THE INAUGURAL MSR CHALLENGE EVALUATION PROTOCOL

The MSR Challenge Organizers

Last updated on November 14^{st}, 2025

1. INTRODUCTION

Traditional Music Source Separation (MSS) systems operate under the assumption that musical mixtures are linear combinations of unprocessed instrument stems [1]. However, this model inadequately represents real-world audio production pipelines where recordings undergo extensive signal processing including equalization, dynamic range compression, reverberation, harmonic distortion, mastering, lossy transmission, and storage degradation [2]. While MSS systems can separate individual sources, their subtractive nature prohibits recovery of original signals before these transformations are applied.

The Music Source Restoration (MSR) Challenge addresses this limitation by requiring systems to recover original, unprocessed source signals from fully mixed and mastered audio. This necessitates a fundamental shift from subtractive separation methods to generative restoration approaches. The challenge connects academic research with practical audio engineering needs, including archival preservation, professional remixing, recovery of historical recordings, and enhancement of live performances affected by venue acoustics.

This paper presents the comprehensive evaluation protocol for the inaugural MSR Challenge, detailing our evaluation methodology, datasets, metrics, and ranking procedures. We establish two distinct evaluation paths: an objective path with two tracks optimized for signal reconstruction and semantic alignment, and a subjective path evaluating perceptual quality through professional assessment of real-world degradation scenarios.

2. CHALLENGE OVERVIEW

The MSR Challenge evaluates systems on their ability to restore eight target instrument stems: vocals, guitars, keyboards, bass, synthesizers, drums, percussion, and orchestral elements. The commonly found "others" stem in MSS is intentionally excluded due to its inherent variability and lack of clear definition.

2.1. Timeline

- Registration Opens: August 15th, 2025
- Validation Set Release: September 1st, 2025
- Baseline System Release: September 15th, 2025
- Test Set Release: November 25th, 2025
- Final Submission Deadline: November 27th, 2025
- Data Submission Deadline: December 3th, 2025
- Results Announcement: December 4th, 2025
- 2-page Paper Due: December 7th, 2025

2.2. Data Policy

Participants may use any open-source academic datasets for training, including multitrack music datasets (MUSDB18-HQ [3], MoisesDB [4], MedleyDB [5], RawStems [2]), single-instrument recordings (URMP [6], MAESTRO [7]), and noise datasets (WHAM! [8], Freesound [9]). Participants may also create synthetic training data. All newly created datasets and generation pipelines must be submitted and shared by November 26, 2025. This deadline, occurring after the final submission deadline, allows participants to maintain data advantages during the competition while ensuring eventual sharing for the benefit of the research community.

Important Note on RawStems: While RawStems is available as a training resource, participants should be aware that it contains significant data quality issues, including incomplete time alignment and instrument leakage. RawStems is best used as a starting point, and participants are strongly encouraged to source additional clean stems for each instrument category to achieve competitive performance.

3. DATASET

3.1. Validation Set: MSRBench

The validation set consists of MSRBench, a professionally curated dataset specifically developed for benchmarking MSR systems. MSRBench contains 250 10-second audio clips with corresponding individual stem pairs for each of the eight target instruments with 12 additional mixture degradation types. All audio is provided in stereo at 48 kHz sampling rate. The dataset was created using professional mixing techniques, and is publicly available at `https://huggingface.co/datasets/yongyizang/MSRBench` and may be used entirely or partially for system development. Participants are encouraged to report findings on MSRBench in academic publications regardless of challenge participation.

3.2. Non-Blind Test Set

The non-blind test set contains 1000 10-second stereo clips at 48 kHz, extracted from professionally mixed and mastered commercial songs with previously unreleased stems. The distribution of this test set is similar to that of the validation set (MSRBench), ensuring consistency in evaluation conditions. Ground-truth unprocessed stems are available, allowing calculation of intrusive objective metrics. These clips are custom mixed and undergo a range of simulated audio degradations including:

- Equalization and frequency shaping
- Dynamic range compression and limiting
- Spatial processing (reverb, delay, stereo widening)
- Harmonic distortion and saturation

- Mastering chain effects
- Lossy encoding artifacts

3.3. Blind Test Set

The blind test set contains 500 10-second stereo clips at 48 kHz representing four real-world degradation scenarios where ground-truth signals are unavailable:

1. **Historical Recordings (125 clips)**: Digitized cylinder recordings from the UCSD Cylinder Audio Archive, representing storage degradation from early recording media.
2. **Live Recordings (125 clips)**: Concert performances sourced from YouTube, affected by venue acoustics, crowd noise, and environmental degradation.
3. **FM Radio Broadcast (125 clips)**: Songs recorded through FM radio transmission, exhibiting analog transmission artifacts and degradation.
4. **Lossy Streaming (125 clips)**: Music transmitted under low bitrates and lossy codecs, representing digital transmission degradation common in streaming services.

