Cinematic Source Separation with Dialogue-Driven Sidechain Ducking

Atoof Shakir [1] **Florian Grötschla** [1] **Luca A. Lanzendörfer** [1] **Roger Wattenhofer** [1]

Abstract

Cinematic audio source separation, the task of extracting dialogue, music, and effects stems from film soundtracks, is a prerequisite for speech-centric post-production workflows including dubbing, accessibility captioning, and broadcast ASR, yet is limited by the lack of realistic training and evaluation data. Existing synthetic datasets sum stems linearly without production artifacts such as sidechain ducking, where music and effects are attenuated in the presence of speech to protect dialogue intelligibility. Real mixes apply this ducking on the mix bus, so the original stems no longer sum to the output mix. We present CineAudioGen, an agentic pipeline that generates four-stem cinematic training data using a Director-Critic architecture. We also release CineAudioDB, a dataset of real cinematic audio with ground-truth stems from independent filmmakers. Fine-tuning Bandit v2, MRX, and HTDemucs on our data yields considerable improvements on cinematic content compared to baselines.

1. Introduction

Source separation of cinematic audio into stems is a prerequisite for speech-centric post-production tasks including multilingual dubbing, broadcast ASR, and accessibility services, as well as creative workflows such as remixing and localization. While music source separation has benefited from datasets such as MUSDB18 (Rafii et al., 2017) and dedicated challenges (Uhlich et al., 2024), the cinematic domain remains underserved (Pons et al., 2024). Real production stems are proprietary, and synthetic alternatives such as Divide and Remaster (DnR) (Petermann et al., 2021), which combines LibriSpeech (Panayotov et al., 2015), FMA (Defferrard et al., 2017), and FSD50K (Fonseca et al., 2022), construct mixtures by summing sources at fixed gains. DnR v3 (Watcharasupat et al., 2024) improves this with realistic per-stem loudness distributions and EBU R 128 (European Broadcasting Union, 2023) peak limiting, and DnR-nonverbal (Hasumi & Fujita, 2025) adds vocal diversity. However, both still omit production effects such as compression, reverb, and sidechain ducking. Sidechain ducking attenuates music and effects during dialogue to preserve intelligibility, shaping the spectral and temporal relationship between dialogue and competing sources. Professional mixing further complicates separation because sidechain compression, equalization, and mastering introduce non-linear interactions between stems, so the original stems no longer sum to the output mix. This violates a core assumption of many current separation architectures. Mask-based models such as Bandit v2 (Watcharasupat et al., 2024) estimate stems by redistributing mixture energy across sources, making it difficult to reconstruct stems that exceed the mixture in energy. Other architectures such as HTDemucs (Rouard et al., 2023) are less directly constrained by this formulation, but also degrade on non-additive cinematic mixtures. To close the gap between current research and real productions, we present CineAudioGen, an agentic pipeline for generating cinematic source separation training data (Wang et al., 2025; 2024). A *Director* agent analyzes dialogue recordings via a multi-modal LLM and plans scene composition, a DSP *Engine* renders four stems with sidechain ducking and spatial processing, and a *Critic* agent evaluates the mix and feeds corrections back to the Engine. Since audio LLMs often rely on lexical content over acoustic features (Chen et al., 2025; Zang et al., 2025), the agents receive structured metadata and loudness metrics alongside audio. To bridge previous research with real productions, we propose *linear mixing*: ducking is applied to individual stems *before* summation, so training targets sum exactly to the mixture while retaining realistic gain modulation. Dialogue stems draw from datasets spanning acted emotions (Livingstone & Russo, 2018), conversational styles (Nguyen et al., 2023), nonverbal vocalizations (Borisov et al., 2025), and underrepresented languages such as Amharic (Retta et al., 2023). We fine-tune Bandit v2 (Watcharasupat et al., 2024), MRX (Petermann et al., 2022), and HTDemucs (Rouard et al., 2023) on our generated data and evaluate on synthetic and real cinematic test sets.

[1]ETH Zurich, Switzerland. Correspondence to: Atoof Shakir <ashakir@ethz.ch>.

Machine Learning for Audio Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

2. Methodology

Generating realistic cinematic training data requires contextually plausible stem selection and a clear definition of the separation target. CineAudioGen[1] addresses both by planning scene-dependent dialogue, music, ambience, and SFX, and by treating each post-ducking stem contribution as the target source. This preserves additivity for current separation models while also allowing release-rendered mixes for non-additive evaluation.

