Audio Engineering Society

Convention Paper 10232

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

A Scalable Two-Stage Automatic Mixing System Integrating Machine Learning and Domain Knowledge

Jinjie Shi[1], Kunzhu Xie[2], Yinghao Ma[1], and Joshua Reiss[1]

[1]*Centre for Digital Music, Queen Mary University of London, UK*
[2]*Wuhan University of Communication, China*

Correspondence should be addressed to Jinjie Shi (jinjie.shi@qmul.ac.uk)

ABSTRACT

Music mixing involves transforming clean, individual tracks into a cohesive final mix using audio effects and expert knowledge. While rule-based and machine learning methods have shown promise, scaling them to real-world situations remains challenging. We propose a two-stage mixing architecture that combines domain knowledge with deep learning, enabling the system to handle over 100 input tracks with high perceptual quality.

The first stage uses a rule-based level balancing system to mix grouped tracks into stems. The second stage employs a differentiable mixing style transfer model guided by a reference mix. To enhance intra-group (within subgroup) robustness, we refine loudness estimation by incorporating spectral centroid and fundamental frequency features, addressing limitations of Loudness Units relative to Full Scale (LUFS) on narrowband signals.

Subjective listening tests demonstrate that our enhanced intra-group mixing approach consistently outperforms LUFS-based baselines across multiple musical genres. Furthermore, our proposed two-step system enables deep learning to successfully handle projects with over 100 tracks for the first time, achieving mixing results that significantly surpass those of traditional rule-based systems. Code and audio examples are available at https://doi.org/10.5281/zenodo.17171082.

1 Introduction

Over the past two decades, the accessibility of music production tools has dramatically increased, enabling a growing number of musicians and creators to engage in professional-grade audio production [1]. However, achieving high-quality multitrack mixing remains a complex, skill-intensive process, often requiring years of experience [2]. To address this challenge, automatic mixing has emerged as a field of research focused on developing systems capable of autonomously balancing levels, applying effects, and shaping the overall mix.

1.1 Challenges in Automatic Mixing

Automatic mixing research has evolved through multiple phases, from early knowledge-engineering approaches to more recent machine learning models. Despite significant progress, the field still faces fundamental challenges that limit real-world applicability. In recent years, comprehensive systems have emerged in both knowledge engineering and machine learning [3, 4, 5] within this domain. It is an opportune time to revisit past advancements, identify their limitations, and explore future directions for the field.

In the context of automatic mixing, knowledge engineering refers to the explicit encoding of expert audio engineering practices into computational rules and systems. Instead of learning from data, these approaches rely on formalized representations of mixing knowledge — such as psychoacoustic principles, production guidelines, and rule-based decision trees — to automate tasks like track grouping, loudness balancing, equalization, and dynamic processing.

Knowledge-based systems, while effective in specific cases, often lack flexibility across different musical styles and production contexts [6]. Data-driven models, such as end-to-end deep learning approaches [2, 7], show improved generalization but frequently struggle with scalability. Many systems are trained on fixed configurations with limited track counts, making them less adaptable to the variable structure of real-world sessions, which often include 30 to 60 tracks. Moreover, these models can produce artifacts and lack interpretability due to limited grounding in traditional audio processing. Optimization-based approaches, such as those using genetic algorithms for gain control [8], face similar scalability and generalization issues.

A major obstacle in advancing automatic mixing is the lack of large-scale multitrack datasets containing both dry (unprocessed) and wet (processed) versions of individual tracks. Most public datasets, such as MedleyDB [9], MUSDB18 [10], and the Cambridge-MT dataset, provide isolated stems or unmixed multitracks but lack corresponding post-production multitracks. While efforts like the Open Multitrack Testbed [11] and the Mix Evaluation Dataset [12] include dry/wet pairs, they are limited in scale and no longer actively maintained. As a result, no comprehensive public dataset currently exists for supervised learning of realistic mixing scenarios, severely limiting the development of robust, data-driven mixing systems.

Scalability remains a major limitation for most deep learning–based mixing systems. For example, prior works [13, 14, 15] support only a limited number of input tracks and struggle to generalize to real-world scenarios with diverse structures and instrumentation. Martinez et al. [13] attempted to address this by testing on out-of-domain material, but performance dropped significantly.

Steinmetz et al. [4] proposed a scalable architecture using differentiable digital signal processing, weight sharing, and a sum/difference stereo loss. They extended this with Diff-MST[5], a framework combining a Transformer controller with a differentiable mixing console. Given a reference mix and raw multitracks, Diff-MST estimates per-track parameters—gain, EQ, compression, and panning—without requiring explicit source labels. Its self-attention allows it to handle arbitrary track counts, and shared MLPs ensure parameter consistency.

