Finetuning Multimodal LLMs for Relative Level Analysis: Anchor-Conditioned Advice for Multitrack Music Mixing

Pratham Vadhulas
Advisor: Alexander Lerch
Fall 2025 Project Proposal

Georgia Tech · College of Design
Center for Music Technology

Brief Introduction

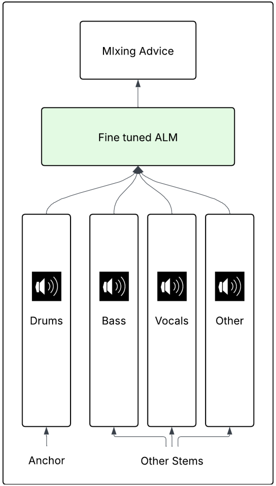
overview

- **Music mixing** requires a complex, relational understanding of multiple audio tracks, and collaboration.
- This research investigates a framework to fine-tune an **Audio-Language Model (ALM)** to generate actionable mixing advice.
- As a starting point, we condition the model on an **"anchor track"** (e.g., bass) to teach it how to balance the levels of other instruments relative to that **stable reference point**.

Brief Introduction

overview

- **Music mixing** requires a complex, relational understanding of multiple audio tracks, and collaboration.
- This research investigates a framework to fine-tune an **Audio-Language Model (ALM)** to generate actionable mixing advice.
- As a starting point, we condition the model on an **"anchor track"** (e.g., bass) to teach it how to balance the levels of other instruments relative to that **stable reference point**.

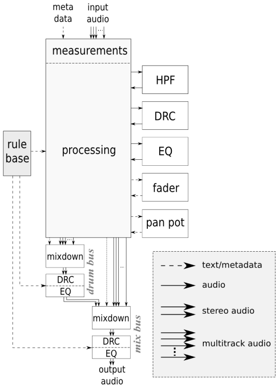


Automatic Mixing Review

Rule-Based & Deep Learning Approaches

Rule-Based and Traditional Machine Learning Systems

- Knowledge-engineered autonomous mixing [1]
- A machine-learning approach for instrument-specific application of artificial reverberation. [2]



Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing [3]
- Differentiable mixing console with neural effects [4]

Automatic Mixing Review
Rule-Based & Deep Learning Approaches

Rule-Based and Traditional Machine Learning Systems

- Knowledge-engineered autonomous mixing [1]
- A machine-learning approach for instrument-specific application of artificial reverberation. [2]

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing [3]
- Differentiable mixing console with neural effects [4]

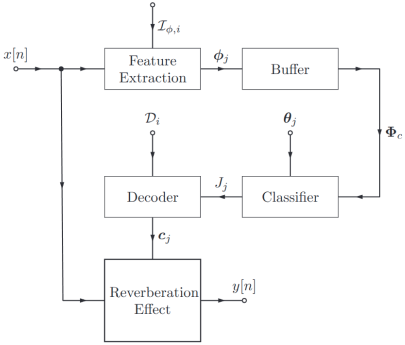


Fig. 1. Reverb application.

Automatic Mixing Review

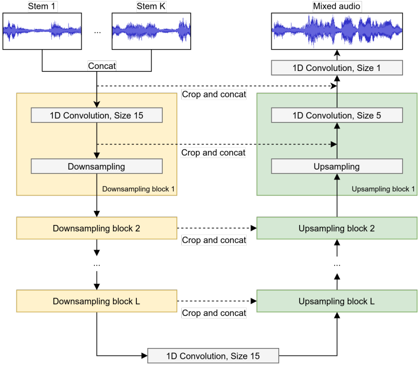
Rule-Based & Deep Learning Approaches

Rule-Based and Traditional Machine Learning Systems

- Knowledge-engineered autonomous mixing [1]
- A machine-learning approach for instrument-specific application of artificial reverberation. [2]

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing [3]
- Differentiable mixing console with neural effects [4]



Automatic Mixing Review

Rule-Based & Deep Learning Approaches

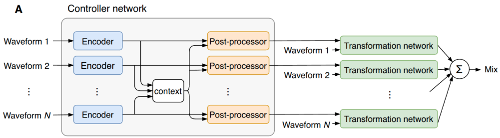
Georgia Tech · College of Design
Center for Music Technology

