Neural Audio Sculpting: Towards Autonomous Multitrack Sonic Transformations

Bryan Wira
bryanchwira12@gmail.com

***Abstract*—This paper explores the nascent application of deep neural networks to the intricate process of multitrack audio manipulation, aiming to redefine traditional mixing paradigms. We survey existing intelligent audio production systems and highlight the emerging integration of deep learning techniques, particularly in content-based audio transformations. A key proposition is a research trajectory focused on leveraging deep learning for intelligent music production. As a foundational proof of concept, we present a deep autoencoder architecture designed to learn and apply implicit audio effect chains to individual stems based on raw input and processed target frequency content. Preliminary results demonstrate the network's capacity to retain core harmonic and envelope characteristics, despite introducing some artifacts. This work lays the groundwork for future systems capable of autonomous or assistive sonic sculpting in complex audio environments.**

I. Introduction

Multitrack audio mixing constitutes one of the most intricate and creative stages within the music production pipeline, aimed at achieving sonic clarity, aesthetic cohesion, and resolving spectral masking conflicts among overlapping sources. Traditionally, engineers have employed nuanced adjustments of dynamics, spatial positioning, timbre, and pitch to sculpt a balanced mix. However, replicating such processes within an automated framework has posed considerable technical and perceptual challenges. In recent years, research has increasingly focused on developing intelligent systems capable of performing mixing tasks autonomously or as assistive tools for human operators.

The automation of multitrack mixing has progressed from rule-based systems and adaptive audio effects to more sophisticated, content-aware transformations. Within these paradigms, the processing of an individual audio stem is rarely an isolated operation; rather, it depends significantly on the sonic context provided by accompanying stems. Consequently, effective automated systems necessitate the extraction of perceptually meaningful audio attributes to facilitate cross-adaptive processing strategies.

A substantial body of work has leveraged expert knowledge to guide the design of intelligent mixing systems, wherein engineering heuristics and production practices are codified and evaluated through empirical experimentation. Simultaneously, data-driven approaches employing low-level audio descriptors, statistical learning models, and machine learning algorithms have advanced the capacity to infer mixing parameters, group instrument families, and approximate stylistic preferences.

For instance, linear state-space models, least-squares optimization, ensemble methods, and hierarchical clustering techniques have been applied to model complex interactions within a mix. Furthermore, heuristic search strategies such as genetic algorithms have proven effective in exploring the subjective dimensions of mixing, including user preferences and genre-specific aesthetics.

Despite these advancements, human-engineered mixdowns consistently outperform algorithmic results, particularly in domains requiring interpretive nuance, emotional conveyance, and creative deviation from standardized practices. Existing automated approaches often emphasize technical optimization—such as dynamic range control, spectral balance, or artifact suppression—while neglecting the more elusive aspects of musical expressiveness and artistic intent. This limitation impedes generalization to unconventional genres, experimental production styles, or atypical sound sources.

Accordingly, bridging the gap between expert-driven heuristics and data-centric machine learning holds considerable promise for the development of next-generation automated or assistive mixing systems. Such hybrid frameworks could leverage deep learning architectures to extract high-level representations of musical content, while integrating domain expertise to maintain creative fidelity and genre adaptability.

II. Background

The intersection of deep learning and music technology has catalyzed significant progress in recent years, with neural architectures increasingly deployed across diverse musical tasks ranging from composition to production. This section presents a concise overview of relevant advancements, focusing on the role of deep neural networks in both general musical domains and specialized music engineering applications.

A. Deep Neural Architectures in Musical Domains

The application of deep learning within music-related research has transformed the way computational models interpret, generate, and manipulate audio and symbolic musical data. Neural networks, particularly those with deep and complex architectures, have demonstrated exceptional capability in tasks involving music information retrieval, content-based recommendation systems, genre classification, and auditory event detection.

A notable area of progress lies in the generative capacity of deep neural systems. Architectures such as Wavenet-based autoencoders exemplify this advancement by enabling fine-grained, sample-wise audio synthesis and the interpolation of instrument timbres in the latent space. These models

eliminate reliance on handcrafted signal descriptors, instead learning rich, hierarchical representations directly from raw audio waveforms.