However, the model still struggles to generalize beyond its training data. Loudness distributions learned from four-track datasets may not apply to large-scale sessions (16–64 tracks), where diversity and complexity are much greater. For instance, having only seen mixed vocals during training, the model may fail to balance finer-grained vocal tracks, such as lead vocals and background vocals.

1.2 Our Contributions

Real-world automatic music mixing presents several key challenges:

- The number and variety of instruments in a production are highly unpredictable, requiring systems to handle an arbitrary number of input tracks.
- Due to the limited diversity of training datasets, systems must generalize to unseen musical styles and structures.

To address these challenges, we propose a two-step automatic mixing system that combines knowledge engineering with deep learning. Our key contributions are as follows:

- **Two-step mixing framework:** We decompose the mixing task into intra-group and inter-group stages. The first stage applies a knowledge-engineered system for level balancing within subgroups. The second stage employs a deep learning–based differentiable mixing style transfer model to finalize the mix.
- **Enhanced loudness computation:** We refine traditional loudness estimation by incorporating spectral centroid and fundamental frequency features, addressing limitations of ITU-R BS.1770 (LUFS) [16] in narrowband signals. This improves level balancing robustness across different musical styles.

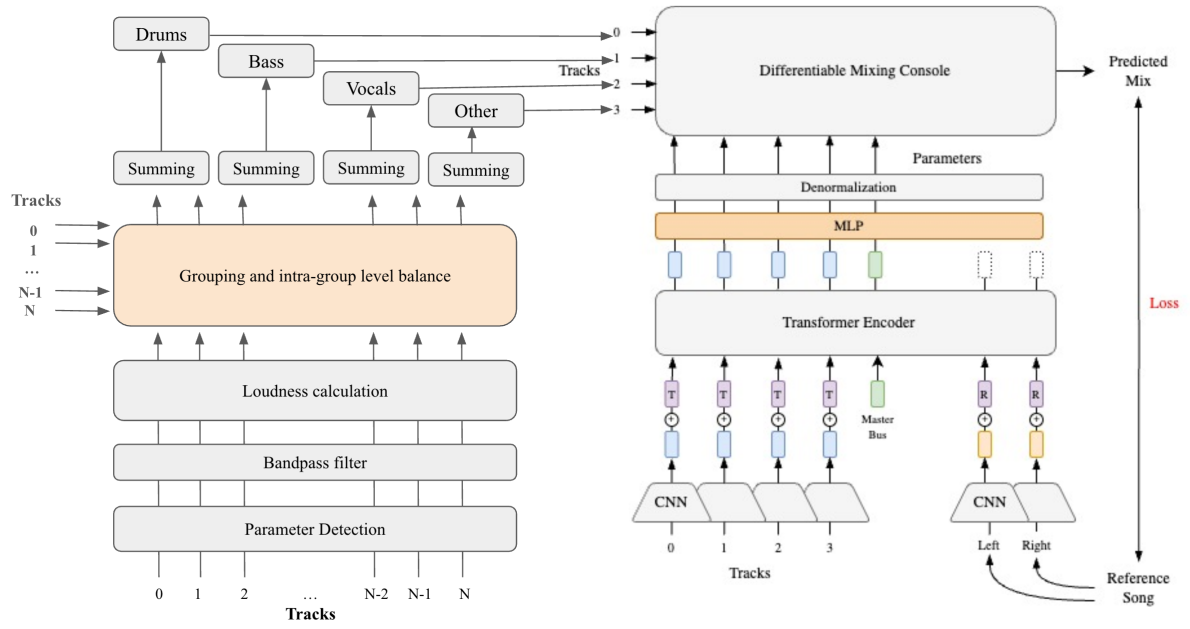

Fig. 1: Schematic diagram of the two-step automatic mixing system integrating intra-group automatic level balancing, and a differentiable mixing style transfer framework. The structure on the right is based on the Diff-MST model proposed in [5]

Subjective listening tests demonstrate that our intra-group mixing approach consistently outperforms LUFS-based baselines across multiple genres. Furthermore, our two-step system enables deep learning models to handle projects with over 100 tracks for the first time, achieving mixing results that significantly surpass traditional knowledge-engineering systems and approach the quality of mixing engineers.

2 Methods

2.1 Problem Formulation

Modern music productions typically consist of multiple individual tracks recorded, sampled, or synthesized separately. The goal of automatic mixing is to estimate appropriate processing parameters for each track to produce a cohesive and balanced final mix.