2.1. Pipeline Overview

We propose CineAudioGen, a pipeline split into different components. The Director's scene plan specifies a music mood tag, an ambience description, and timestamped SFX events. Music is retrieved via tag-based lookup from approximately 2k FMA (Defferrard et al., 2017) tracks tagged for mood, instrumentation, and tempo using Music Flamingo (Ghosh et al., 2025). SFX and ambience assets are retrieved via CLAP (Wu et al., 2023) embedding similarity over FSD50K (Fonseca et al., 2022). The Engine renders each stem through a signal processing chain (high-pass filter, compression, reverb, gain) at 48 kHz. We define 14 reverb presets spanning indoor, venue, industrial, and outdoor environments, and six scene archetypes with distinct mixing profiles to ensure diversity in the data. Before the Critic is invoked, the Engine builds a *rough* mix from hard-coded initial settings (speech compression at -15 dB threshold with ratio 3:1, music and SFX at -9 dB gain, ambience at -6 dB). Sidechain ducking is applied on the mix bus and the result is mastered (-3 dB headroom, compression, limiting, loudness normalization to -27 dB LUFS). This rough mix is provided to the Critic alongside per-stem loudness metrics. The Critic then outputs delta gain and reverb corrections relative to these initial settings. Each scene is rendered in two modes, namely in *linear*, where Critic-adjusted ducking is applied to individual stems before summation (Section 2.2), serving as the training target and *release*, where the mastering chain is applied to the Critic-adjusted summed signal.

2.2. Linear Mixing

Sidechain ducking is central to cinematic mixing: music and effects are attenuated during dialogue to preserve speech intelligibility. In professional workflows, ducking is combined with mastering processes such as compression, limiting, and loudness normalization, so the original stems no longer sum to the output mix. Since Bandit v2's mask-based formulation ($\hat{\mathbf{s}}_k = \mathbf{x} \odot \mathbf{m}_k$) failed to train on such non-additive mixtures, we adopt a linear mixing strategy that applies ducking before summation:

$$\begin{aligned} \text{music}_{\text{ducked}} &= \text{duck}(\text{music}, \text{speech}) \\ \text{sfx}_{\text{ducked}} &= \text{duck}(\text{sfx}, \text{speech}) \\ \text{mix} &= \text{speech} + \text{music}_{\text{ducked}} + \text{sfx}_{\text{ducked}}. \end{aligned}$$

This guarantees $\text{mix} = \sum_k \text{stem}_k$ while retaining dialogue-dependent gain modulation. We implement envelope-following sidechain compression with slow attack (15–30 ms) and release (200–300 ms); music receives stronger ducking (-12 dB reduction, -20 dB threshold) than SFX (-8 dB reduction, -25 dB threshold). After rendering, we verify that $\max(|\text{mix} - \sum_k \text{stem}_k|) < 10^{-4}$. Linear and release renderings let us test whether this additive approximation transfers to real production conditions.

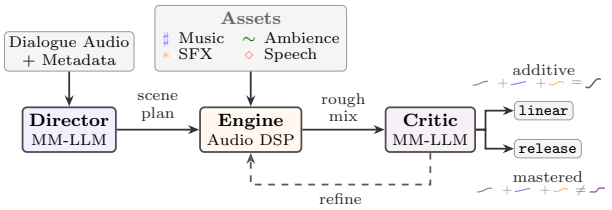

Figure 1. CineAudioGen pipeline. The Director plans scenes from dialogue, the Engine applies DSP with sidechain ducking, and the Critic refines settings. `linear` preserves mix$=\sum$(stems) while `release` applies mastering, breaking additivity. 'MM-LLM' stands for Multimodal-LLM

2.3. CineAudioDB

We assemble CineAudioDB,[2] an evaluation set containing nine real productions from two sources totalling 2.5h of audio content. First, two film universities in Switzerland provided productions from their degree programs, where students collaborate with professional recording studios for mixing and mastering. These range from short narrative films and comedies to documentaries and animation, with multilingual dialogue (English, French, German) and diverse vocal content including non-verbal sounds. Second, we include professionally mixed open-source Blender films. All productions were mixed in varying surround audio formats. For consistency, we obtained the stereo downmix from each source. Because professional mixes include compression, equalization, spatial effects, and cross-stem processing, delivered stems do not necessarily sum to the mastered stereo mix. We therefore release two CineAudioDB versions: **production-mix**, using the original mastered mix as input, and **linear-mix**, obtained by directly summing the reference stems. The former reflects realistic deployment, while the latter isolates separation performance from mixing-process mismatch.

