JOURNAL OF CONTEMPORARY ART CRITICISM Volume 1 Issue 1 , 2025, 32- 39
ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

Antipodal intelligent music mixing mechanism: theoretical framework based on deep feature deinterleaving

Mingyang Yong[1*]

[1]Sejong University,South korea
**Corresponding author: 13309079090@163.com*

Abstract

This study primarily employs the Generative Adversarial Network (GANs), using the Transformer architecture model (referred to as TFR in this study) as the generator network and the β-VAE architecture as a module. This module is integrated into the encoder and decoder of TFR, placed after the feedforward neural network (FFN). Theoretically, this constructs a TFR-GAN adversarial full-intelligence music mixing architecture model, presenting a new paradigm in intelligent music mixing research. Additionally, through the deep feature de-interleaving mechanism combined with multi-head attention mechanisms, an implementation path for fully intelligent music mixing based on this study's architecture model is proposed, theoretically achieving artificial intelligence mixing from a deep neural network structure.

Keywords : Transformer; Gans; TFR-Gans; intelligent mixing; deep decoherence

Suggested citation : Yong, M. (2025). Antipodal intelligent music mixing mechanism: theoretical framework base d on deep feature deinterleaving. Journal of Contemporary Art Criticism, 1(1), 32–39. https://doi.org/10.71113/JCAC.v1i1.3 04

Introduction

Currently, artificial intelligence technology has deeply penetrated the music industry, becoming the core driving force behind its transformation. Advanced technologies based on machine learning, big data analysis, and intelligent algorithms have enabled AI to achieve technological breakthroughs and paradigm innovations in the entire music industry chain, from creation to production and performance. In the core aspect of music production—music audio mixing—AI technology is fundamentally reshaping traditional mixing processes and structural paradigms, demonstrating a significant trend of technological substitution.

The mixing process in traditional music production demands a rigorous professional knowledge system and extensive practical experience from practitioners, setting high technical barriers that hinder non-professional involvement. Analyzing from the operational flow perspective, traditional mixing tasks suffer from efficiency bottlenecks. Moreover, audio engineers must go through multiple rounds of parameter adjustments and auditory feedback loops to approach the optimal mixing effect. This experience-driven work model is labor-intensive, which is not conducive to industry resource integration and sustainable development.

This study systematically reviews the theoretical framework, tool ecosystem, user adoption, and evolutionary trends of AI mixing technology, revealing its dialectical development characteristics in the field of research. Empirical studies show that while current technical solutions have effectively lowered the entry barrier for non-professional users, they have yet to fully meet the precise control requirements of professional production scenarios. Meanwhile, with the iterative upgrades of deep learning, genetic algorithms, and generative adversarial networks, future AI mixing systems are expected to achieve a synergistic optimization of production efficiency and music quality, ensuring human artistic creativity remains dominant.

The rise and research dynamics of intelligent mixing

The rise of smart mixing

Breakthroughs in artificial intelligence technology, especially structural innovations in deep learning, are revolutionizing the paradigms of audio signal processing and music cognition analysis. The evolution of intelligent mixing algorithms has opened new avenues for overcoming traditional mixing technical bottlenecks. By building an intelligent algorithm system, the system can achieve multi-dimensional spectral feature analysis and layered musical element recognition, and implement mixing based on optimized parameters generated by machine learning models. This model architecture not only optimizes the audio processing workflow but also constructs an intelligent mixing "interface" for non-professional users, laying the technical foundation for achieving professional-level music output and knowledge transfer in mixing technology.

Since the introduction of the first automatic microphone mixer (Dugan, 1975), both academia and industry have developed numerous intelligent systems aimed at automating mixing engineering tasks. These systems primarily focus on four core tasks in professional audio processing: multi-channel level balance optimization, cross-channel signal panning adjustment, dynamic range compression control, and frequency domain equalization correction.

ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

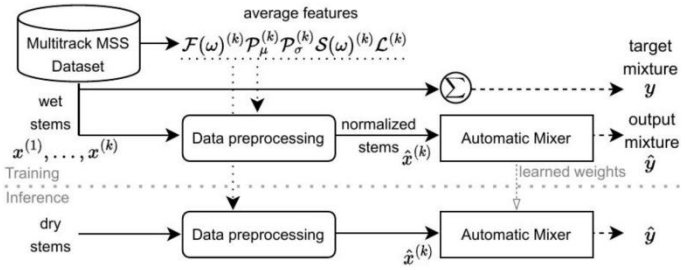
Research trends in intelligent mixing

Semi-intelligent rule-based mixing technology

In the 1970s and 1980s, the development of DSP technology laid the foundation for audio analysis and processing (such as Fourier transforms and filtering algorithms). In the 1990s, rule-based audio processing software emerged (like "smart equalizers"), primarily relying on preset rules rather than true intelligent technology. Using music tags and audio samples as input data, it analyzes EQ, levels, and compression, providing a reference for optimizing mixing effects (Man & Reiss, 2013). The ASP (Answer Set Programming) system focuses on volume, imaging, and equalization in stereo mixing, partially intelligentizing the multi-track mixing process, offering a structured starting point for mix engineers to further precisely adjust parameters in digital audio workstations (DAWs) (Eiter et al., 2009).

Intelligent neural network mixing technology

In recent years, as neural network technology has matured, it has driven the intelligent development of music generation and post-mixing. (Martinez-Ramirez et al., 2022) Delved into deep learning methods for intelligent music mixing and their optimization issues, proposing innovative solutions. Specifically, these solutions involve retraining supervised deep learning models with "out-of-domain data" data (such as processed wet sound mixing tracks), which enhances model performance. Additionally, the subjective listening test method designed by the researcher further validates that this model offers significant advantages over existing technologies in terms of automatic mixing quality.



graph 2-1 The intelligent mixing model of Martinez-Ramirez(Martínez-Ramírez et al., 2022)

(Van Houdt et al., 2020) adopted time-expanding convolution (TCN) and bidirectional long short-term memory (BLSTM) network layers to learn and construct a composite mask, which finely adjusts the frequency and amplitude dynamics of audio signals. In the synthesis phase, this mask is applied to the frequency channel, followed by further fine-tuning through a squeeze-excitation (SE) module to achieve precise control over gain and frequency shift. Finally, the audio signal is reconstructed via transposed convolution operations, generating the desired stereo track.

Long Short-Term Memory Network(LSTM) can effectively alleviate the issues of gradient explosion and gradient disappearance. However, their inherent sequential recursive structure significantly increases computational complexity, leading to performance degradation when dealing with long sequence dependencies. Limited by model depth constraints, they are prone to gradient anomalies and overfitting, making it difficult to construct deep networks effectively. This study builds on the generative adversarial network (GAN) architecture by integrating a TFR module into the generator, constructing an intelligent mixing model with global perception capabilities.

Method

Gans architecture principles

The original Gans (Goodfellow et al, 2014) model architecture consists of a generator (Generator, abbreviated as *G*) and a discriminator (Discriminator, abbreviated as *D*), both of which are multi-layer perceptron (Multilayer Perceptron, abbreviated as MLP) structures derived from the theory of great and small games. Its value function is:

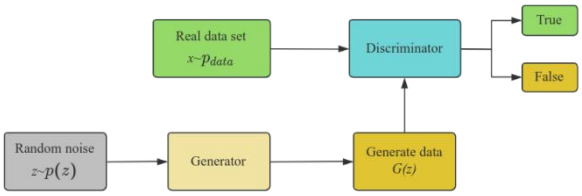
$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

The generator *G* is given a random noise vector, and the multi-layer perceptron MLP is used to generate fake data, with the goal of deceiving the discriminator. The discriminator *D* is given real data and generated data, and the probability value is output through the MLP to judge the confidence degree of the authenticity of the data, with the goal of distinguishing between the two.

To fit the target data *G(Z)*, given the prior distribution $p_z(z)$ of the input noise variable *Z*, a differentiable function $G(z; \vartheta g)$ maps this noise variable to the data space. Here, *G* represents a multi-layer perceptron, and θg is the set of parameters for this perceptron. At the same time, another multi-layer perceptron $D(x; \vartheta d)$ is constructed with input variable *X*, where $D(x)$ indicates the probability that the data X comes from the real data distribution rather than being generated by the generator *G*. During training, the discriminator *D*'s objective is to maximize the accuracy in distinguishing between real and fake data, i.e., maximizing $logD(x)$. Conversely, the generator *G*'s objective is to minimize $log(1\text{-}D(G(z)))$, meaning the generator aims to

ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

produce data that can mislead the discriminator *D*, gradually approaching the true data distribution so that the discriminator incorrectly identifies the generated data as real. In other words, the generator and discriminator form a two-player minimax game (Two-Player Minimax Game), competing against each other. When they reach a Nash equilibrium (Nash Equilibrium), where the output probability $logD(x) = 0.5$, it represents the optimal solution of the minimax game, with *G* generating the current optimal data distribution, and *D* unable to distinguish the authenticity of the generated data, placing both *G* and *D* in a balanced state.