Existing approaches fall into two categories, each with limitations. Knowledge engineering methods are robust to varying input track counts but generalize poorly to unseen data. In contrast, machine learning approaches excel in constrained scenarios—such as four-stem mixes resembling their training data—but struggle to handle the diversity and complexity of real-world multitrack sessions.

A key insight from the literature is that automatic mixing can benefit from a two-stage process: intra-group and inter-group mixing. This motivates our hybrid system design, where knowledge engineering techniques provide robust intra-group level balancing, and deep learning methods model complex inter-group relationships. Specifically, we propose a two-step system combining automatic level balancing with a differentiable mixing style transfer framework, as illustrated in Figure 1.

The intra-group module, shown on the left side of Figure 1, computes Characteristic Frequency Band (CFB) loudness, a measure of track energy within perceptually relevant frequency regions, and adjusts gain for each track. At this stage, tracks are manually grouped into instrument-based subgroups, as the impact of different grouping strategies remains under investigation.

The right side of Figure 1 illustrates the differentiable mixing style transfer model, adapted from Diff-MST [5]. This component combines CNN (Convolu-

tional Neural Network)-based encoders and a Transformer controller to estimate per-stem mixing parameters. The CNNs extract spectral features from the input stems and the reference mix, which are fused by the Transformer to predict gain, equalizer, compression, and panning parameters for each subgroup. These parameters are applied through a differentiable mixing console, enabling end-to-end training and flexible style transfer aligned with the reference mix. Each system component is evaluated independently through listening tests.

Formally, the intra-group level balancing process is defined as:

$$s_j = \sum_{i \in G_j} \lambda_i \cdot a_i \tag{1}$$

where s_j is the mixed stem for the j-th subgroup, G_j is the set of tracks in subgroup j, λ_i is a time-invariant scaling coefficient, and a_i is the full audio signal of track i.

Once all subgroups are mixed into stems, the final mixing stage estimates a parameter matrix P, where each row p_j contains the processing parameters for stem s_j. Given a set of subgroup stems $S = s_1, s_2, \ldots, s_M$ and a reference mix M_r, we define:

$$P = g(f(S), f(M_r)) \tag{2}$$

where $f(\cdot)$ denotes the feature extractors for stems and the reference mix, and $g(\cdot)$ is the Transformer-based controller that predicts the audio effect parameters for each input stem.

Finally, the differentiable mixing console applies the estimated parameters P to the stems to generate the final mix M_p:

$$M_p = h(S, P) \tag{3}$$

where $h(\cdot)$ represents the differentiable mixing console that performs gain control, equalizer, compressor, and panning.

This formulation ensures that the mixing process is data-driven, interpretable, and flexible, allowing the system to generalize to different input track configurations while preserving the reference mix characteristics.

2.2 Intra-group level balance

The goal of intra-group level balancing is to establish appropriate loudness relationships among tracks within each subgroup. Setting volume levels using faders is a fundamental task in audio mixing, with a significant impact on the final musical outcome [17]. However, fully automatic intra-group mixing remains challenging due to the limitations of current machine learning systems in modeling complex inter-track relationships. At minimum, it is essential to ensure accurate intra-group level balance across all input tracks.

Existing automatic level balancing systems that meet our requirements—namely, the ability to process arbitrary numbers of input tracks without requiring reference stems—are primarily based on knowledge engineering. These systems aim to maximize inter-channel clarity, typically using using equal loudness process [18, 19]. While these approaches have shown good performance in listening tests, especially those using LUFS, several studies have highlighted LUFS limitations in multi-source scenarios [20, 21, 22]. In particular, LUFS estimates can deviate by several decibels from subjective assessments, especially for high-frequency narrowband signals.

To address these issues, some researchers have explored incorporating middle-ear transfer functions [23], but results suggest that simpler energy-based models may be preferred by users. In fact, complex psychoacoustic models do not consistently outperform LUFS-based approaches [24]. Consequently, we forgo personalized psychoacoustic models and instead propose a refined LUFS-based strategy.

Fenton [25, 26] demonstrated that manually optimizing loudness model parameters—such as filter type and integration window length—for different instrument types yields better alignment with human mixes than generic configurations like K-weighted filters or the models proposed by Pestana et al. [21]. Building on this insight, we propose an automated parameter selection scheme to improve scalability and generalization across diverse mixing scenarios.

