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

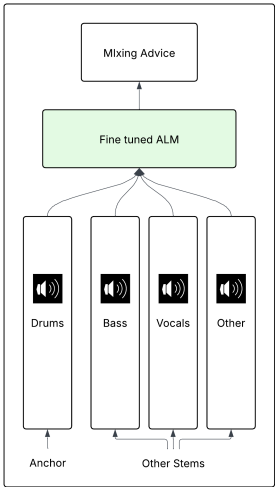
overview

- **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**.



Related Work

Rule-Based & Deep Learning Approaches

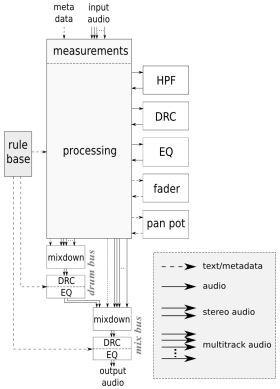
Georgia Tech · College of Design
Center for Music Technology

Rule-Based and Traditional Machine Learning Systems

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

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing **placeholder-waveunet**
- Differentiable mixing console with neural effects [2]



Related Work

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. **placeholder-reverb**

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing **placeholder-waveunet**
- Differentiable mixing console with neural effects [2]

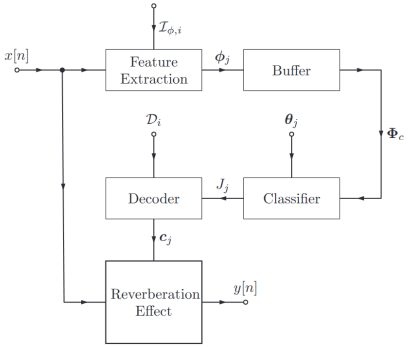


Fig. 1. Reverb application.

Related Work

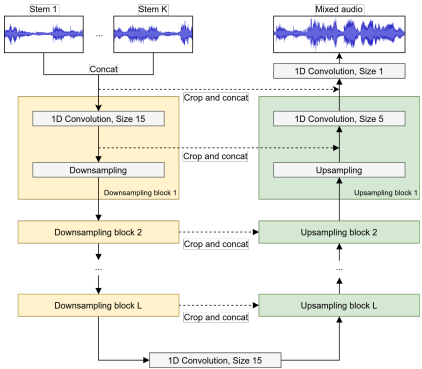
Rule-Based & Deep Learning Approaches

Rule-Based and Traditional Machine Learning Systems

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

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing **placeholder-waveunet**
- Differentiable mixing console with neural effects [2]



Related Work

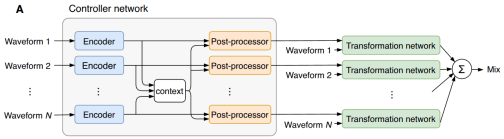
Rule-Based & Deep Learning Approaches

Rule-Based and Traditional Machine Learning Systems

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

Deep Learning Architectures

- Wave-U-Net autoencoders for automatic mixing **placeholder-waveunet**
- Differentiable mixing console with neural effects [2]



Related Work
Semantic Approaches

Language-Audio Integration

- Word-embedding approaches linking audio and language for effects/EQ recommendations **placeholder-semantic-mixing**, [3], [4]
- Text-driven interfaces mapping natural language to effect parameters and mix actions **Clemens˙Marasovic˙2025**, [5], [6]

Related Work
Semantic Approaches

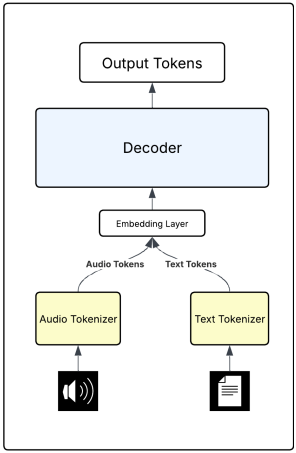
Language-Audio Integration

- Word-embedding approaches linking audio and language for effects/EQ recommendations **placeholder-semantic-mixing**, [3], [4]
- Text-driven interfaces mapping natural language to effect parameters and mix actions **Clemens˙Marasovic˙2025**, [5], [6]

Multi-Modal LLMs

Architectural Approaches for Audio-Language Models

- 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].