To ensure fair evaluation and prevent overfitting, participants receive the complete test set 48 hours before the submission deadline for final inference and submission.

4. EVALUATION METRICS

4.1. Objective Metrics

4.1.1. Multi-Mel Spectrogram Signal-to-Noise Ratio (Multi-Mel-SNR)

For the Signal Reconstruction track, we employ Multi-Mel-SNR, which measures spectro-temporal reconstruction accuracy while avoiding the phase oversensitivity inherent in complex spectrogram or waveform metrics. This metric evaluates magnitude-only reconstruction across multiple time-frequency resolutions.

Scale-invariant normalization. To isolate reconstruction quality from loudness differences, we first apply scale-invariant normalization to the waveform domain. For each segment pair consisting of reference stem $\mathbf{s}$ and predicted stem $\hat{\mathbf{s}}$, we compute the optimal scaling factor:

$$\alpha^* = \frac{\langle \mathbf{s}, \hat{\mathbf{s}} \rangle}{\langle \hat{\mathbf{s}}, \hat{\mathbf{s}} \rangle} \tag{1}$$

and obtain the scaled prediction $\tilde{\mathbf{s}} = \alpha^* \hat{\mathbf{s}}$. This normalization follows the scale-invariant principle used in SI-SNR [10], ensuring that global amplitude differences do not affect the metric.

Multi-resolution mel spectrogram analysis. From the reference waveform $\mathbf{s}$ and scaled prediction $\tilde{\mathbf{s}}$, we compute power mel spectrograms (squared magnitude) using three configurations with different time-frequency resolutions:

- Configuration A: 512-sample window, 256-sample hop, 80 mel bins
- Configuration B: 1024-sample window, 512-sample hop, 128 mel bins
- Configuration C: 2048-sample window, 1024-sample hop, 192 mel bins

All configurations use $f_{\text{min}} = 0$ Hz, $f_{\text{max}} = 24$ kHz, and are computed via `torchaudio.transforms.MelSpectrogram` with default settings.

SNR computation. For each configuration $i \in \{A, B, C\}$, let $\mathbf{M}_i$ and $\tilde{\mathbf{M}}_i$ denote the reference and scaled prediction mel spectrograms. We compute:

$$\text{SNR}_i = 10 \log_{10} \frac{\sum_{t,f} \mathbf{M}_i(t,f)^2}{\sum_{t,f} (\mathbf{M}_i(t,f) - \tilde{\mathbf{M}}_i(t,f))^2} \tag{2}$$

where the summation is over all time frames t and mel frequency bins f. The Multi-Mel-SNR for a single segment is:

$$\text{Multi-Mel-SNR} = \frac{1}{3} \sum_{i \in \{A,B,C\}} \text{SNR}_i \tag{3}$$

The final track score is the arithmetic mean of Multi-Mel-SNR across all eight target stems and all test clips. Higher values indicate better reconstruction quality.

4.1.2. Zimtohril

For the Generation Quality track, we employ Zimtohrli [11], a recently developed full-reference audio similarity metric is grounded in psychoacoustic principles. Zimtohril combines a 128-bin gammatone filterbank that models cochlear frequency resolution with a non-linear resonator model to imitate the human eardrum's response. The metric computes similarity by comparing perceptually-mapped spectrograms using modified Dynamic Time Warping (DTW) and Neurogram Similarity Index Measure (NSIM) algorithms enhanced with non-linearities to better align with human perception. The Zimtohril implementation can be found at `https://github.com/google/zimtohrli`.

For each separated stem, we compute Zimtohrli scores for all test segments and report the mean score. Higher Zimtohrli scores indicate better generation quality. The final track ranking is determined by the average Zimtohrli score across all eight target stems and all test clips.

4.1.3. Fréchet Audio Distance with CLAP (FAD-CLAP)

For the Semantic Alignment track, we employ FAD-CLAP, which computes Fréchet Audio Distance over CLAP (Contrastive Language-Audio Pretraining) embeddings [12]. FAD-CLAP is calculated as:

$$\text{FAD-CLAP} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{4}$$

where μ_r, Σ_r and μ_g, Σ_g are the mean and covariance of CLAP embeddings for reference and generated audio, respectively. Lower FAD-CLAP values indicate better semantic similarity and structural preservation.

All objective metrics are calculated on 10-second non-overlapping windows, with final rankings based on mean scores across all windows.

