Fabbro、Masato Hirano、Shusuke Takahashi、Gordon Wichern、Jonathan Le Roux、Dipam Chakraborty、Sharada Mohanty、Kai Li、Yi Luo 等，《2023年声音解混挑战赛电影音频解混赛道》，*arXiv*预印本 *arXiv:2308.06981*，2023年。[8]Roman Solovyev、Alexander Stempkovskiy 和 Tatiana Habruseva，《声音解混任务的基准与排行榜》，*arXiv*预印本 *arXiv:2305.07489*，2023年。[9]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。[10] Igor Pereira、Felipe Ara´ujo、Filip Korzeniowski 和 Richard Vogl，《MoisesDB：一个包含4个以上分轨的源分离数据集》，载于第*24*届国际音乐信息检索学会会议（*ISMIR*），2023年。[11] Yongyi Zang、Jiarui Hai、Wanying Ge、Qiuqiang Kong、Zheqi Dai、Helin Wang、Yuki Mitsufuji 和 Mark D Plumbley，《MSRBench：音乐源恢复基准测试数据集》，*arXiv*预印本 *arXiv:2510.10995*，2025年。[12] Rub´en Tortosa、Jose M Jim´enez、Juan R Diaz 和 Jaime Lloret，《音频流的最优编解码器选择算法》，载于*IEEE Globecom* 研讨会，2014年，第237–242页。[13] Jyrki Alakuijala、Martin Bruse、Sami Boukortt、Jozef Marus Coldenhoff 和 Milos Cernak，《Zimtohrli：一种高效的心理声学音频相似度指标》，*arXiv*预印本 *arXiv:2509.26133*，2025年。[14] Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Tay-lor Berg-Kirkpatrick 和 Shlomo Dubnov，《采用特征融合和关键词到字幕增强的大规模对比式语言-音频预训练》，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2023年。[15] Daniel Stoller、Sebastian Ewert 和 Simon Dixon，《Wave-U-Net：一种用于端到端音频源分离的多尺度神经网络》，载于第*19*届国际音乐信息检索学会会议（*ISMIR*），2018年。[16] Andreas Jansson、Eric Humphrey、Nicola Montecchio、Rachel Bittner、Aparna Kumar 和 Till man Weyde，《采用深度U-Net卷积网络的歌声分离》，载于第*18*届国际音乐信息检索学会会议（*ISMIR*），2017年。[17] Wei-Tsung Lu、Ju-Chiang Wang、Qiuqiang Kong 和 Yun-Ning Hung，《采用频带分割RoPE变换器的音乐源分离》，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2024年，第481–485页。[18] Ju-Chiang Wang、Wei-Tsung Lu 和 Minz Won，《用于音乐源分离的 Mel 频带 RoFormer 》，*arXiv*预印本 *arXiv:2310.01809*，2023年。[19] Yi Luo 和 Jianwei Yu，《采用频带分割RNN的音乐源分离》，*IEEE/ACM*音频、语音与语言处理汇刊，第31卷，第1893–1901页，2023年。[20] Junyu Chen、Susmitha Vekkot 和 Pancham Shukla，《基于轻量级深度学习框架的音乐源分离（DTTNET：双路径TFC-TDF U-Net）》，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2024年，第656–660页。[21] Rachel M Bittner、Justin Salamon、Mike Tierney、Matthias Mauch、Chris Cannam 和 Juan Pablo Bello，《MedleyDB：一个用于注释密集型音乐信息检索研究的多轨数据集》，载于第*15*届国际音乐信息检索学会会议（*ISMIR*），2014年。[22] Bochen Li、Xinzhao Liu、Karthik Dinesh、Zhiyao Duan 和 Gaurav Sharma，《创建用于多模态音乐分析的多轨古典音乐演奏数据集：挑战、见解与应用》，*IEEE*多媒体汇刊，第21卷第2期，第522–535页，2018年。[23] Curtis Hawthorne、Andriy Stasyuk、Adam Roberts、Ian Simon、Cheng-Zhi Anna Huang、Sander Dieleman、Erich Elsen、Jesse Engel 和 Douglas Eck，《借助MAESTRO数据集实现因子化钢琴音乐建模与生成》，载于国际学习表征会议（*ICLR*），2019年。