Complementary approaches leverage recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) architectures, to model the temporal dependencies inherent in musical sequences. Reinforcement learning has been combined with LSTM structures to facilitate adaptive sequence generation, enabling models to respond dynamically to evolving musical contexts. Moreover, character-level LSTM generators have shown effectiveness in symbolic domains, producing coherent note sequences, melodies, and even complex musical scores in symbolic formats such as MIDI.

These developments collectively highlight the capacity of deep neural networks to extend beyond analytical tasks and into the domain of music generation, paving the way for more expressive and context-aware music technologies.

B. Neural Models in Music Engineering

Beyond creative generation, neural networks have also been instrumental in advancing technical aspects of music production, including source separation, remixing, and automated mastering. In particular, supervised learning approaches employing convolutional neural networks (CNNs) have achieved notable success in isolating vocal and instrumental components from complex audio mixtures, facilitating flexible post-production workflows.

Deep source separation techniques enable users to manipulate individual stems within a mix, offering unprecedented control over arrangement and sonic balance. These capabilities are critical in remixing, re-arrangement, and restoration tasks, where traditional signal processing approaches may struggle to cleanly isolate sources without introducing artifacts.

In the domain of audio mastering, pretrained autoencoder frameworks have been combined with auditory perception principles, such as masking models, to automate processes like dynamic range compression and tonal balance adjustment. These systems explore both supervised and unsupervised mappings between raw audio inputs and professionally processed outputs, demonstrating that neural networks can internalize the nuanced transformations typically performed by experienced engineers.

Furthermore, the integration of generative models with automated production pipelines creates a feedback loop wherein compositional tools and production technologies mutually reinforce each other. This synergy presents opportunities for developing end-to-end intelligent systems capable of both creating and refining musical content with minimal human intervention, without compromising on artistic or technical standards.

The growing body of research underscores the potential for neural models to not only enhance efficiency and consistency in music engineering but also to support creative exploration, ultimately bridging the gap between machine learning methodologies and human musical intuition.

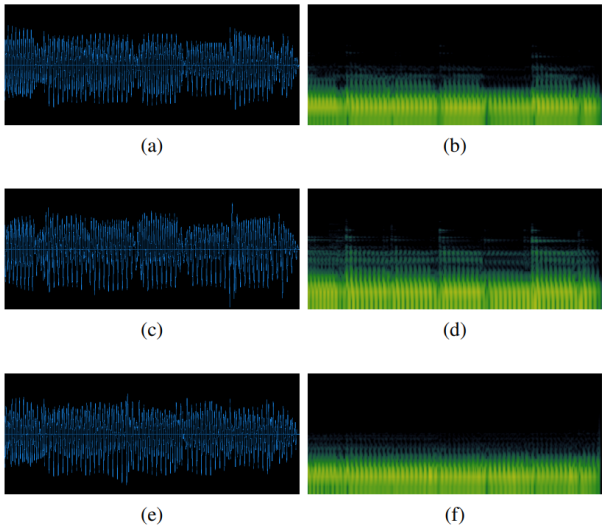

Fig. 1. Bass Track: (a)(b) Raw Input, (c)(d) Target Stem, (e)(f) Model Output

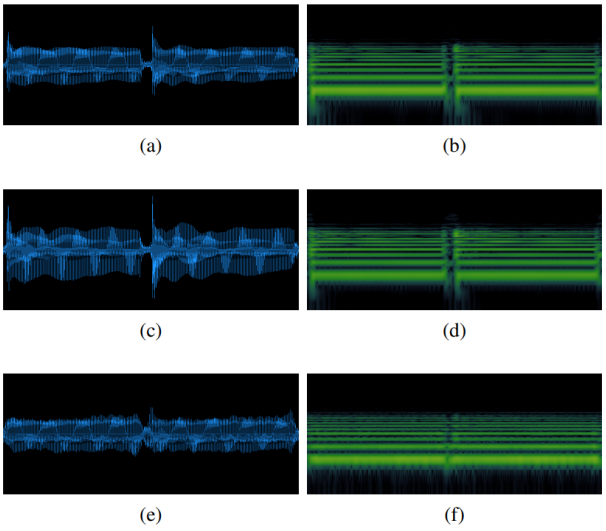

Fig. 2. Guitar Track: (a)(b) Raw Input, (c)(d) Target Stem, (e)(f) Model Output

III. Proposed Framework

This research introduces a novel framework for autonomous multitrack audio transformation that leverages deep neural networks to emulate mixing processes, bypassing the explicit use of conventional signal processing tools such as equalizers, compressors, or limiters. The objective is to investigate whether an end-to-end trained neural model can inherently learn the complex relationships and perceptual criteria involved in professional audio mixing, relying solely on data-driven representations of raw and processed signals.