[1] https://github.com/ETH-DISCO/cineaudiogen

[2] https://huggingface.co/datasets/disco-eth/cineaudiodb

3. Experimental Evaluation

CineAudioSynth Data Generation. As LLM, we chose Google Gemini 3-Flash since it natively supports audio input, text input and reasoning, while being cost-effective for large-scale generation and offering an API. (Team et al., 2025) We generated 326 scenes (22 hours)[3] split into 280 training, 33 validation, and 13 test samples. Dialogue stems draw from four datasets following (Hasumi & Fujita, 2025): Expresso (Nguyen et al., 2023) (long-form conversational speech, 7 styles), RAVDESS (Livingstone & Russo, 2018) (acted emotional clips assembled into pseudo-dialogues), ASED (Retta et al., 2023) (Amharic emotional speech), and NonverbalTTS (Borisov et al., 2025) (laughs, coughs, sighs). The distribution is 30% Expresso, 30% RAVDESS, 30% ASED, 10% NonverbalTTS. Music comprises 2k FMA tracks (20+ hours) tagged via Music Flamingo (Ghosh et al., 2025); SFX and ambience are drawn from FSD50K via CLAP retrieval. Each scene is rendered in *linear* (reconstruction SI-SNR ≈ 74 dB) and *release* (reconstruction SNR ≈ -6 dB) modes. We merge ambience into SFX for all experiments to match the three-stem convention used by DnR. CAS refers to the linear mix if not specified otherwise.

Models and Training. We evaluate Bandit v2 (Watcharasupat et al., 2024), MRX (Petermann et al., 2022), and HTDemucs (Rouard et al., 2023). Bandit v2 estimates complex-valued time-frequency masks, while MRX estimates magnitude masks at multiple STFT resolutions. Both are fine-tuned from their DnR v3 pretrained checkpoints for 30 epochs, either on *CineAudioSynth (CAS)* alone or on *CAS + DnR*, where cinematic and DnR v3 samples are mixed at a 2:1 duration ratio as data replay (Chaudhry et al., 2019). We denote pretrained checkpoints by PT and fine-tuned checkpoints by FT. HTDemucs predicts sources directly rather than estimating masks. Since its available pretrained checkpoint targets music stems, we train it from scratch for 30 epochs. For release-rendered data, where additivity does not hold, we add an auxiliary residual output to absorb mastering artifacts and discard it at inference.

Evaluation Protocol. We report mean track-level SI-SNR (Le Roux et al., 2019). For non-additive rendering modes (release), we use mix-consistent references that redistribute the ground-truth residual equally across stems, preventing models from being penalized for energy introduced by the mastering chain. For real cinematic content, where objective metrics are unreliable due to non-linear mixing, we additionally run a listening test. We evaluate on (i) the DnR v3 validation set, (ii) the CineAudioSynth testset across both rendering modes, and (iii) CineAudioDB.

Separation Quality under Domain Shift. Three findings emerge from Table 1. First, the domain gap between general-purpose and cinematic separation is severe and asymmetric across stems. The pretrained model achieves reasonable dialogue quality but produces negative SI-SNR on music and near-zero on effects, indicating that speech characteristics transfer across domains while music and effects do not. Second, replaying the pretraining data during fine-tuning (CAS + DnR) consistently outperforms cinematic-only training across all stems, suggesting that data replay acts as an effective regularizer against catastrophic forgetting of useful pretrained representations. Third, comparing Bandit v2 and HTDemucs under cinematic-only training reveals comparable performance despite their architectural differences, indicating that the domain of the training data, rather than the model family, is the primary bottleneck. Across all configurations, music remains the most challenging stem, reflecting the spectral diversity of cinematic scores and their substantial overlap with both dialogue and effects. These gains do not transfer freely. Table 2 evaluates all six variants on the DnR validation set. Cinematic-only fine-tuning causes catastrophic forgetting for both architectures, with Bandit v2 losing more in absolute terms on MX and FX. Mixed training fully eliminates this forgetting and even surpasses pretrained performance for both models, confirming that replaying DnR data during cinematic fine-tuning is useful. Bandit v2 outperforms MRX across all training strategies.

