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

Pratham Vadhulas and Alexander Lerch

Fall 2025 Project Proposal

Georgia Tech · College of Design
Center for
Music Technology

Brief Introduction

overview

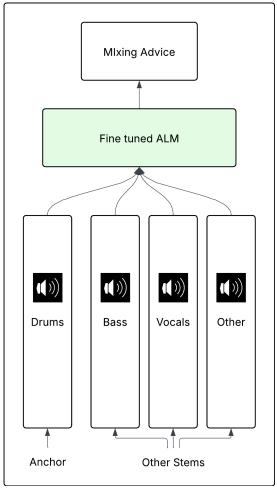
Georgia Tech · College of Design
Center for Music Technology

- **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**.



Related Work
Rule-Based Systems

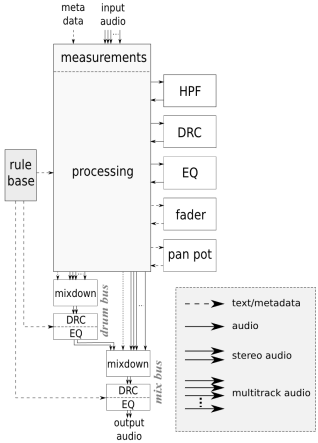
Georgia Tech · College of Design
Center for Music Technology

Expert-Derived Rules

- A knowledge-engineered autonomous mixing system [1]

Instrument-Specific Processing

- A machine-learning approach to intelligent artificial reverberation **placeholder-reverb**



Related Work
Rule-Based Systems

Expert-Derived Rules

- A knowledge-engineered autonomous mixing system [1]

Instrument-Specific Processing

- A machine-learning approach to intelligent artificial reverberation **placeholder-reverb**

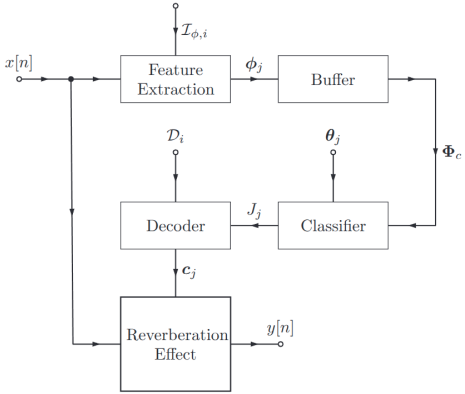


Fig. 1. Reverb application.

Related Work

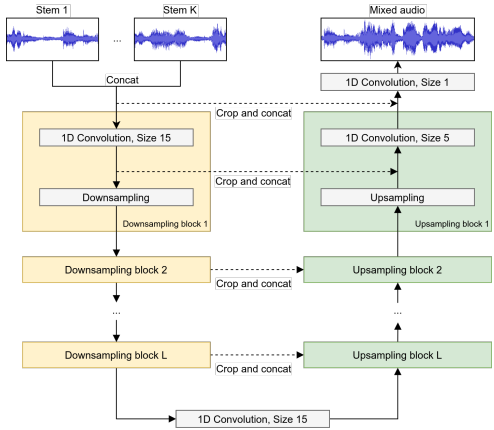
Deep Learning Architectures

Wave-U-Net and Autoencoders

- Automatic music signal mixing with 1D Wave-U-Net autoencoders **placeholder-waveunet**

Differentiable Mixing Consoles

- Automatic multitrack mixing with a differentiable mixing console of neural audio effects [2]



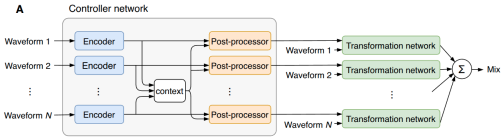
Related Work
Deep Learning Architectures

Wave-U-Net and Autoencoders

- Automatic music signal mixing with 1D Wave-U-Net autoencoders **placeholder-waveunet**

Differentiable Mixing Consoles

- Automatic multitrack mixing with a differentiable mixing console of neural audio effects [2]



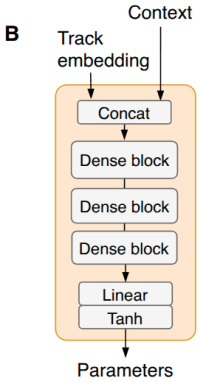
Related Work
Deep Learning Architectures

Wave-U-Net and Autoencoders

- Automatic music signal mixing with 1D Wave-U-Net autoencoders **placeholder-waveunet**

Differentiable Mixing Consoles

- Automatic multitrack mixing with a differentiable mixing console of neural audio effects [2]



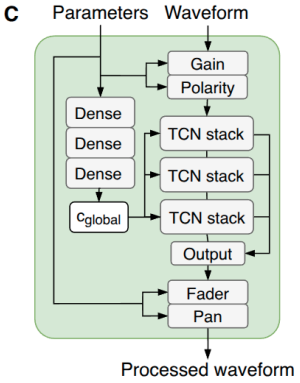
Related Work
Deep Learning Architectures

Wave-U-Net and Autoencoders

- Automatic music signal mixing with 1D Wave-U-Net autoencoders **placeholder-waveunet**

Differentiable Mixing Consoles

- Automatic multitrack mixing with a differentiable mixing console of neural audio effects [2]



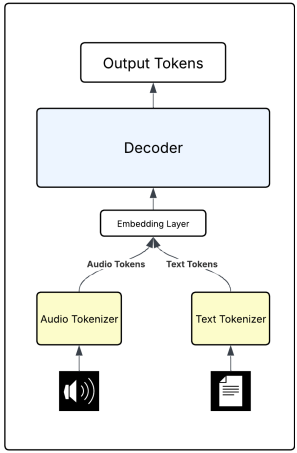
Related Work
Semantic Approaches

- **Word-embedding approaches** learn semantic spaces linking audio and language to steer effects/EQ recommendations and retrieval-guided processing **placeholder-semantic-mixing**, [3], [4].
- **Text-driven interfaces** map natural-language instructions to effect parameters and mix actions, enabling promptable, co-creative workflows **Clemens˙Marasovic˙2025**, [5], [6].

Multi-Modal Audio LLMs

Architectural Approaches

- 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 (e.g., AudioPaLM, LauraGPT, SpeechGPT) [7], [8], [9].
- 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^{2}UGen; LTU placeholder) **placeholder-ltu**, [10].

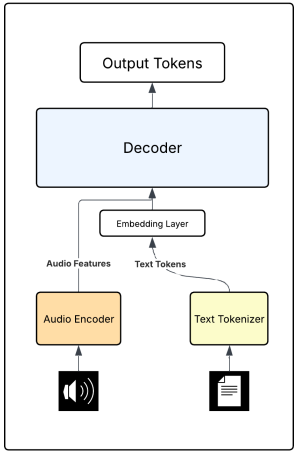


Unified Approach

Multi-Modal Audio LLMs

Architectural Approaches

- 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 (e.g., AudioPaLM, LauraGPT, SpeechGPT) [7], [8], [9].
- 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^{2}UGen; LTU placeholder) **placeholder-ltu**, [10].



Cascade Approach

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, and how does its performance vary across different musical genres?
- **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, and how does its performance vary across different musical genres?
- **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, and how does its performance vary across different musical genres?
- **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, and how does its performance vary across different musical genres?
- **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?

references

[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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

references

[9] 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.

[10] 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.