2.3 Loudness of Characteristic Frequency Band

In multitrack mixing, inter-track masking often leads listeners to perceive only the most salient narrow frequency band from each track. We refer to this salient

Table 1: Comparison of Loudness-Grouping Relationships

Group	Track	LUFS	CFB Loudness
Lead	VOC	-19	-24
Accompany	AG PICK	-21	-27
Accompany	BV	-22	-27
Pad	EG CHO	-24	-32
Pad	EG DIST	-32	-33
Pad	PIANO	-35	-32
Pad	ORGAN	-33	-33

region as the Characteristic Frequency Band (CFB). To identify this band, we propose two methods: (1) the spectral centroid $\pm$ half the spectral bandwidth, and (2) two to four times the track's fundamental frequency.

To enhance level balancing, we apply pre-filtering based on each track's CFB before gain adjustment. For each audio track a_i in a subgroup, the gain coefficient λ_i is computed by first estimating the loudness l_i of the pre-filtered a_i. We then apply a fixed gain of $b - l_i$ to the original signal, where b is the target group loudness, defined as $-0.1 - p$ to prevent clipping, with p being the global peak value within the subgroup.

Using the CFB method, we analyzed multitrack mixes from several engineers and identified patterns that were overlooked by standard LUFS analysis (Table 1). We asked engineers to label the musical roles of tracks based solely on listening. The results show that tracks with the same role tend to exhibit similar CFB loudness, whereas their LUFS values can differ by up to 9 dB. This inconsistency may explain why earlier studies did not recognize loudness-grouping phenomena.

2.4 Automatic Level Balance System with Pre-filtering

We enhance the traditional automatic level balance framework by integrating pre-filtering based on the Characteristic Frequency Band (CFB). Using the 2–4 times fundamental frequency filtering as an example, the system first detects the fundamental frequency of each track using the PESTO algorithm [27]. It then applies a band-pass filter around 2–4 times the detected fundamental frequency and computes loudness within this band.

Loudness estimation is performed using the pyloudnorm package [28], following the ITU-R BS.1770-4 standard as defined by EBU R-128. The system then adjusts the gain of each original (unfiltered) track to align its loudness with a predefined target. Finally, the processed tracks are rendered as individual audio files and combined into a mixdown for evaluation.

2.5 Subgrouping for Automatic Level Balance

Knowledge engineering-based automatic level balancing systems typically aim to maximize inter-channel clarity through techniques such as LUFS normalization and equal-loudness contour-based gain adjustment [18, 19]. While these methods perform well in listening tests and are widely used in industry, they face limitations in multitrack scenarios. For example, Jillings and Stables [29] found that in listener-preferred mixes, lead vocals were on average 11 LU higher than the overall mix level, highlighting a systematic bias not addressed by existing methods [24].

However, multiple studies have highlighted their limitations, though solutions were not proposed at the time due to technical constraints[24]. For instance, Jillings and Stables[29] conducted a survey involving 71 participants and found that, even in attempts to create balanced mixes, vocals tended to dominate, with lead vocal levels averaging 11 LU higher than the overall mix level.

Identifying grouping strategies that align with perceptual loudness principles is critical for the practical deployment of knowledge engineering-based systems. Previous research has explored how engineers create subgroups [30], apply effects, and how subgrouping correlates with mixing preferences [31]. Some attempts at automatic subgrouping have used random forests based on audio features [32], while others have relied on simple instrument label-based grouping for loudness normalization [2]. However, the impact of different grouping strategies on automatic mixing performance—particularly for level balancing—remains underexplored.

Given the overlap between datasets for automatic mixing and source separation, we adopt the most common source separation grouping scheme—drums, bass, vocals, and other—as a starting point. If this grouping strategy proves effective, it opens the door to using source separation techniques to generate additional training material for inter-group mixing tasks.

3 Evaluation and Results

The listening test comprised two components: intra-group level balance and overall mixing. Due to track count limitations in current machine learning systems, the intra-group evaluation compared only the proposed knowledge engineering-based auto-balancing system with human mixes. The overall mixing evaluation included three systems: knowledge engineering, knowledge engineering combined with Diff-MST, and a human mix. Six multitrack sessions from the Cambridge-MT dataset were selected, covering five musical styles with 40–100 tracks per project.

3.1 Intra-group Level Balance

The intra-group level balance evaluation tested the effectiveness of our proposed *CFB loudness* and *sub-grouping* methods on real-world multitrack material. We compared our improved knowledge engineering-based level balancing approach with both a traditional LUFS-based method and a human mix. The main goal was to assess whether the proposed system could deliver stable and perceptually good results across diverse musical styles.