4.2. Subjective Metrics

Professional mixing engineers and music producers assess restoration quality through blind and non-blind listening tests. Each 10-second segment receives ratings on a 5-point scale (1 = very poor, 5 = excellent) across three dimensions:

- **MOS-Separation**: Evaluates whether the output contains only the complete target instrument with no interference from other sources.
- **MOS-Restoration**: Assesses how well the target instrument has been restored to its original, undegraded state, even if other elements are present or the target is incomplete.
- **MOS-Overall**: Reflects overall perceptual quality and similarity to ground truth, determining final system rankings in the Perceptual Quality track.

MOS-Separation and MOS-Restoration provide analytical insights, while MOS-Overall determines final rankings based on professional audio standards. Individual rater scores will be released for future research.

Development Recommendation: Participants are encouraged to self-test their systems using audiobox-aesthetics as a partial surrogate for human raters during development. While not a perfect substitute for professional evaluation, this can provide useful guidance on perceptual quality before final submission.

5. ZIMTOHRIL IMPLEMENTATION

While Zimtohril provides a Python wrapper, its documentation requires clarification. The implementation can be achieved through the following procedure. First, the repository is cloned from `https://github.com/google/zimtohrli`, and the package is installed via `pip install .` from within the cloned directory.

A minimal implementation example is provided in Listing 1. The metric requires both reference and degraded audio signals as single-precision floating-point arrays. Audio files are loaded using the `soundfile` library, and the Zimtohril distance is computed through the `Pyohrli` class interface, which returns a perceptual distance measure between the two signals. The following example assumes both audio clips are mono, 48 kHz and equal in length.

Listing 1. Minimal Zimtohril usage example.

```
import numpy as np
import soundfile as sf
import pyohrli

ref_path = "ground-truth.wav"
deg_path = "estimated.wav"

ref, sr_ref = sf.read(ref_path)
deg, sr_deg = sf.read(deg_path)

ref = np.asarray(ref, dtype=np.float32)
deg = np.asarray(deg, dtype=np.float32)

metric = pyohrli.Pyohrli()
zimt_distance = metric.distance(ref, deg)
```

6. EVALUATION TRACKS AND RANKING

The challenge features two distinct evaluation paths: **Objective Evaluation** (non-blind test set) and **Subjective Evaluation** (blind test set). Within the objective path, we recognize two different optimization strategies, resulting in three total leaderboards.

6.1. Objective Evaluation Path

The objective path uses the non-blind test set where ground-truth signals are available, allowing computation of intrusive metrics. This path includes two tracks optimizing for different aspects of restoration:

6.1.1. Track 1: Signal Reconstruction

This track emphasizes accurate recovery of original stem waveforms with focus on spectro-temporal fidelity. To address the oversensitivity of phase information in complex spectrogram metrics, we employ a Multi-Mel Spectrogram Signal-to-Noise Ratio (Multi-Mel-SNR) metric that evaluates reconstruction quality across multiple time-frequency resolutions. Systems are ranked exclusively by this metric in Track 1.

6.1.2. Track 2: Generation Quality

This track evaluates the perceptual quality and naturalness of separated stems using Zimtohril [11], a learning-based metric designed to assess audio generation quality. This track rewards systems based on perceptual quality, complementing the signal-level reconstruction metric in Track 1.

6.1.3. Track 3: Semantic Alignment

This track prioritizes preservation of musical content and semantic coherence using the Fréchet Audio Distance with CLAP embeddings (FAD-CLAP). FAD-CLAP measures the distributional similarity between separated stems and reference stems in the CLAP semantic embedding space, capturing high-level musical characteristics such as timbre, instrumentation, and musical context.

Systems are ranked exclusively by FAD-CLAP performance, averaged across all eight target stems and all test clips. Lower FAD-CLAP scores indicate better semantic alignment. This track recognizes systems that maintain musical meaning and structural coherence, even when waveform-level precision differs from the reference.

6.1.4. Overall Ranking

To provide a comprehensive assessment, we compute an overall ranking by taking the macro-average of each system's rankings across all three tracks. Specifically, if a system ranks r_1, r_2, and r_3 in Tracks 1, 2, and 3 respectively, its overall ranking score is $(r_1 + r_2 + r_3)/3$. Systems are then ordered by this overall ranking score, with lower scores indicating better overall performance. This approach ensures balanced consideration of signal reconstruction fidelity, generation quality, and semantic alignment.

6.2. Subjective Evaluation Path

This track is based exclusively on subjective evaluations using the blind test set. Professional audio engineers rate restoration quality across all four real-world degradation scenarios (historical recordings, live performances, FM radio broadcast, and lossy streaming).

