Exploring Vocal Percussion Feature Extraction Methods

Pratham Vadhulas
Music Informatics Group, Georgia Institute of Technology

Georgia Tech · College of Design
Center for
Music Technology

Brief Introduction

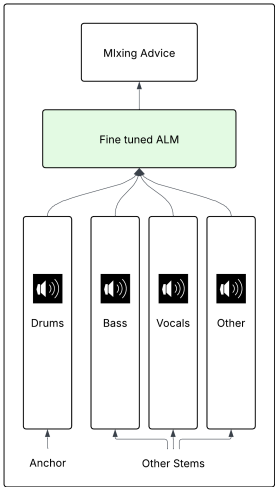
overview

- **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**.



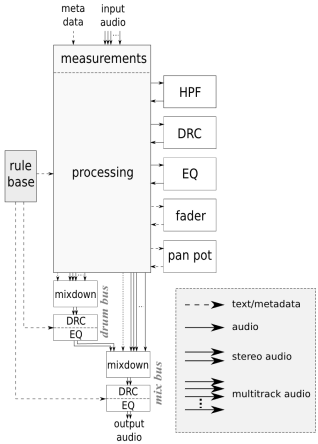
Related Work
Rule-Based Systems

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**

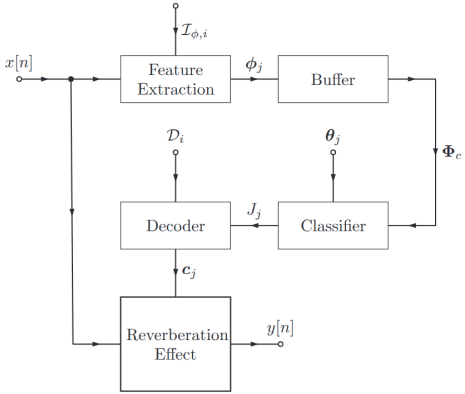


Fig. 1. Reverb application.

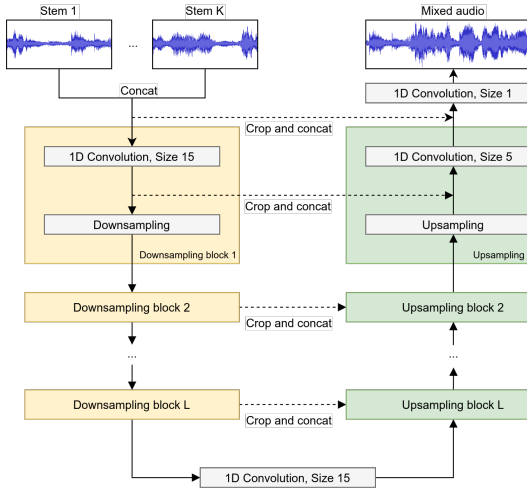
Related Work
Deep Learning Architectures

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
Semantic Approaches

Word Embeddings

- Text2FX: Harnessing CLAP Embeddings for Text-Guided Audio Effects [3]
- A Semantic Approach to Autonomous Mixing **placeholder-semantic-mixing**
- Word Embeddings for Automatic Equalization in Audio Mixing [4]

Text-Driven Interfaces

- Can Large Language Models Predict Audio Effects Parameters from Natural Language? [5]
- SonicMaster: Towards Controllable

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.