For each multitrack session, a graduate student in music mixing grouped tracks based on their musical roles and exported 15-second audio clips for evaluation. From the six sessions, we selected seven instrument groups with the highest number of tracks, mainly the Drums and Other groups. Each group was processed using the three different methods, and the mix engineer also provided a reference level balance for comparison.

For each audio track within the selected groups, we applied automatic level balancing with the following configurations. Apart from the loudness computation method, all other processing variables were kept constant:

- **Centroid**: Loudness was calculated after applying a pre-filter using the range defined by the spectral centroid half the spectral bandwidth.
- f_0: Loudness was calculated after applying a pre-filter defined by 2 to 4 times the average fundamental frequency of the track.
- **LUFS (Baseline)**: Standard LUFS-based loudness equalization was applied without pre-filtering.
- **Human Mix**: A graduate student in music mixing manually balanced the levels within the group.

3.2 Overall Mixing

For the overall mixing task (intra-group + inter-group), we tested the effectiveness of combining our intra-group level balancing method with Diff-MST, an inter-group mixing model based on differentiable mixing style transfer. The tested configurations are listed in Table 2. Rules 1–3 correspond to purely knowledge engineering-based approaches without any learning components, serving as baselines for the full mixing task involving more than 100 tracks. Rule 4 represents our proposed Two-Step Mixing architecture, which combines the best-performing Centroid-based intra-group balancing method with Diff-MST-based inter-group loudness adjustment. Rule 5 corresponds to the human-mixed reference.

Table 2: Comparison rules for intra-group and inter-group methods.

Rule	Intra-group	Inter-group
Rule 1 (Centroid):	Centroid	Centroid
Rule 2 (f_0):	f_0	f_0
Rule 3 (LUFS):	LUFS	LUFS
Rule 4 (Two-Step Mixing):	Centroid	Diff-MST
Rule 5 (Human Mix):	Human	Human

3.3 Testing Procedure

A total of 27 participants took part in the study, including 23 mixing engineers, as well as researchers and music enthusiasts. Since the primary aim was to evaluate participants' professional judgment in realistic mixing decisions rather than strictly comparing ratings across individuals, participants were asked to conduct the listening tests remotely in their familiar mixing environments, which primarily consisted of professional mixing studios or high-quality headphones.

The listening tests consisted of three phases: a preliminary trial, volume calibration, and the main experiment.

All clips in listening test were normalized to -30 LUFS. During calibration, participants adjusted playback to a comfortable level. The main experiment consisted of eight groups comparing intra-group automatic level balancing, and one group comparing the two-step auto mixing of the complete track.

The main experiment was conducted in a blind, MUSHRA-like format: participants did not know

which system produced each audio clip, and the order of clips was randomized. Among the eight intra-group evaluations, the fourth and seventh presented the same set of audio clips; this duplication was used to assess rating consistency.

Participants rated the clarity of each clip. To ensure a broad range of perceptual responses, participants were required to assign at least one high score (80–100) and one low score (0–20).

4 Data Analysis

4.1 Intra-group Performance Analysis

For the intra-group listening test, we included a set of repeated hidden reference items to evaluate listener consistency. Pearson correlation coefficients were calculated to screen for reliable participants, and responses with coefficients below 0.378 were excluded. To visualize the relationship between listener experience and test reliability, we plotted Pearson correlation coefficients against years of experience for all 27 participants in Figure 2. A horizontal threshold line at 0.378 indicates the cutoff used for inclusion. The majority of excluded participants had fewer years of experience, suggesting a potential correlation between listening experience and response consistency. As a result, we retained data from 17 out of 27 participants, with most of the excluded ones being less experienced mixing engineers or amateur listeners. This suggests that level balancing remains a challenging task, even for trained mixing engineers.

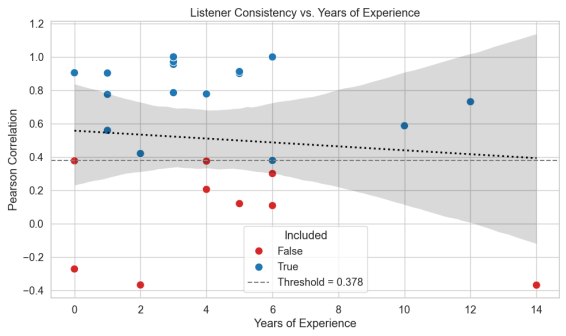

Fig. 2: Listener Consistency vs. Years of Experience

Violin plots in Figures 3 and 4 visualize the score distributions for different methods. Wider sections indicate where score values are more densely concentrated, and the central lines show medians and quartiles. This allows for an intuitive comparison of score spread and central tendency across methods.