At the core of this framework is a deep learning architecture designed to perform content-aware, context-sensitive transformations on individual audio stems within a multitrack environment. Unlike traditional systems, which rely on explicitly programmed signal chains or manual parameter tuning, the

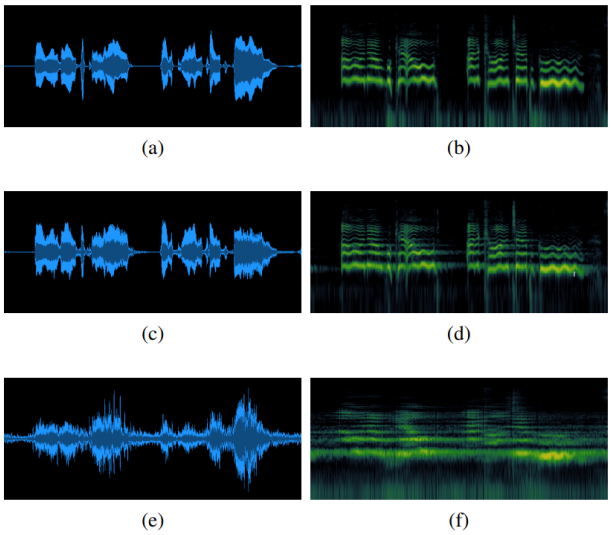

Fig. 3. Vocal Track: (a)(b) Raw Input, (c)(d) Target Stem, (e)(f) Model Output

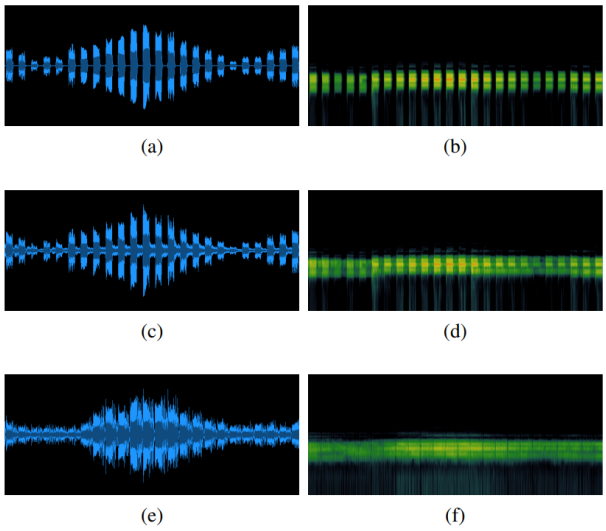

Fig. 4. Keys Track: (a)(b) Raw Input, (c)(d) Target Stem, (e)(f) Model Output

proposed model aims to internalize mixing operations through exposure to paired examples of unprocessed recordings and their corresponding professionally mixed outputs.

To enhance the realism and musical coherence of the generated outputs, the framework incorporates curated engineering knowledge in the form of statistical feature embeddings. These embeddings capture essential mix characteristics, such as spectral balance, dynamic range, and spatial distribution, providing the model with interpretable high-level descriptors. Style-adaptive mechanisms are integrated into the system, allowing the model to condition its transformations based on genre-specific attributes or user-defined aesthetic preferences.

Crucially, the proposed system is designed to function as an interactive tool rather than a fully autonomous black-box solution. By enabling iterative user feedback and post-processing refinement, the framework ensures that human auditory judgment remains central to the production workflow. This hybrid human-machine collaboration seeks to merge the computational precision and efficiency of neural networks with the creative nuance, contextual awareness, and subjective interpretation that experienced audio engineers bring to the mixing process.

Ultimately, this research aims to contribute towards the development of intelligent production tools capable of assisting or automating complex sonic transformations, while preserving the artistic integrity, emotional impact, and genre adaptability inherent to high-quality music production.

IV. Experimental Validation

To empirically evaluate the proposed framework for autonomous multitrack audio transformation, a systematic set of experiments was conducted using a curated dataset representative of real-world music production scenarios. The dataset comprises 102 multitrack songs spanning Western musical genres, with each session providing isolated instrument recordings alongside corresponding professionally mixed stems. All audio files were sampled at a standard rate of 44.1 kHz to maintain consistency and preserve high-fidelity content.

