MixingBuddy: A Multimodal LLM for Mix Critique and Advice

Pratham Vadhulas, Alexander Lerch

***Abstract*—this is the abstract. here is some more text. here is some more text.**

I. Introduction

Automatic Mixing, a key subfield of Music Informatics Research (MIR), aims to automate the complex and subjective task of music mixing. This area of study is pivotal to the modern music production and audio engineering market. To date, research in this field has made significant progress, largely by leveraging deep learning. Sophisticated models, such as U-Nets or generative frameworks, have been developed to make mixing systems accurate (predicting parameters that match professional mixes), controllable (allowing for high-level parameters to be set), and diverse (accomadating different genres and styles).

However, a critical limitation of these approaches is their "black box" nature. They can perform the mix, but they cannot explain their reasoning. The recent promise of large language models (LLMs) and multi-modal "agentic" systems introduces a new, necessary paradigm: explainability. We can now envision a tool that reasons about and discusses a mix.

This potential for co-creative, linguistic feedback brings us to our core research question: To what extent can an audio language model, when given a flawed mix, provide correct and useful advice?

II. Related Work

This section reviews the evolution of automatic mixing systems, from rule-based approaches to deep learning architectures, and examines the emerging paradigm of semantic and language-driven methods. We then survey multimodal audio-language models that enable explainable, reasoning-based audio processing, positioning our work within this landscape.

A. Automatic Mixing Systems

1) Rule-Based and Traditional Machine Learning Approaches: Early automatic mixing research centered on systems that captured expert knowledge through explicit mixing rules and heuristics. [1] developed an autonomous mixing system based on knowledge engineering principles, demonstrating that rule-based approaches could achieve reasonable results for specific mixing tasks. Subsequent work explored machine learning techniques for instrument-specific effects, such as [2]'s approach to intelligent artificial reverberation application. While these methods provided interpretable control and domain-specific optimization, they were limited in their ability to generalize across diverse musical styles and lacked the flexibility to adapt to novel mixing scenarios.

2) Deep Learning Architectures: The advent of deep learning brought significant advances in automatic mixing, with models capable of learning complex mappings from raw audio to mixing parameters. [3] introduced Wave-U-Net autoencoders for automatic mixing, demonstrating that end-to-end neural architectures could produce professional-quality mixes. [4] further advanced the field with a differentiable mixing console incorporating neural audio effects, enabling gradient-based optimization of mixing parameters. These deep learning approaches achieved notable success in terms of accuracy (matching professional mixes), controllability (allowing high-level parameter adjustment), and diversity (accommodating different genres and styles). However, a critical limitation of these systems is their "black box" nature: while they can perform the mix, they cannot explain their reasoning or provide linguistic feedback about mixing decisions.

B. Semantic and Language-Driven Approaches

1) Language-Audio Integration: Recognizing the need to bridge the semantic gap between audio processing and human understanding, researchers began exploring word-embedding approaches that link natural language descriptions to audio effect parameters. [5] demonstrated word embeddings for automatic equalization in audio mixing, while [6] developed Text2FX, which harnesses CLAP embeddings for text-guided audio effects. Early semantic mixing approaches [7] laid the groundwork for understanding how high-level, semantic knowledge could inform mixing decisions. These methods represented initial attempts to make mixing systems more interpretable and user-friendly by connecting linguistic descriptions to audio processing parameters.

2) Prompt-Driven Interfaces: Building on language-audio integration, recent work has explored prompt-driven interfaces that map natural language instructions directly to mixing tasks. [8] investigated whether large language models can predict audio effects parameters from natural language, while [9] developed SonicMaster, a controllable all-in-one music restoration and mastering system. [10] introduced MixAssist, an audio-language dataset for co-creative AI assistance in music mixing, demonstrating the potential for collaborative human-AI mixing workflows. These approaches represent an evolution toward more natural interaction paradigms, where users can express mixing intentions in natural language rather than manipulating low-level parameters.