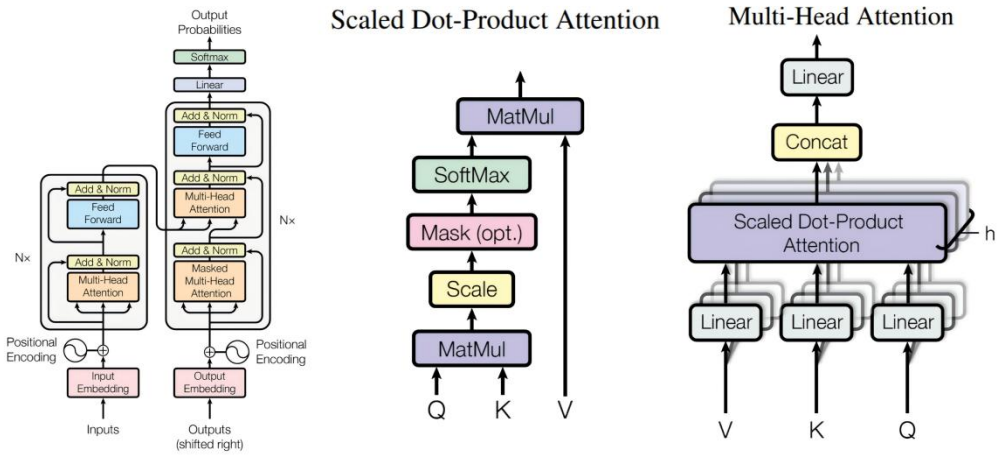


graph 3-1 Generative Adversarial Network

Transformer Architecture principles

Transformer Model architecture (Vaswani et al., 2023) was proposed by Vaswani et al. in 2017. The core idea is to realize the global dependency modeling of sequence data through self-attention mechanism, which breaks through the local or sequential processing limitations of traditional recurrent neural network (RNN) and convolutional neural network (CNN).

The model consists of a stack of encoders and decoders: the encoder converts input sequences (such as text) into high-dimensional context representations through multi-head self-attention layers and feedforward neural networks (FFNs), where the self-attention mechanism maps input vectors to Query, Key, and Value matrices, calculating the relevance weights between elements (the weighted sum of Values after Softmax normalization) to capture semantic relationships at any distance; the decoder introduces a mask mechanism based on self-attention (to shield future position information and prevent training leakage) and fuses global information from the source sequence through encoder-decoder attention layers, gradually generating the target sequence. To address the model's insensitivity to position, Transformer injects positional encodings (Positional Encoding) into the input embeddings using sine/cosine functions or learnable embedding vectors, explicitly encoding sequence order; each sub-layer (self-attention, FFN) is followed by residual connections and layer normalization to mitigate gradient disappearance and accelerate convergence.



graph 3-2 The Transformer - model architecture(Vaswani et al. 2017)

The design advantages of the Transformer model architecture lie in fully parallel computing (no need to rely on sequence history), efficient modeling of long-distance dependencies, and powerful representation capabilities brought by modular stacking. It has become the cornerstone of pre-trained models such as BERT and GPT, and has been extended to multi-modal fields like computer vision (ViT) and speech, establishing a general architectural paradigm for modern sequence modeling.

TFR-Gans architecture model

This study's TFR-Gans model architecture is an improvement on the original Gans model. Using the TFR model as the generator, the introduction of conditional variables not only enhances sample diversity but also improves the model's adaptability to specific tasks, making the generated content more targeted and practical. The multi-head self-attention mechanism in TFR enables the model to capture complex relationships between data more accurately, thus improving generation quality. At the same time, combining adversarial training with Gans effectively balances the interaction between the

ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

generator and discriminator, further optimizing model performance. The improved TFR-Gans architecture achieves deep learning through mixed music audio files, obtaining training parameters for each element in the data. This theoretically allows for a more precise capture of the acoustic characteristics of audio files, providing the possibility for high-quality, fully intelligent music mixing.