Rule-Based and Traditional Machine Learning Systems

- Knowledge-engineered autonomous mixing [1]
- A machine-learning approach for instrument-specific application of artificial reverberation. [2]

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing [3]
- Differentiable mixing console with neural effects [4]



Automatic Mixing Review
Semantic Approaches

Georgia Tech · College of Design
Center for Music Technology

Language-Audio Integration

- Word-embedding approaches linking audio and language for effects/EQ recommendations [5], [6], [7]
- Text-driven interfaces mapping natural language to effect parameters and mix actions [8], [9], [10]

Automatic Mixing Review

Semantic Approaches

Language-Audio Integration

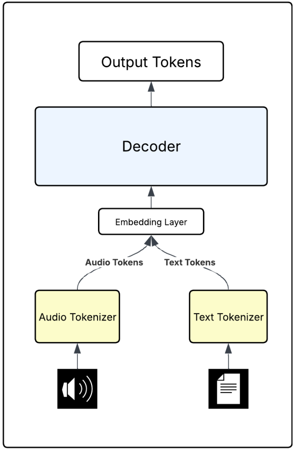
- Word-embedding approaches linking audio and language for effects/EQ recommendations [5], [6], [7]
- Text-driven interfaces mapping natural language to effect parameters and mix actions [8], [9], [10]

Automatic Mixing Review

Architectural Approaches for Audio-Language Models

Georgia Tech · College of Design
Center for Music Technology

- **Direct Tokenization (Unified Approach)**: converts raw audio into discrete tokens via audio codecs; tokens are flattened into a 1D sequence as LLM input; the LLM vocabulary is extended to include audio tokens [11], [12], [13].
- **Feature Extraction (Cascade Approach)**: uses audio-specific encoders/decoders with the LLM as a central backbone; high-level features are passed between modules (e.g., M^2UGen) [14], [15].



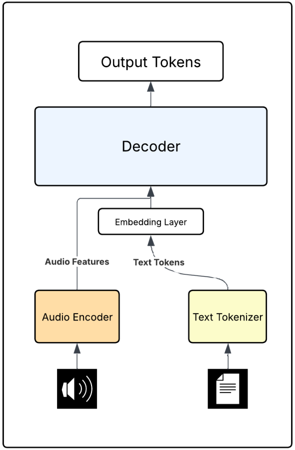
Unified Approach

Automatic Mixing Review

Architectual Approaches for Audio-Language Models

Georgia Tech College of Design Center for Music Technology

- **Direct Tokenization (Unified Approach)**: converts raw audio into discrete tokens via audio codecs; tokens are flattened into a 1D sequence as LLM input; the LLM vocabulary is extended to include audio tokens [11], [12], [13].
- **Feature Extraction (Cascade Approach)**: uses audio-specific encoders/decoders with the LLM as a central backbone; high-level features are passed between modules (e.g., M^2UGen) [14], [15].



Cascade Approach

Automatic Mixing Review

The Roles of LLMs in Audio Language Models

Georgia Tech · College of Design
Center for Music Technology

- **LLM as Backbone**: The LLM acts as the core processing engine, unifying audio and text into a single model [11], [12], [13].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models [16].
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks [17], [18], [19].

Automatic Mixing Review

The Roles of LLMs in Audio Language Models

- **LLM as Backbone**: The LLM acts as the core processing engine, unifying audio and text into a single model [11], [12], [13].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models [16].
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks [17], [18], [19].

Automatic Mixing Review

The Roles of LLMs in Audio Language Models

- **LLM as Backbone**: The LLM acts as the core processing engine, unifying audio and text into a single model [11], [12], [13].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models [16].
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks [17], [18], [19].

Research Questions

primary

Primary Research Question

- To what extent can an **Audio-Language Model**, conditioned on an **anchor track**, learn the **relative gain** relationships among multitrack stems and generate musically effective gain-balancing **advice**?