In the drum group, the LUFS method performed comparably to the human mix, while our Centroid method significantly outperformed LUFS and slightly outperformed the human mix. This is consistent with the violin plots, where both the *Centroid* and f_0 methods show higher mean scores and a greater density of high ratings, indicated by the top-heavy shape of the plots.

In the "Other" group, both f_0 and *Centroid* methods yielded slightly higher mean scores than LUFS, with a greater proportion of high ratings. While none of the automatic systems outperformed the human mix here, the performance gap was within an acceptable range.

Overall, in both the drum and "Other" instrument groups, the *Centroid* and f_0 methods outperformed the LUFS baseline in intra-group balancing tasks.

A t-test revealed statistically significant differences between both proposed methods and the baseline. Table 3 summarizes the results of one-sided and two-sided t-tests comparing the performance of different mixing approaches (LUFS, Human Mix, f_0, and *Centroid*) in the drum and "Other" groups. Statistically significant results are marked in bold.

Table 3: T-Test Results for Intra-group Level Balance

Instrument Group	Alternative Hypothesis	p-value
Drums	LUFS $<$ Human	0.301
	Centroid $>$ Human	0.145
	$f_0 < Centroid$	0.361
	$f_0 > LUFS$	**0.028**
	Centroid $>$ LUFS	**0.047**
Other	LUFS $\neq$ Human	**0.016**
	Centroid $\neq$ Human	**0.011**
	$f_0 \neq Centroid$	0.771
	$f_0 \neq LUFS$	0.844
	LUFS $\neq$ Centroid	0.943

While the f_0-based method requires fundamental frequency estimation and incurs a higher computational cost, the *Centroid*-based method is more efficient and may be preferable in time-sensitive applications.

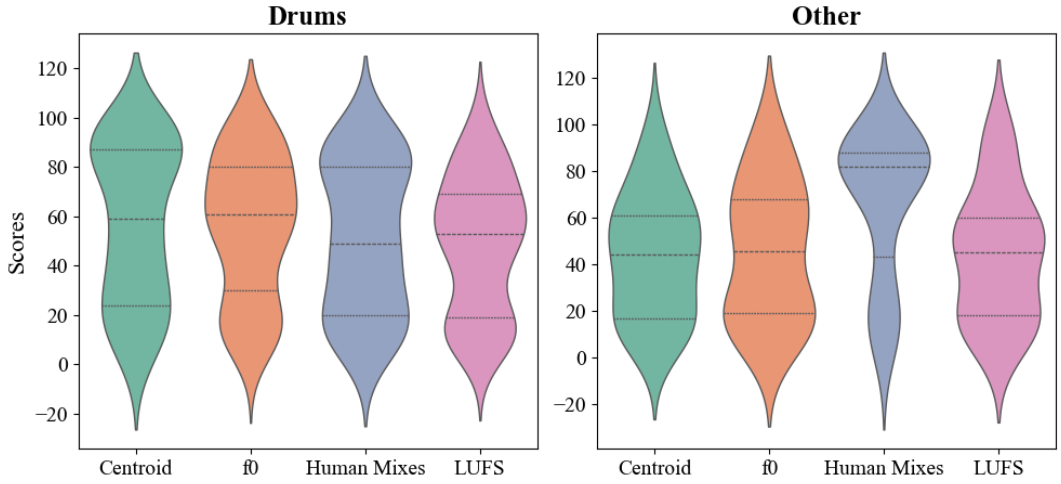

Fig. 3: Violin plot for Percussion and Other instruments

4.2 Overall Mixing Performance Analysis

Results for the complete mixing task, incorporating both intra- and inter-group balancing, are shown in Figure 4. The pure knowledge engineering system showed a large performance gap compared to both the human mixes and our proposed two-step system. Notably, the two-step method achieved mean scores up to four times higher than the LUFS-based baseline and outperformed the human mix in some cases, indicating strong effectiveness.

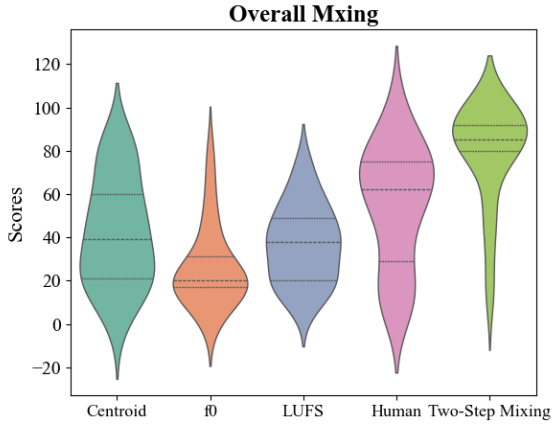

Fig. 4: Violin plot for inter-group level balance

Table 4: Summary of T-Test Results for Overall Mixing Methods

Alternative Hypothesis	p-value
$f_0 < Centroid$	**0.0059**
$f_0 < LUFS$	0.071
LUFS $< Centroid$	**0.033**
Centroid $< Human$	0.188

Table 4 shows that the *Centroid* method outperformed both f_0 and LUFS in overall mixing, and no statistically significant difference was found between *Centroid* and the human mix.