The final ranking for Track 3 is computed by equally weighting all three Mean Opinion Scores:

- MOS-Separation (33.3% weight)
- MOS-Restoration (33.3% weight)
- MOS-Overall (33.3% weight)

Scores are averaged across all eight target stems, all four degradation scenarios, and all test clips. This track reflects real-world professional standards and end-user perceptual quality in challenging restoration scenarios.

6.3. Ranking Procedure

For each track, scores are computed separately for each of the eight target stems. All scores are normalized to a 0–1 scale across all stem-metric combinations to ensure fair comparison. Final track rankings are determined by averaging the normalized scores across all stems.

This produces three distinct leaderboards: Track 1 (Signal Reconstruction), Track 2 (Semantic Alignment), and Track 3 (Perceptual Quality). Each leaderboard recognizes different strengths in restoration systems. The top 5 participants in the combined rankings of Tracks 1 and 2 (the objective evaluation path) will be invited to submit a 2-page short paper detailing their methodology.

7. SUBMISSION GUIDELINES

7.1. Platform

Due to timeout issues with handling large audio objects, the testing phase of the challenge will be based on Google Drive. Candidates will receive a zip file for all mixtures for separation, and an example submission zip file (as detailed in submission format section); candidates need to replace the dummy results in example zip file, and upload the zip file to a publically accessible Google Drive link. Candidates can submit the google drive link for their submission through a Google Form that'll be available during the testing phase; the final submission will be used to rate their system. Registration and development phases are managed separately.

7.2. Submission Format

Each submission must be a zip file containing folders named by song ID. Each folder must contain restored stems as separate FLAC files named `{stem_name}.flac`, where stem names correspond to the eight target instruments (vocals, guitars, keyboards, bass, synthesizers, drums, percussion, orchestral). All output files must be stereo at 48 kHz sampling rate with exactly 10 seconds duration.

An example submission structure will be provided alongside the test set release to ensure proper formatting. Submissions that do not conform to the required format will be rejected automatically.

7.3. Submission Limits

Participants can change the google drive link used for final submission during the final testing period at any time before the end of the testing phase. Each submission must include results for both the blind and non-blind test sets.

7.4. Reproducibility Requirements

All participants must open-source their training and evaluation code to ensure reproducibility and advance the field. Final submissions must include a `README.md` file containing:

- Brief system description and methodology overview
- Link to publicly accessible code repository
- Instructions for reproducing results

Participants who fail to provide accessible code by the submission deadline will be disqualified. This requirement ensures the challenge contributes lasting value to the research community.

8. BASELINE SYSTEMS

We provide complete training and evaluation code for two baseline systems alongside the validation set release:

1. **U-Net** [13]: A multi-scale neural network operating on complex spectrograms, originally designed for end-to-end audio source separation. This baseline demonstrates a straightforward approach to the restoration task.
2. **BSRNN** [14]: Band-Split RNN architecture for high-fidelity enhancement, also operating in the complex spectrogram domain. This baseline represents a more sophisticated frequency-domain approach.

Both baselines are trained on the RawStems dataset using the provided synthetic degradation pipeline. Participants may use these baselines as starting points or comparison references for their own system development. We provide example code for training, inference and evaluation at `https://github.com/yongyizang/MSRKit`, and pre-trained checkpoints at `https://huggingface.co/yongyizang/MSRChallengeBaseline`.

9. REFERENCES

[1] M. Mirbeygi et al., "Speech and music separation approaches—a survey," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 21155–21197, 2022.

[2] Y. Zang et al., "Music source restoration," *arXiv preprint arXiv:2505.21827*, 2025.

[3] Z. Rafii et al., "MUSDB18-HQ—an uncompressed version of MUSDB18," 2019.

[4] I. Pereira et al., "MoisesDB: A dataset for source separation beyond 4-stems," in *Proc. ISMIR*, 2023.

[5] R. M. Bittner et al., "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. ISMIR*, 2014, vol. 14.

[6] B. Li et al., "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.

[7] C. Hawthorne et al., "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. ICLR*, 2019.

[8] G. Wichern et al., "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019.

[9] E. Fonseca et al., "FreeSound datasets: A platform for the creation of open audio datasets," in *Proc. ISMIR*, 2017.

[10] J. Le Roux et al., "SDR—half-baked or well done?," in *Proc. ICASSP*, 2019.

[11] Jyrki Alakuijala, Martin Bruse, Sami Boukortt, Jozef Marus Coldenhoff, and Milos Cernak, "Zimtohrli: An efficient psychoacoustic audio similarity metric," *arXiv preprint arXiv:2509.26133*, 2025.

[12] B. Elizalde et al., ," .

[13] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. ISMIR*, 2018.

[14] J. Yu et al., "High fidelity speech enhancement with band-split RNN," in *Proc. Interspeech*, 2023.