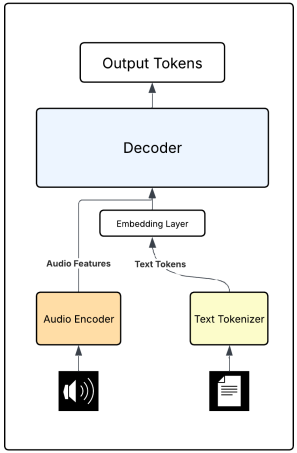


Unified Approach

Multi-Modal LLMs

Architectural Approaches for Audio-Language Models

- 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

Multi-Modal LLMs

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 [7], [8], [9].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models **placeholder-musicgen**, **placeholder-tango**.
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks **placeholder-audiogpt**, **placeholder-wavjourney**, **placeholder-musicagent**.

Multi-Modal LLMs

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 [7], [8], [9].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models **placeholder-musicgen**, **placeholder-tango**.
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks **placeholder-audiogpt**, **placeholder-wavjourney**, **placeholder-musicagent**.

Multi-Modal LLMs

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 [7], [8], [9].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models **placeholder-musicgen**, **placeholder-tango**.
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks **placeholder-audiogpt**, **placeholder-wavjourney**, **placeholder-musicagent**.

Multi-Modal LLMs

The Roles of LLMs in Audio Language Models

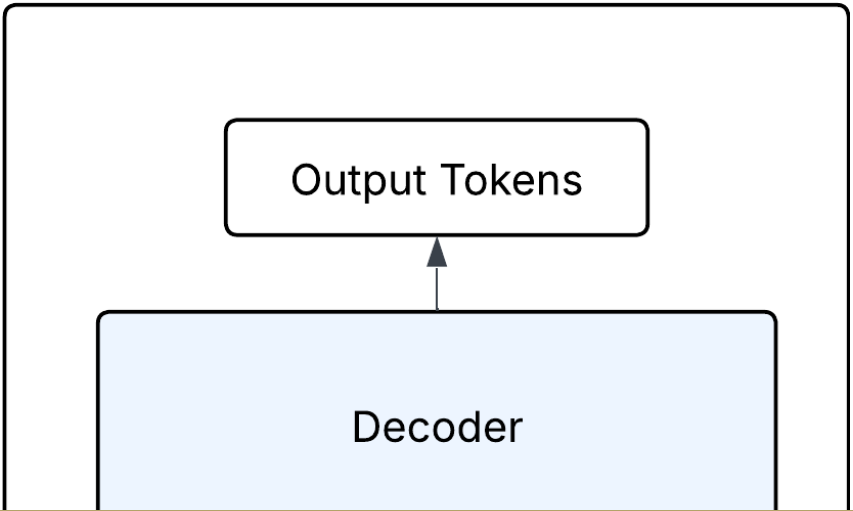
- **LLM as Backbone**: The LLM acts as the core processing engine, unifying audio and text into a single model [7], [8], [9].
- **LLM as Conditioner**: The LLM converts text instructions into embeddings to guide audio generation models **placeholder-musicgen**, **placeholder-tango**.
- **LLM as Agent**: The LLM acts as an intelligent controller, orchestrating multiple specialized AI tools to execute complex audio tasks **placeholder-audiogpt**, **placeholder-wavjourney**, **placeholder-musicagent**.

Proposed Method

- **Multimodal LLM Architecture**
  - Audio encoder for multitrack analysis
  - Text encoder for natural language instructions
  - Cross-modal attention mechanisms
- **Anchor-Conditioned Training**
  - Reference track conditioning
  - Relative level prediction
  - Context-aware mixing advice
- **Training Data**
  - Professional mixing examples
  - Synthetic multitrack datasets
  - Expert annotations and feedback

Proposed Method

Architecture Overview



Proposed Method

Key Components

Audio Processing

- Multitrack feature extraction
- Temporal context modeling
- Spectral analysis

Integration

- Cross-modal attention mechanisms
- Joint representation learning
- Contextual advice generation

Text Processing

- Natural language understanding
- Mixing terminology mapping
- Instruction parsing

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?

Evaluation
Human Evaluation

Study with Semi-Professional Audio Engineers

- **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

Automated Evaluation

Complementary Metrics to Human Ratings

- **LLM-as-a-Judge**: Using language models to evaluate the quality and relevance of generated advice
- **Semantic Similarity**: Measuring similarity between generated advice and expert annotations
- **Gain Advice Accuracy**: Evaluating the direction and magnitude of gain recommendations
  - Direction accuracy: Does the model suggest increasing or decreasing levels correctly?
  - Magnitude accuracy: How close are the suggested gain changes to optimal values?

Evaluation

Evaluation Framework

Human Evaluation

- Real-world applicability
- Professional judgment
- Workflow integration

Automated Evaluation

- Scalable assessment
- Objective metrics
- Reproducible results

Combined Approach: Human evaluation provides ground truth for real-world effectiveness, while automated metrics enable systematic model comparison and iteration.

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.