Overall, our knowledge engineering system achieved competitive results with human engineers in intra-group tasks across multiple genres and track counts (ranging from 40 to 100). However, it fell short in overall mixing, especially in inter-group balance. This performance gap was effectively addressed through our proposed machine learning–based inter-group mixing system.

The clear performance disparity between intra-group and inter-group results highlights the rationale behind our two-stage architecture. The system leverages the stability and scalability of knowledge-based methods for intra-group tasks while compensating for their lack

of generalizability in inter-group mixing through data-driven models. This hybrid design extends the capability of automatic mixing systems to handle large-scale projects without compromising mix quality.

5 Summary

Applying machine learning models to real-world mixing scenarios remains a major challenge in automatic mixing. In this work, we proposed a two-stage mixing architecture that combines domain-specific knowledge with machine learning, enabling the system to handle over 100 input tracks while maintaining robust performance across diverse musical styles.

From a practical music production standpoint, we addressed two key limitations of traditional knowledge engineering approaches: the inadequacy of LUFS in capturing the perceptual loudness of narrowband high-frequency signals, and the lack of empirical validation for instrument grouping strategies in automatic mixing.

Listening test results confirmed that our knowledge-based system performs reliably for intra-group level balancing, even with large and genre-diverse sessions. However, its performance declines in inter-group balancing tasks. Integrating a machine learning model for inter-group mixing (Diff-MST) within our two-step framework effectively overcomes this limitation.

We recommend using the *Centroid* and f_0 methods for intra-group level balancing, particularly in large sessions. For full mixing tasks, combining intra-group balancing with a learned inter-group model like Diff-MST offers a practical and scalable solution. This hybrid approach demonstrates the strength of uniting expert-driven rules with data-driven inference to achieve high-quality, consistent mixes at scale.

Acknowledgments

This work was supported by the China Scholarship Council.

References

[1] Walzer, D. A., "Independent Music Production: How Individuality, Technology and Creative Entrepreneurship Influence Contemporary Music Industry Practices," *Creative Industries Journal*, 10(1), pp. 21–39, 2017, doi:10.1080/17510694.2016.1247626.

[2] Scott, J., "Automated Multi-Track Mixing and Analysis of Instrument Mixtures," in *Proc. 22nd ACM Int. Conf. Multimedia*, pp. 651–654, Orlando, FL, USA, 2014, doi:10.1145/2647868.2654859.

[3] Man, B. D., Stables, R., and Reiss, J. D., *Intelligent Music Production*, Focal Press, Oxford, U.K., 2019.

[4] Steinmetz, C. J., Pons, J., Pascual, S., and Serrà, J., "Automatic Multitrack Mixing With A Differentiable Mixing Console of Neural Audio Effects," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 71–75, 2021, doi: 10.1109/ICASSP39728.2021.9414364.

[5] Vanka, S. S., Steinmetz, C., Rolland, J.-B., Reiss, J., and Fazekas, G., "Diff-MST: Differentiable Mixing Style Transfer," in *Proc. 25th International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, USA, 2024.

[6] De Man, B. and Reiss, J. D., "A Knowledge-Engineered Autonomous Mixing System," in *Proc. Audio Eng. Soc. Conv. 135*, 2013.

[7] Moffat, D. and Sandler, M., "Machine Learning Multitrack Gain Mixing of Drums," in *Proc. Audio Eng. Soc. Conv. 147*, 2019.

[8] Kolasinski, B., "A Framework for Automatic Mixing Using Timbral Similarity Measures and Genetic Optimization," in *Proc. Audio Eng. Soc. Conv. 124*, 2008.

[9] Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. P., "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in *Proc. Int. Soc. Music Information Retrieval Conf. (ISMIR)*, volume 14, pp. 155–160, 2014.

[10] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R., "MUSDB18-HQ - an uncompressed version of MUSDB18," 2019, doi:10.5281/zenodo.3338373.