Research Questions
primary

Primary Research Question

- To what extent can an **Audio-Language Model**, conditioned on an **anchor track**, learn the **relative gain** relationships among multitrack stems and generate musically effective gain-balancing **advice**?

Research Questions

secondary

Secondary Questions

- **Model Understanding**: What model architecture best represents and reasons about multitrack stems and anchor tracks for learning relative gain relationships?
- **Mixing Conventions & Genre Awareness**: To what extent does the model's advice reflect established mixing conventions?
- **Communication & Actionability**: How effectively does the model communicate its advice in a way that is clear, actionable, and distinct from simply being "correct"?
- **Human Evaluation & Usefulness**: How do audio engineers and producers evaluate the effectiveness, musicality, and real-world usefulness of the advice in their workflows?

Research Questions

secondary

Secondary Questions

- **Model Understanding**: What model architecture best represents and reasons about multitrack stems and anchor tracks for learning relative gain relationships?
- **Mixing Conventions & Genre Awareness**: To what extent does the model's advice reflect established mixing conventions?
- **Communication & Actionability**: How effectively does the model communicate its advice in a way that is clear, actionable, and distinct from simply being "correct"?
- **Human Evaluation & Usefulness**: How do audio engineers and producers evaluate the effectiveness, musicality, and real-world usefulness of the advice in their workflows?

Research Questions

secondary

Secondary Questions

- **Model Understanding**: What model architecture best represents and reasons about multitrack stems and anchor tracks for learning relative gain relationships?
- **Mixing Conventions & Genre Awareness**: To what extent does the model's advice reflect established mixing conventions?
- **Communication & Actionability**: How effectively does the model communicate its advice in a way that is clear, actionable, and distinct from simply being "correct"?
- **Human Evaluation & Usefulness**: How do audio engineers and producers evaluate the effectiveness, musicality, and real-world usefulness of the advice in their workflows?

Research Questions

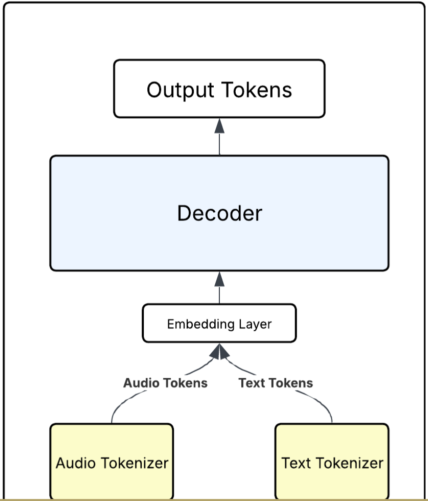
secondary

Secondary Questions

- **Model Understanding**: What model architecture best represents and reasons about multitrack stems and anchor tracks for learning relative gain relationships?
- **Mixing Conventions & Genre Awareness**: To what extent does the model's advice reflect established mixing conventions?
- **Communication & Actionability**: How effectively does the model communicate its advice in a way that is clear, actionable, and distinct from simply being "correct"?
- **Human Evaluation & Usefulness**: How do audio engineers and producers evaluate the effectiveness, musicality, and real-world usefulness of the advice in their workflows?

Proposed Method
Overview

- **Base Model**: A pretrained Audio LLM like Qwen-Audio [20] as the base model.
- **Dataset**: A multitrack dataset like MUSDB18 [21] for synthesizing the prompt-response pairs.
- **Input**: Anchor track and the rest of the tracks.
- **Output**: A structured response containing advice for the user to balance the levels of the tracks.
- **Architecture**: Cascade approach, with the LLM as the backbone.
- **Training strategy**: LoRA [22] or QLoRA [23] to fine-tune the model.



Proposed Method
Dataset Synthesis

- **Per-Sample Prompt**: We will have 4 categories of per-sample prompts:
  - **Audio-only prompts**: Response will default to giving advice.
  - **Incorrect description prompts**: Response will correct the description in addition to the advice.
  - **Advice-seeking prompts**: Response will be only advice.
  - **Correct description prompts**: Response will confirm the description in addition to the advice.
- **Audio**:
  - Chunk the song into 10-second segments.
  - Inject an error of $\pm n$ dB on the non-anchor track.
- **Response Formulation**: We will programmitically create structured responses.
  - We will first pick a template based on the prompt, containing both description and solution placeholders.
  - We populate the placeholders with the description and solution.