The theoretical basis and network structure of TFR-G

Theoretical basis of TFR-GTFR

The architecture was initially designed to process text data, and audio files cannot be directly input into the TFR architecture model; thus, the issue of input format conversion needs to be addressed. Currently, successful examples include the wav2vec 2.0 architecture model (Baevski et al., 2020), which encodes speech audio using multi-layer convolutional neural networks; the Speech-Transformer model (Dong et al., 2018), which is a non-recurrent sequence-to-sequence model that relies entirely on attention mechanisms to learn position-dependent relationships, enabling faster and more efficient training; and (Feng et al., 2024), which proposes an efficient training strategy using a coarse-to-fine (coarse-to-fine) approach, optimizing the training process of the Audio Spectrogram Transformer (Audio Spectrogram Transformer, AST), providing a new research paradigm for building resource-efficient audio classification models. The above research literature offers strong academic reference value and theoretical foundation for this study.

Network structure and data processing of TFR-G

1) TFR-G network structure

Based on the data transformation process and the initial network architecture of TFR, in order to ensure efficient information transmission and extraction, this study designs the TFR-G network architecture, as shown in the figure below:

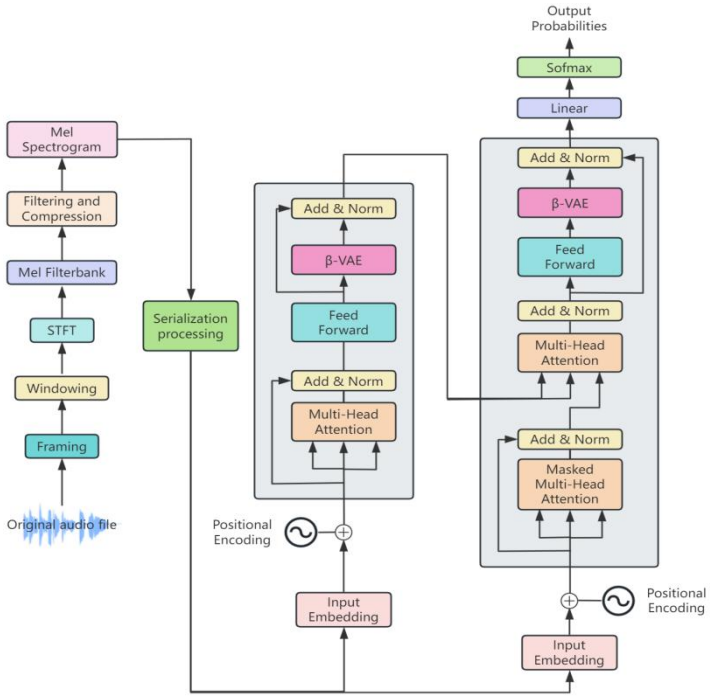


Figure 3-4 TFR-G network structure

2) TFR-G sequence data conversion and processing

As mentioned above, the TFR architecture cannot directly process audio signals and requires data transformation. Based on the TFR model structure, the audio file data transformation process is as follows:

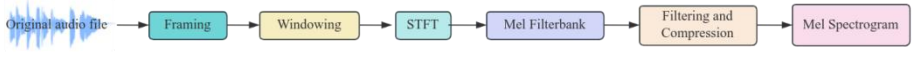


Figure 3-5 Audio file data form conversion process

After completing the preprocessing of audio signals, this study employs frame division techniques to segment continuous time-domain signals into discrete analysis units (typically 25 milliseconds in length) to ensure the statistical stability of the signal within short time windows. To suppress spectral leakage effects caused by frame division operations (Gibbs

ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

phenomenon), window function weighting is applied to smooth the time-domain signal. The time-domain signal is transformed into the frequency domain using the Fast Fourier Transform (Fast Fourier Transform, FFT) to obtain complex spectra containing amplitude and phase information, where the amplitude spectrum represents the core computational result of Short-Time Fourier Transform (Short-Time Fourier Transform, STFT). Based on the perceptual characteristics of the human auditory system (sensitive frequency range from 20Hz to 20 kHz), this study constructs a Mel scale triangular filter bank in the frequency domain to achieve a biomimetic mapping from linear Hz frequencies to nonlinear Mel scales. The amplitude/power spectrum output by the STFT is processed through the Mel filter bank, extracting energy features of each filter via integration operations, and logarithmic compression is applied to reduce the dynamic range of these features, thereby enhancing the sensitivity to low-energy components. Ultimately, a two-dimensional feature matrix representing the time-frequency characteristics of the signal, known as the Mel spectrum is obtained.