Pratham Vadhulas is with the Georgia Institute of Technology, Atlanta, GA, USA (email: pvadhulas3@gatech.edu).

Alexander Lerch is with the Georgia Institute of Technology, Atlanta, GA, USA (email: alexander.lerch@gatech.edu).

C. Multimodal Audio-Language Models

1) Unified Tokenization Approaches: Multimodal audio-language models have emerged as a powerful paradigm for combining audio understanding with language reasoning capabilities. One architectural approach, direct tokenization (also known as the unified approach), converts raw audio into discrete tokens via audio codecs and extends the LLM vocabulary to include these audio tokens. This enables the language model to process audio and text within a unified framework. Key works in this direction include [11]'s AudioPaLM, [12]'s LauraGPT, and [13]'s SpeechGPT. The unified approach offers the advantage of treating audio and text as first-class citizens within the same model architecture, potentially enabling more seamless cross-modal reasoning.

2) Cascade Feature Extraction Approaches: An alternative architectural paradigm, the cascade (or feature extraction) approach, uses audio-specific encoders and decoders with the LLM serving as a central backbone. In this framework, audio is first encoded into feature representations that are then processed by the language model, which can generate text responses or guide audio generation. Examples include [14]'s M^2UGen and [15]'s Listen, Think, and Understand (LTU). The cascade approach allows for specialized audio processing while leveraging the reasoning capabilities of large language models, making it particularly relevant for tasks requiring both audio understanding and linguistic explanation, such as mix critique and advice.

D. Gap Analysis and Positioning

While existing approaches have made significant progress in automatic mixing, deep learning architectures, and language-audio integration, there remains a critical gap: the lack of systems that can provide explainable, reasoning-based mix critique. Rule-based and deep learning mixing systems can perform mixing tasks but cannot explain their decisions. Semantic approaches connect language to parameters but do not enable conversational reasoning about mixes. Multimodal audio-language models show promise for explainable audio processing, but their application to mix critique and advice remains largely unexplored.

Our work, MixingBuddy, addresses this gap by leveraging multimodal audio-language models to provide linguistic feedback and reasoning about mixes. By combining the audio understanding capabilities of cascade architectures with the reasoning and explanation abilities of large language models, we enable a new paradigm where AI systems can not only perform mixing tasks but also discuss, critique, and advise on mixing decisions in natural language. This positions MixingBuddy as a step toward explainable, co-creative mixing systems that bridge the gap between automated processing and human understanding.

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] E. Chourdakis and J. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, p. 56–65, Feb. 2017.

[3] ——, "Automatic music signal mixing system based on one-dimensional wave-u-net autoencoders," *EURASIP Journal on Audio, Speech, and Music Processing*, 2022. [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]. [Online]. Available: http://arxiv.org/abs/2010.10291

[5] S. Venkatesh, D. Moffat, and E. R. Miranda, "Word embeddings for automatic equalization in audio mixing," *Journal of the Audio Engineering Society*, vol. 70, no. 9, p. 753–763, Nov. 2022.

[6] 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]. [Online]. Available: http://arxiv.org/abs/2409.18847

[7] E. Chourdakis and J. Reiss, "A semantic approach to autonomous mixing," 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]. [Online]. Available: http://arxiv.org/abs/2505.20770

[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]. [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, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, "Audiopalm: A large language model that can speak and listen," no. arXiv:2306.12925, Jun. 2023, arXiv:2306.12925 [cs]. [Online]. Available: http://arxiv.org/abs/2306.12925

[12] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng, C. Zhou, Z. Yan, and S. Zhang, "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," no. arXiv:2310.04673, Jul. 2024, arXiv:2310.04673 [cs]. [Online]. Available: http://arxiv.org/abs/2310.04673

[13] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," no. arXiv:2305.11000, May 2023, arXiv:2305.11000 [cs]. [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]. [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]. [Online]. Available: http://arxiv.org/abs/2305.10790