Evaluation Framework

Human Evaluation

- **Participants**: Semi-professional audio engineers and producers
- **Evaluation Criteria**:
  - **Effectiveness**: How well does the advice address the mixing challenge?
  - **Actionability**: How clear and implementable is the advice?
  - **Adherence to Conventions**: How well does the advice follow established mixing practices?
- **Methodology**: Rating scales and qualitative feedback collection

Evaluation Framework

Automated Evaluation

- **LLM-as-a-Judge**: Use LLMs to rate advice quality and relevance
- **Semantic Similarity**: Compare advice to expert annotations
- **Gain Advice Accuracy**:
  - Direction accuracy: Increase vs. decrease correctness
  - Magnitude accuracy: Proximity of suggested dB changes to optimal

- picture here

Limitations

- **Focus on Gain Only**: The model's scope is limited to gain-balancing advice; it does not address other effects like EQ, compression, or spatial effects.
- **Advisory, Not Prescriptive**: Evaluation focuses on the usefulness of the textual advice, not the numeric accuracy of specific gain predictions.
- **Dataset Dependency**: The project relies on the MUSDB18 dataset for valid "ground truth" for professional mixes.

Timeline

Georgia Tech · College of Design Center for Music Technology

Tasks Leading to Nov. 28th Submission

- ~~Dataset preprocessing and JSONL format conversion.~~
- ~~Initial codebase and data loading pipeline setup.~~
- ~~Partial fine-tuning experiments and architecture testing.~~
- Finalize and execute all remaining experiments.
- Submit and obtain IRB approval for human studies.
- Conduct the human evaluation study with audio professionals.
- Complete the final paper, web interface, and Hugging Face deployment.

references

Georgia Tech · College of Design
Center for Music Technology

[1] E. Pérez-González and J. D. Reiss, "A knowledge-engineered autonomous mixing system," in *Audio Engineering Society Convention 135*, Oct. 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16953.

[2] E. Chourdakis and J. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," en, *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 56–65, Feb. 2017, ISSN: 15494950. DOI: 10.17743/jaes.2016.0069.

[3] E. Chourdakis and J. Reiss, "Automatic music signal mixing system based on one-dimensional wave-u-net autoencoders," en, 2022. DOI: 10.1186/s13636-022-00266-3. [Online]. Available: https://www.researchgate.net/publication/366902955_Automatic_music_signal_mixing_system_based_on_one-dimensional_Wave-U-Net_autoencoders.

[4] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects,", no. arXiv:2010.10291, Oct. 2020, arXiv:2010.10291 [eess]. DOI: 10.48550/arXiv.2010.10291. [Online]. Available: http://arxiv.org/abs/2010.10291.

[5] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2fx: Harnessing clap embeddings for text-guided audio effects,", no. arXiv:2409.18847, Feb. 2025, arXiv:2409.18847 [eess]. DOI: 10.48550/arXiv.2409.18847. [Online]. Available: http://arxiv.org/abs/2409.18847.

[6] S. Venkatesh, D. Moffat, and E. R. Miranda, "Word embeddings for automatic equalization in audio mixing," en, *Journal of the Audio Engineering Society*, vol. 70, no. 9, pp. 753–763, Nov. 2022, ISSN: 15494950. DOI: 10.17743/jaes.2022.0047.

[7] E. Chourdakis and J. Reiss, *A semantic approach to autonomous mixing*, en, 2016. [Online]. Available: https://www.researchgate.net/publication/273574043_A_Semantic_Approach_To_Autonomous_Mixing.