[11] De Man, B., Morfi, V., and Reiss, J. D., "The open multitrack testbed," in *Audio Eng. Soc. Conv. 137*, Audio Engineering Society, 2014.

[12] De Man, B. and Reiss, J. D., "The mix evaluation dataset," in *Proc. 20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

[13] Martínez-Ramírez, M. A., Liao, W.-H., Fabbro, G., Uhlich, S., Nagashima, C., and Mitsufuji, Y., "Automatic Music Mixing with Deep Learning and Out-of-Domain Data," in *Proc. 23rd International Society for Music Information Retrieval Conference (ISMIR)*, pp. 411–418, Bengaluru, India, 2022, doi:10.5281/zenodo.7316688.

[14] Martínez-Ramírez, M. A., Stoller, D., and Moffat, D., "A Deep Learning Approach to Intelligent Drum Mixing with the Wave-U-Net," *Journal of the Audio Engineering Society*, 69(3), pp. 142–151, 2021, doi: 10.17743/jaes.2020.0031.

[15] Koo, J., Martínez-Ramírez, M. A., Liao, W.-H., Uhlich, S., Lee, K., and Mitsufuji, Y., "Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[16] European Broadcasting Union, "EBU R 128: Loudness Normalization and Permitted Maximum Level of Audio Signals," Technical report, Geneva, Switzerland, 2011.

[17] Bromham, G., "How Can Academic Practice Inform Mix-Craft?" in *Mixing Music*, pp. 265–276, Routledge, 2016.

[18] Perez-Gonzalez, E. and Reiss, J., "Automatic Gain and Fader Control for Live Mixing," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2009.

[19] Mansbridge, S., Finn, S., and Reiss, J. D., "Implementation and Evaluation of Autonomous Multi-Track Fader Control," in *Proc. Audio Eng. Soc. Conv. 132*, 2012.

[20] Pestana, P. D. and Barbosa, A., "Accuracy of ITU-R BS.1770 Algorithm in Evaluating Multitrack Material," in *Proc. Audio Eng. Soc. Conv. 133*, 2012.

[21] Pestana, P. D., Reiss, J. D., and Barbosa, A., "Loudness Measurement of Multitrack Audio Content Using Modifications of ITU-R BS.1770," in *Proc. Audio Eng. Soc. Conv. 134*, 2013.

[22] Ma, Z., Reiss, J. D., and Black, D. A. A., "Partial Loudness in Multitrack Mixing," in *Proc. AES 53rd Int. Conf.: Semantic Audio*, 2014.

[23] Ward, D., Reiss, J. D., and Athwal, C., "Multitrack Mixing Using a Model of Loudness and Partial Loudness," in *Proc. Audio Eng. Soc. Conv. 133*, 2012.

[24] Wichern, G., Wishnick, A., Lukin, A., and Robertson, H., "Comparison of Loudness Features for Automatic Level Adjustment in Mixing," in *Proc. Audio Eng. Soc. Conv. 139*, 2015.

[25] Fenton, S. and Lee, H., "Alternative Weighting Filters for Multi-Track Program Loudness Measurement," in *Proc. 143rd Audio Eng. Soc. Conv.*, pp. 71–80, New York, USA, 2017.

[26] Fenton, S., "Automatic mixing of multitrack material using modified loudness models," in *Audio Eng. Soc. Conv. 145*, 2018.

[27] Riou, A., Lattner, S., Hadjeres, G., and Peeters, G., "PESTO: Pitch Estimation with Self-supervised Transposition-equivariant Objective," in *Proc. 24th International Society for Music Information Retrieval Conference, ISMIR*, 2023.

[28] Steinmetz, C. J. and Reiss, J., "pyloudnorm: A Simple Yet Flexible Loudness Meter in Python," in *Proc. Audio Eng. Soc. Conv. 150*, 2021.

[29] Jillings, N. and Stables, R., "Investigating Music Production Using a Semantically Powered Digital Audio Workstation in the Browser," in *AES International Conference on Semantic Audio*, 2017.

[30] Ronan, D., De Man, B., Gunes, H., and Reiss, J. D., "The Impact of Subgrouping Practices on the Perception of Multitrack Music Mixes," in *Proc. Audio Eng. Soc. Conv. 139*, 2015.

[31] Ronan, D. M., Gunes, H., and Reiss, J. D., "Analysis of the Subgrouping Practices of Professional Mix Engineers," in *Proc. Audio Eng. Soc. Conv. 142*, 2017.

[32] Ronan, D., Gunes, H., Moffat, D., and Reiss, J. D., "Automatic Subgrouping of Multitrack Audio," in *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, Trondheim, Norway, 2015.