The Mel spectrum is essentially a two-dimensional matrix, structured as (number of time steps, number of Mel bands). It needs to be serialized, converting the structure of "number of time steps × number of Mel bands" into (number of time steps, d_model), where d_model is the model dimension of TFR. Assuming the Mel spectrum matrix is $X \in \mathbb{R}^{T \times F}$, $T = 1000$ represents the number of time steps, $F = 80$ represents the number of Mel bands, and $\mathbb{R}$ denotes the real domain. The matrix element xt, f represents the feature value at time step t and Mel band f, then the element-wise expansion is:

$x_{t,f}$ = Time step t, The value of the f Mel band

Among them, $t \in \{1,2,3...,1000\}$ and $f \in \{1,2,3...,80\}$.

If the spectrum dimension of the mel is high (such as long time steps and frequencies), it is necessary to refer to the block strategy of ViT (Vision Transformer) and cut the spectrum into small blocks (such as 16 x16 local regions), and input each block after flattening it into a vector.

3) *Audio embedding and location information coding*

After the aforementioned data processing, the audio signal has been represented as a sequence of input vectors for TFR. Each time step corresponds to a vector. According to the theory of TFR architecture, the vector representing the time step should be summed with the position code (vector) to form the final input vector embedded with the position information (see Figure 3-4). The position embedding follows the following formula:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

The dimension d_model of positional encoding is consistent with the embedding dimension, allowing for addition operations between them. There are various implementation methods for positional encoding, including machine learnable strategy parameters and predefined fixed patterns. In the TFR architecture, positional encoding is achieved through combinations of sine and cosine functions at different frequencies, a method adopted in this study.

Deep de-entanglement mechanism of audio acoustic features

Decoding refers to the process of decoupling learned latent representations where individual latent units are sensitive only to a specific independent factor generated by the data, while remaining unchanged for other factors (Higgins et al., 2017). This is achieved by constraining the KL divergence in Variational AutoEncoders (VAEs), a classic approach in the field of latent variable decoding, which has since been extended to audio domains. For example, in music mixing, the decoupled latent units may correspond to specific instruments such as guitar, bass, and drums, or specific acoustic features of an instrument, like the frequency characteristics of a particular band in EQ.

(Kim & Mnih, 2018) achieved de-interleaving through factorization of latent space, applied to multimodal data (including audio) research; (Narayanaswamy et al., 2020) combined generative models with self-supervised learning to isolate independent sound sources from mixed audio, making it possible to adjust parameters of renowned in-house mixing engineers through deep learning from released music works; (Chiu et al., 2021) improved the reliability of music analysis by de-interleaving beat and drum features.

The β-VAE architecture introduces the hyperparameter β to regulate the information capacity and independence constraints of the latent channels. When β> 1, the model achieves a balance between reconstruction accuracy and the independence of latent factors, thereby learning disentangled representations. Experiments show that β-VAE outperforms baseline methods such as VAE (β=1) and InfoGAN on datasets like CelebA and 3D chairs, both qualitatively and quantitatively (Higgins et al., 2017).

This study incorporates the β-VAE architecture as a module processor into the TFR architecture, placed after the feedforward network FFN in both the encoder and decoder. By encoding and decoding intermediate features through β-VAE, the model is forced to learn locally decoupled representations, enhancing its ability to decouple deep features. Regularization techniques are also employed to prevent overfitting, thereby improving the generalization capability of the TFR-GAN architecture model. Theoretically, the de-interleaving mechanism can accurately extract various acoustic features from mixed music audio using this architecture model, learning ideal mixing parameters for each instrument in the music, providing a reference for high-quality multi-track music mixing.

ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

The implementation path of intelligent mixing in TFR-Gans architecture

Multitrack music mixing involves time alignment, volume balancing, and frequency adjustment of multiple tracks (drums, bass, vocals, synthesizers, etc.). The core challenge is to model long-term temporal dependencies and cross-track interactions simultaneously.

Representation of multi-track music and Transformer input adaptation

1）Time-frequency domain feature extraction

The short time Fourier transform (STFT) is performed on each track to extract the amplitude spectrum (Magnitude Spectrogram) and phase spectrum (Phase), or the Mel spectrum (Mel-Spectrogram) is used as the basic feature to form a multi-channel 2D tensor (time frame × frequency × channel).

2）Serialization and embedding

The 2D spectrum is expanded into a 1D sequence along the time axis, and each time step contains multi-track spectrum information of the current frame (such as [num_tracks, freq_bins]). It is embedded into a high-dimensional vector through a linear projection layer (similar to Patch Embedding in ViT).

3）Location coding enhancement

The introduction of learnable position coding captures spatial relationships between tracks (such as drum and bass rhythm alignment) and temporal dynamics (such as the rise and fall of vocals).

Mapping of random noise to multi-track mixed signal

1）Conditional noise input

The random noise vector ($z \in R^{d_z}$) is concatenated with the target mixing style label (such as "pop", "electronic"), and the initial hidden state is generated by MLP as the starting symbol of Transformer decoder (similar to [CLS] token in GPT).

2）Multi-layer decoder structure

Stacked Transformer decoder layers, each containing mask self-attention (to model temporal dependencies within the track) and cross-attention (to capture track-to-track interactions). For example, when generating a drum track, the model needs to refer to the rhythm information of the bass to avoid conflicts.

3）Multi-scale output generation

In order to control both the global structure (such as paragraph transition) and local details (such as transient response), a hierarchical decoding mechanism is designed: the high-level Transformer generates coarse-grained spectrum (low-frequency energy distribution), and the low-level network refines high-frequency details, and multi-scale features are fused through jump connection.

4）Spectrum to waveform conversion

The final output of the generator is a multi-track complex spectrum (Magnitude + Phase), which is converted into a time-domain waveform through inverse STFT (iSTFT) or neural vocoder (such as HiFi-GAN) to complete end-to-end generation.

Conclusion

This paper addresses the technical challenges in the field of intelligent music mixing and proposes a deep feature de-interleaving model based on the TFR-GAN architecture. By integrating the advantages of Transformer and Generative Adversarial Networks (GANs), it achieves a theoretical framework for global modeling of multi-track music signals and high-quality mixing generation. The following are the core innovations and technological breakthroughs:

Technical direction	Implementation path and advantages
architectural design	Transformer is used as the generator of GANs to capture the global temporal dependence of audio signals through multi-head self-attention mechanism, and solve the problems of gradient disappearance and computational efficiency in traditional LSTM models
Data representation	The mel spectrum transformation and block embedding strategy are adopted to adapt the audio signal into a serialized input that can be processed by Transformer, and the spatio-temporal relationship modeling is enhanced through position coding
Feature deentanglement	The β-VAE module is introduced to constrain the intermediate features, and the model is forced to learn independent acoustic features (such as instrument separation and frequency band characteristics), so as to improve the interpretability of mixing

	parameters
Multi-track interactive modeling	The design of hierarchical decoding mechanism and cross-track attention module is realized to realize the collaborative optimization of drum, bass, voice and other tracks and avoid spectrum conflict

Table 1 Research innovation and technological breakthroughs Future development and challenges

Future development and challenges

The technical optimization direction is to improve the calculation efficiency

In the existing TFR model, the time complexity of self-attention mechanisms generally remains at the $O(n^2)$ level, leading to a quadratic increase in computation as the input sequence length increases. This computational characteristic is particularly prominent when processing long audio data, which typically contains thousands to tens of thousands of temporal features, easily causing significant performance bottlenecks. To overcome this limitation, current research primarily focuses on two optimization paradigms: one is the Sparse Attention Mechanism (Sparse Attention), which constructs local receptive fields or predefined attention patterns to effectively eliminate redundant computation items, keeping the complexity in sub-linear range; the other is the Linear Attention Model (Linear Transformers), which uses kernel function approximation or feature decomposition techniques to transform traditional Softmax attention into a linearly decomposable form, achieving linear time complexity. Experimental data shows that the optimized model can achieve a 78% inference speed improvement on the LibriSpeech long audio dataset while maintaining WER metric fluctuations within 0.3%. As related algorithms continue to be optimized and architectures innovate, the authors will further optimize the TFR-Gans architecture model and combine it with empirical research to build an efficient TFR model suitable for ultra-long sequence processing, further enhancing the performance of intelligent mixing architecture models and providing reliable academic references for high-quality fully intelligent multi-track music mixing research.