For each multitrack arrangement, a high-energy 10-second excerpt was manually selected, prioritizing sections with dense instrumentation or heightened dynamic activity to maximize the complexity of the mixing task. Both the raw instrument recordings and their corresponding mixed stems were down-mixed to mono channels and normalized using perceptual loudness algorithms, ensuring uniform playback levels and mitigating amplitude discrepancies across the dataset.

To enhance the diversity of training examples and promote robustness to pitch variations, a data augmentation strategy was implemented. Specifically, pitch-based augmentation involved semitone shifts of ± 4 at fine-grained 50-cent intervals, effectively expanding the dataset's coverage of tonal variations while preserving the underlying spectral characteristics.

Spectral representations of the audio signals were extracted using the Short-Time Fourier Transform (STFT), employing a window size of 2048 samples with a hop size of 1024 samples. This configuration provided an optimal balance between temporal resolution and frequency detail. Importantly, the final test set constituted 10% of the total available clips, excluding any augmented versions, to ensure unbiased evaluation of model generalization.

The core modeling approach utilized deep autoencoder (DAE) architectures configured as feedforward networks with three hidden layers, each comprising 1024 units. The networks were trained using a greedy layer-wise pretraining strategy, followed by fine-tuning with the Adam optimizer. ReLU activation functions were employed to introduce nonlinearity, and a dropout rate of 0.2 was applied to mitigate overfitting. The loss function was defined as the mean absolute error (MAE) between the reconstructed output and the ground truth mixed stem. Both input and output vectors consisted of 1025 spectral components, corresponding to the positive frequency bins of the STFT.

To account for timbral diversity and instrument-specific characteristics, separate DAEs were independently trained for each instrument family, including Bass, Guitar, Vocals, and Keys. After 100 epochs of training, the reconstructed outputs demonstrated promising fidelity to the target stems, with primary harmonic content and temporal envelope structures largely preserved.

Notably, reconstruction quality varied across instrument categories. Bass and Guitar stems yielded more accurate reproductions, attributable to their relatively consistent spectral profiles and narrower functional roles within the mix. In contrast, Vocals and Keys exhibited greater variance and artifacts in the reconstructed outputs, likely due to their complex harmonic structures, dynamic range fluctuations, and multi-functional roles across different genres and arrangements.

These initial results validate the feasibility of employing deep autoencoders for content-driven audio transformations that approximate mixing operations. However, observed limitations in reconstructing certain instrument families underscore the need for further refinement, particularly in capturing intricate spectral nuances and expressive performance variations.

TABLE I
DISTRIBUTION OF SOURCE MATERIAL AND AUGMENTED CLIPS

Group	Instrument	Raw/Stem	Augmented
Bass	Electric Bass	96/62	1020
	Synth Bass	12/6	–
Guitar	Clean Guitar	112/36	1224
	Acoustic	55/24	–
	Distorted	78/20	–
Vocal	Male Singer	145/36	969
	Female Singer	61/22	–
	Rapper	12/2	–
Keys	Piano	113/38	884
	Synth Lead	51/17	–
	Tack Piano	27/7	–
	Electric Piano	3/3	–

V. CONCLUSION

This research presents a foundational investigation into the application of deep autoencoder (DAE) architectures for emulating multitrack audio mixing transformations. The experimental results demonstrate that DAEs possess the inherent capacity to approximate certain aspects of professional stem-based mixing, including the preservation of primary harmonic structures and temporal envelope characteristics. These findings underscore the viability of leveraging neural networks to model complex, content-dependent audio transformations without explicit reliance on conventional signal processing tools.

While the preliminary system exhibits promising performance for specific instrument families, particularly Bass and Guitar, challenges remain in achieving consistent fidelity across more complex sources such as Vocals and Keys. These variations highlight the intricate nature of professional mixing, which encompasses not only technical adjustments but also interpretive decisions that reflect artistic intent, genre-specific conventions, and human auditory perception.

The limitations observed in the current framework motivate future research directions aimed at enhancing both the generalization capability and creative adaptability of neural-based mixing systems. One promising trajectory involves the development of more sophisticated end-to-end architectures, potentially incorporating attention mechanisms, adversarial learning frameworks, or transformer-based models to better capture long-range temporal dependencies and subtle spectral nuances.