Table 1. Separation quality on the testset of CineAudioSynth (CAS) with *linear* rendering where mix $= \sum$(stems). DX = dialogue, MX = music, SFX = sound effects. Values are mean SI-SNR in dB. Higher is better.

Model	Training Data	DX	MX	SFX
Bandit v2-FT	CAS + DnR	**25.69**	**−13.39**	**4.37**
Bandit v2-FT	CAS	23.56	−13.04	−0.37
HTDemucs-FT	CAS + DnR	23.35	−4.54	0.55
MRX-FT	CAS + DnR	20.78	−20.23	−1.29
MRX-FT	CAS	17.80	−28.20	−4.22
Bandit v2-PT	DnR v3	17.35	−23.46	−5.97
MRX-PT	DnR v3	15.04	−28.51	−6.91

Table 2. Mean SI-SNR in dB on cross-domain evaluation on the DnR v3 validation set ($n = 600$). Higher is better. We find that cinematic-only fine-tuning causes forgetting, whereas mixed training recovers and exceeds pretrained performance.

Model	Training	DX	MX	SFX
Bandit v2-PT	DnR v3	14.42	9.38	8.51
Bandit v2-FT	CAS	11.30	0.70	1.83
Bandit v2-FT	CAS + DnR	**15.77**	**10.36**	**9.91**
MRX-PT	DnR v3	9.03	2.56	3.41
MRX-FT	CAS	7.06	−0.89	0.77
MRX-FT	CAS + DnR	11.73	5.97	6.37
HTDemucs	CAS + DnR	4.51	−24.58	−2.54
HTDemucs	Release	4.75	−15.93	−4.04

[3] https://huggingface.co/datasets/disco-eth/cineaudiosynth

Table 3. Effect of mastering on separation quality evaluated on the CineAudioSynth testset. 'CAS' refers to CineAudioSynth. Gains from fine-tuning vanish as additivity decreases. Mean SI-SNR in dB, higher is better.

Architecture	Training	Inference	DX	MX	SFX
Bandit v2-PT	DnR v3	Linear	17.4	−23.5	−6.0
Bandit v2-PT	DnR v3	Release	14.1	−24.2	−14.6
Bandit v2-FT	CAS (linear)	Linear	23.6	−13.0	−0.4
Bandit v2-FT	CAS (linear)	Release	19.1	−27.4	−9.1
Bandit v2-FT	CAS + DnR	Linear	**25.7**	−13.4	**4.4**
Bandit v2-FT	CAS + DnR	Release	20.1	−20.5	−9.9
HTDemucs	CAS (linear)	Linear	23.3	**−4.5**	0.6
HTDemucs	CAS (linear)	Release	22.0	−7.1	−1.1
HTDemucs	CAS (release)	Linear	18.6	−24.3	−7.1
HTDemucs	CAS (release)	Release	15.9	−27.1	−15.2

Table 4. Mean SI-SNR in dB on CineAudioDB (mean across samples). Higher is better. *Production*: original mastered mix; *Linear*: ground-truth stems summed directly. Mixed training achieves the best overall Bandit v2 results.

		Production Mix			Linear Mix		
Model	Training	DX	MX	SFX	DX	MX	SFX
Bandit v2-PT	DnR v3	−1.95	4.04	−8.29	5.49	**5.07**	1.31
Bandit v2-FT	CAS	−0.43	0.02	−10.91	4.88	2.48	−0.84
Bandit v2-FT	CAS + DnR	−1.69	2.97	−8.07	**5.98**	**5.07**	**2.60**
MRX-PT	DnR v3	−2.61	2.27	−8.67	−18.33	3.55	1.83
MRX-FT	CAS	−1.57	−0.69	−11.34	0.96	0.49	−0.82
MRX-FT	CAS + DnR	−1.35	2.76	−7.17	4.53	3.75	2.54
HTDemucs	CAS + DnR	−3.31	−4.59	−12.44	−1.80	−4.89	−7.66

Cross-Mode Degradation. To isolate the effect of mastering on separation quality, we evaluate all model configurations on the CineAudioSynth test split rendered in both modes with decreasing reconstruction additivity, from *linear* (SI-SNR 67 dB) to *release* (SI-SNR 29.8 dB) (see Table 3). Fine-tuning on CineAudioSynth yields large gains on linear data across all stems. On release data, Bandit v2 retains a clear dialogue advantage over the pretrained baseline but loses most of its music and effects gains. Mixed training (CAS + DnR) outperforms cinematic-only training in both modes. HTDemucs trained on linear data generalizes best to release, retaining reasonable separation across all stems, while training directly on release-rendered data degrades performance in both modes.