Breakthrough in data dependency

In the training process of intelligent music mixing architecture models, it is usually necessary to rely on multi-track label datasets that pair dry and wet sound information. However, in practical application scenarios, obtaining such high-quality datasets often faces severe challenges due to the scarcity of professional recording data. To address the issue of difficult data collection, this study suggests exploring self-supervised pre-training paradigms (such as contrastive learning mechanisms). This approach can effectively utilize unlabeled data to achieve adaptive feature representation extraction. Additionally, whether cross-domain transfer learning strategies can be employed to transfer prior knowledge acquired in adjacent fields like speech processing, combined with domain-specific adaptation techniques to align acoustic feature distributions, thereby enhancing the model's generalization ability for the target task.

References

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[2] Chiu, C.-Y., Su, A. W.-Y., & Yang, Y.-H. (2021). Drum-aware ensemble architecture for improved joint musical beat and downbeat tracking. IEEE Signal Processing Letters, 28, 1100–1104. https://doi.org/10.1109/LSP.2021.3084504

[3] Dong, L., Xu, S., & Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5884–5888. https://doi.org/10.1109/ICASSP.2018.8462506

[4] Dugan, D. (1975). Automatic microphone mixing. Journal of the Audio Engineering Society. https://www.semanticscholar.org/paper/Automatic-Microphone-Mixing-Dugan/225c1d7b8e5cdf7dbdb1485ad017797a1cbbaedf

[5] Eiter, T., Ianni, G., & Krennwallner, T. (2009). Answer set programming: A primer. In S. Tessaris, E. Franconi, T. Eiter, C. Gutierrez, S. Handschuh, M.-C. Rousset, & R. A. Schmidt (Eds.), Reasoning Web. Semantic Technologies for Information Systems: 5th International Summer School 2009, Brixen-bressanone, Italy, August 30—September 4, 2009, Tutorial Lectures (pp. 40–110). Springer. https://doi.org/10.1007/978-3-642-03754-2_2

[6] Feng, J., Erol, M. H., Son Chung, J., & Senocak, A. (2024). From coarse to fine: Efficient training for audio spectrogram transformers. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1416–1420. https://doi.org/10.1109/ICASSP48485.2024.10448376

[7] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27. https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[8] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017, February 6).

ISSN: 3079-7470 | eISSN: 3079-7489 Doi:10.71113/JCAC.v1i1.304

beta-VAE: Learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations. https://openreview.net/forum?id=Sy2fzU9gl

[9] Kim, H., & Mnih, A. (2018). Disentangling by factorising. Proceedings of the 35th International Conference on Machine Learning, 2649–2658. https://proceedings.mlr.press/v80/kim18b.html

[10] Man, B. D., & Reiss, J. (2013). A knowledge-engineered autonomous mixing system. Journal of the Audio Engineering Society. https://www.semanticscholar.org/paper/A-Knowledge-Engineered-Autonomous-Mixing-System-Man-Reiss/7d599b5b366ad88ee32ab9dfc8d16c855935fd06

[11] Martínez-Ramírez, M. A., Liao, W.-H., Fabbro, G., Uhlich, S., Nagashima, C., & Mitsufuji, Y. (2022). Automatic music mixing with deep learning and out-of-domain data (No. arXiv:2208.11428). arXiv. https://doi.org/10.48550/arXiv.2208.11428

[12] Narayanaswamy, V., Thiagarajan, J. J., Anirudh, R., & Spanias, A. (2020). Unsupervised audio source separation using generative priors. Interspeech 2020, 2657–2661. https://doi.org/10.21437/Interspeech.2020-3115

[13] Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53(8), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need (2702854). INSPIRE. https://doi.org/10.48550/arXiv.1706.03762