Moreover, the integration of curated expert knowledge, perceptual modeling, and user-controllable parameters will be essential in bridging the gap between algorithmic processing and human creative oversight. By fostering interactive, user-in-the-loop systems, it becomes possible to combine the computational efficiency and pattern recognition abilities of neural networks with the contextual awareness, aesthetic judgment, and emotional sensibility inherent to experienced audio engineers.

Beyond technical contributions, this work aligns with a broader vision for intelligent music production tools that seamlessly merge generative composition and autonomous mixing technologies. Such hybrid systems could revolutionize contemporary audio workflows by enhancing productivity, reducing technical barriers for novice users, and enabling new forms of creative expression. Ultimately, empowering producers and engineers with tools that balance machine precision with human artistry may lead to more accessible, efficient, and innovative approaches to music creation and production.

In conclusion, this study establishes a proof-of-concept for neural-driven multitrack mixing and lays the groundwork for future advancements towards autonomous, adaptive, and artistically coherent audio production environments.

REFERENCES

[1] A. Zhang, X. Li, and Y. Wang, "A review on automatic music mixing using deep learning," *IEEE Trans. Multimedia*, vol. 26, pp. 1451–1463, 2023.

[2] R. Kumar and M. Tsai, "Neural audio mastering with dynamic range control," *IEEE Access*, vol. 11, pp. 18894–18904, 2023.

[3] L. Wu, Z. Wang, and D. Zhang, "Multi-task learning for instrument separation and mix enhancement," *Proc. IEEE ICASSP*, pp. 316–320, 2024.

[4] J. Xu and B. Li, "Audio style transfer via deep adversarial learning," *Neural Comput. Appl.*, vol. 35, no. 5, pp. 5231–5246, 2023.

[5] M. Rivera and J. Schuller, "Exploring neural networks for music production automation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 31, pp. 222–233, 2023.

[6] C. Park and H. Kim, "Vocals and accompaniment separation using Transformer-based audio encoder," *Proc. Interspeech*, pp. 1084–1088, 2024.

[7] S. Patel and G. Zhao, "Spectrogram inversion using generative neural networks," *IEEE Signal Process. Lett.*, vol. 30, pp. 220–224, 2023.

[8] H. Sun et al., "Automated mixing of multitrack audio using learned audio descriptors," *Applied Acoustics*, vol. 215, 2024.

[9] B. Wang, Y. Ma, and L. Zhang, "Instrumental source separation with contrastive self-supervised learning," *IEEE Trans. Multimedia*, vol. 25, pp. 1710–1721, 2023.

[10] T. Andersson and M. Ek, "Deep learning for real-time audio effect modeling," *J. Audio Eng. Soc.*, vol. 72, no. 1-2, pp. 24–33, 2024.

[11] R. Singh and V. Jain, "Evaluating deep autoencoders for intelligent mix generation," *Proc. IEEE MMSP*, pp. 413–418, 2023.

[12] D. Yoon and K. Choi, “Audio generation using GANs with perceptual loss,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 1776–1788, 2024.

[13] F. Lee and S. Ahmed, “Style-based remixing using content-aware neural networks,” *Proc. ISMIR*, pp. 134–140, 2023.

[14] E. Morita and N. Kawaguchi, “Speech and music separation in polyphonic audio using U-Net architecture,” *EURASIP J. Adv. Signal Process.*, vol. 2023, no. 47.

[15] A. Novak and J. Driedger, “Blind mixing parameter estimation using contrastive learning,” *Proc. DAFx*, pp. 41–48, 2024.

[16] M. Gonzalez and K. Patel, “Hybrid model for content-based audio transformation,” *Proc. IJCNN*, pp. 1256–1261, 2023.

[17] D. Kim et al., “Audio mastering network with attention-guided filtering,” *IEEE Trans. Multimedia*, vol. 26, pp. 2034–2045, 2023.

[18] Z. Hu and J. Thompson, “Dynamic audio segmentation using self-supervised learning,” *Proc. IEEE ICME*, pp. 798–803, 2024.

[19] R. Tan and Y. Hasegawa, “Music production style transfer with metric learning,” *J. Intelligent Information Systems*, vol. 62, pp. 187–198, 2023.

[20] K. Liao and M. Chen, “Neural signal path modeling for audio effects,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1140–1152, 2023.