Evaluation on CineAudioDB. We evaluate Bandit v2 and MRX (three variants each: pretrained, cinematic-only, mixed) on the real cinematic productions with ground-truth stems from independent filmmakers, using both the original production mix and a linear-mix version where ground-truth stems are summed directly (see Table 4). We also evaluate our fine-tuned HTDemucs, which cannot be compared to a pretrained checkpoint since HTDemucs was originally trained for music stem separation. For Bandit v2, removing mastering (linear mix) generally improves separation, with mixed training achieving the strongest results across all stems. This confirms on real data what the cross-mode experiment (see Table 3) demonstrated on synthetic data. We find that MRX clearly performs below Bandit v2 in both conditions, however, finetuning on CineAudioSynth still substantially increases MRX's performance.

Table 5. MUSHRA-like human listening test on CineAudioDB. Fine-tuned models outperform pretrained baselines across all stems. Mean ± 95% confidence interval, higher is better. Models are finetuned on CAS + DnR (2:1 ratio).

Model	DX	MX	SFX
Reference	91.4 ± 2.7	93.0 ± 2.1	89.8 ± 3.6
Bandit v2-FT	$\mathbf{68.8 \pm 5.7}$	$\mathbf{56.3 \pm 5.4}$	$\mathbf{47.5 \pm 6.3}$
Bandit v2-PT (DnR v3)	57.3 ± 6.6	43.0 ± 6.4	23.4 ± 4.7
HTDemucs-FT	50.5 ± 5.7	27.1 ± 4.4	18.3 ± 5.7
MRX-FT	40.1 ± 5.2	39.5 ± 6.0	23.0 ± 5.6
MRX-PT (DnR v1)	15.9 ± 4.7	35.4 ± 6.1	18.6 ± 5.7

Perceptual Evaluation. Separation of non-additive mixes is ill-posed since the mastering chain is non-invertible, so multiple valid separations exist and what matters is whether the output is perceptually useful. We therefore run a hidden-reference listening test following the MUSHRA protocol (itu, 2015) with 20 participants recruited online,[4] rating Bandit v2-PT, Bandit v2-FT, HTDemucs, MRX-PT, and MRX-FT on a 0–100 scale with no anchor conditions. Table 5 shows the results. For both Bandit v2 and MRX, models fine-tuned on CineAudioSynth score higher than their respective pretrained baselines, confirming that the generated training data and fine-tuning improves perceived dialogue separation quality on real mastered content.

4. Conclusion

Existing cinematic separation datasets ignore the production mechanisms that govern how dialogue competes with music and effects. We present CineAudioGen, an agentic pipeline that generates cinematic source separation training data with sidechain ducking and reverb while preserving the additive property required by current separation architectures. Fine-tuning Bandit v2 with mixed-domain training (CAS + DnR v3) yields the best results across all evaluation conditions. A MUSHRA-like study confirms that models trained on our data produce perceptually better separations to pretrained baselines. We release the generation pipeline (CineAudioGen), synthetic dataset (CineAudioSynth), and a real cinematic evaluation set (CineAudioDB) with ground-truth stems in both production-mix and linear-mix versions to support future work in cinematic audio source separation.

[4] https://www.mabyduck.com/

Impact Statement

This paper aims to improve cinematic audio source separation, with potential benefits for multilingual dubbing, accessibility captioning, broadcast transcription, archival restoration, and creative post-production. These applications can make audiovisual content easier to localize, remix, and access. However, improved audio separation may also facilitate unauthorized media manipulation, removal or alteration of speech, or derivative use of copyrighted material. These risks are related to broader concerns around audio editing, voice cloning, and synthetic media. We therefore emphasize that systems and datasets based on this work should respect the rights and consent of content creators, performers, and speakers, and should be used in accordance with applicable licensing and copyright constraints.