[8] S. Doh, J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, J. Nam, and Y. Mitsufuji, "Can large language models predict audio effects parameters from natural language?," no. arXiv:2505.20770, Jul. 2025, arXiv:2505.20770 [cs]. DOI: 10.48550/arXiv.2505.20770. [Online]. Available: http://arxiv.org/abs/2505.20770.

references

Georgia Tech · College of Design Center for Music Technology

[9] J. Melechovsky, A. Mehrish, and D. Herremans, "Sonicmaster: Towards controllable all-in-one music restoration and mastering,", no. arXiv:2508.03448, Aug. 2025, arXiv:2508.03448 [eess]. DOI: 10.48550/arXiv.2508.03448. [Online]. Available: http://arxiv.org/abs/2508.03448.

[10] M. P. Clemens and A. Marasovic, "Mixassist: An audio-language dataset for co-creative AI assistance in music mixing,", 2025. [Online]. Available: https://openreview.net/forum?id=5mICyyD4OF.

[11] P. K. Rubenstein et al., "Audiopalm: A large language model that can speak and listen,", no. arXiv:2306.12925, Jun. 2023, arXiv:2306.12925 [cs]. DOI: 10.48550/arXiv.2306.12925. [Online]. Available: http://arxiv.org/abs/2306.12925.

[12] Z. Du et al., "Lauragpt: Listen, attend, understand, and regenerate audio with gpt,", no. arXiv:2310.04673, Jul. 2024, arXiv:2310.04673 [cs]. DOI: 10.48550/arXiv.2310.04673. [Online]. Available: http://arxiv.org/abs/2310.04673.

[13] D. Zhang et al., "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,", no. arXiv:2305.11000, May 2023, arXiv:2305.11000 [cs]. DOI: 10.48550/arXiv.2305.11000. [Online]. Available: http://arxiv.org/abs/2305.11000.

[14] S. Liu, A. S. Hussain, Q. Wu, C. Sun, and Y. Shan, "M^2ugen: Multi-modal music understanding and generation with the power of large language models,", no. arXiv:2311.11255, Dec. 2024, arXiv:2311.11255 [cs]. DOI: 10.48550/arXiv.2311.11255. [Online]. Available: http://arxiv.org/abs/2311.11255.

[15] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand,", no. arXiv:2305.10790, Feb. 2024, arXiv:2305.10790 [eess]. DOI: 10.48550/arXiv.2305.10790. [Online]. Available: http://arxiv.org/abs/2305.10790.

[16] J. Copet et al., "Simple and controllable music generation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23, New Orleans, LA, USA: Curran Associates Inc., 2023.

references

[17] R. Huang et al., "Audiogpt: Understanding and generating speech, music, sound, and talking head,", no. arXiv:2304.12995, Apr. 2023, arXiv:2304.12995 [cs]. DOI: 10.48550/arXiv.2304.12995. [Online]. Available: http://arxiv.org/abs/2304.12995.

[18] X. Liu et al., "Wavjourney: Compositional audio creation with large language models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2830–2844, 2025, ISSN: 2998-4173. DOI: 10.1109/TASLPRO.2025.3574867.

[19] D. Yu et al., "Musicagent: An ai agent for music understanding and generation with large language models,", no. arXiv:2310.11954, Oct. 2023, arXiv:2310.11954 [cs]. DOI: 10.48550/arXiv.2310.11954. [Online]. Available: http://arxiv.org/abs/2310.11954.

[20] Y. Chu et al., "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,", no. arXiv:2311.07919, Dec. 2023, arXiv:2311.07919 [eess]. DOI: 10.48550/arXiv.2311.07919. [Online]. Available: http://arxiv.org/abs/2311.07919.

[21] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, *Musdb18-hq - an uncompressed version of musdb18*, Aug. 2019. DOI: 10.5281/zenodo.3338373. [Online]. Available: https://doi.org/10.5281/zenodo.3338373.

[22] E. J. Hu et al., "Lora: Low-rank adaptation of large language models,", no. arXiv:2106.09685, Oct. 2021, arXiv:2106.09685 [cs]. DOI: 10.48550/arXiv.2106.09685. [Online]. Available: http://arxiv.org/abs/2106.09685.

[23] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023, arXiv:2305.14314. [Online]. Available: https://arxiv.org/abs/2305.14314.