References

Method for the subjective assessment of intermediate quality level of audio systems. Recommendation BS.1534-3, International Telecommunication Union, 2015.

Borisov, M., Spirin, E., and Diatlova, D. Nonverbaltts: A public english corpus of text-aligned nonverbal vocalizations with emotion annotations for text-to-speech. pp. 104–109, 08 2025. doi: 10.21437/SSW.2025-16.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., and Ranzato, M. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Chen, J., Guo, Z., Chun, J., Wang, P., Perrault, A., and Elsner, M. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. *arXiv preprint arXiv:2510.10444*, 2025.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

European Broadcasting Union. EBU R 128: Loudness normalisation and permitted maximum level of audio signals. Recommendation, EBU, 2023. v5.0.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208.

Ghosh, S., Goel, A., Koroshinadze, L., Lee, S.-g., Kong, Z., Santos, J. F., Duraiswami, R., Manocha, D., Ping, W., Shoeybi, M., et al. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*, 2025.

Hasumi, T. and Fujita, Y. Dnr-nonverbal: Cinematic audio source separation dataset containing non-verbal sounds. In *Proceedings of Interspeech 2025*, pp. 4993–4997, 2025. doi: 10.21437/Interspeech.2025-1148.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. Sdr – half-baked or well done? pp. 626–630, 05 2019. doi: 10.1109/ICASSP.2019.8683855.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 05 2018. doi: 10.1371/journal.pone.0196391. URL https://doi.org/10.1371/journal.pone.0196391.

Nguyen, T., Hsu, W.-N., D'Avirro, A., Shi, B., Gat, I., Fazel-Zarani, M., Remez, T., Copet, J., Synnaeve, G., Hassid, M., Kreuk, F., Adi, Y., and Dupoux, E. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. pp. 4823–4827, 08 2023. doi: 10.21437/Interspeech.2023-1905.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Petermann, D., Wichern, G., Wang, Z.-Q., and Le Roux, J. Divide and remaster (DnR). Zenodo, October 2021. URL https://zenodo.org/records/5574713.

Petermann, D., Wichern, G., Wang, Z.-Q., and Le Roux, J. The cocktail fork problem: Three-stem audio separation for real-world soundtracks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 526–530. IEEE, 2022.

Pons, J., Liu, X., Pascual, S., and Serrà, J. Gass: Generalizing audio source separation with large-scale data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 546–550. IEEE, 2024.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. Musdb18 - a corpus for music separation, December 2017. URL https://doi.org/10.5281/zenodo.1117372.

Retta, E. A., Almekhlafi, E., Sutcliffe, R., Mhamed, M., Ali, H., and Feng, J. A new amharic speech emotion dataset and classification benchmark. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(1), May 2023. ISSN 2375-4699. doi: 10.1145/3529759. URL https://doi.org/10.1145/3529759.

Rouard, S., Massa, F., and Défossez, A. Hybrid transformers for music source separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096956.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., et al. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Uhlich, S., Fabbro, G., Hirano, M., Takahashi, S., Wichern, G., Le Roux, J., Chakraborty, D., Mohanty, S., Li, K., Luo, Y., Yu, J., Gu, R., Solovyev, R., Stempkovskiy, A., Habruseva, T., Sukhovei, M., and Mitsufuji, Y. The sound demixing challenge 2023 – cinematic demixing track. *Transactions of the International Society for Music Information Retrieval*, Apr 2024. doi: 10.5334/tismir.172.

Wang, Z., Tang, C.-K., and Tai, Y.-W. Audio-agent: Leveraging LLMs for audio generation, editing and composition. *arXiv preprint arXiv:2410.03335*, 2024.

Wang, Z., Tang, C.-K., and Tai, Y.-W. ReelWave: Multi-agentic movie sound generation through multimodal LLM conversation. *arXiv preprint arXiv:2503.07217*, 2025.

Watcharasupat, K. N., Wu, C.-W., and Orife, I. Remastering divide and remaster: A cinematic audio source separation dataset with multilingual support. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–10. IEEE, 2024.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zang, Y., O'Brien, S., Berg-Kirkpatrick, T., McAuley, J., and Novack, Z. Are you really listening? Boosting perceptual awareness in music-QA benchmarks. *arXiv preprint arXiv:2504.00369*, 2025.