BACK TO EAR: PERCEPTUALLY DRIVEN HIGH FIDELITY MUSIC RECONSTRUCTION

Kangdi Wang, Zhiyue Wu, Dinghao Zhou, Rui Lin, Junyu Dai, Tao Jiang

ϵar-LAB, initi-AI Ltd

ABSTRACT

Variational Autoencoders (VAEs) are essential for large-scale audio tasks like diffusion-based generation. However, existing open-source models often neglect auditory perceptual aspects during training, leading to weaknesses in phase accuracy and stereophonic spatial representation. To address these challenges, we propose ϵar-VAE, an open-source music signal reconstruction model that rethinks and optimizes the VAE training paradigm. Our contributions are threefold: (i) A K-weighting perceptual filter applied prior to loss calculation to align the objective with auditory perception. (ii) Two novel phase losses: a Correlation Loss for stereo coherence, and a Phase Loss using its derivatives—Instantaneous Frequency and Group Delay—for precision. (iii) A new spectral supervision paradigm where magnitude is supervised by all four MSLR (Mid/Side/Left/Right) components, while phase is supervised only by the LR components. Experiments show ϵar-VAE at 44.1kHz substantially outperforms leading open-source models across diverse metrics, showing particular strength in reconstructing high-frequency harmonics and the spatial characteristics.

Index Terms— VAE, Music, Phase, Perceptual Weighting

1. INTRODUCTION

Achieving perfect, perceptually lossless reconstruction of complex audio signals like music remains a central challenge in audio engineering and machine learning. High-fidelity audio Variational Autoencoders (VAEs) [1] are foundational reconstructive components for many downstream tasks, which fundamentally differs from that of traditional generative VAEs like MusicVAE [2]. While the latter prioritizes the generation of semantically authentic novel content, the former aims to compress and decompress the original signal losslessly. To achieve this, the model prioritize perceptually significant details and discarding imperceptible information. This process relies heavily on psychoacoustic principles, such as utilizing perceptual weighting curves like A-weighting or K-weighting, to model the frequency-dependent sensitivity of human hearing. However, modern audio VAE models fail to integrate such fine-grained perceptual weighting strategies into their training paradigms.

Furthermore, the reconstruction of high-quality music requires the accurate modelling of both phase and spatial information. Spatial information, often parameterized by the Mid/Side (M/S) decomposition, is critical for accurately rendering the stereo image. Concurrently, audio transients and clarity are determined not by the absolute phase of Short-time Fourier Transform (STFT) bins, but by their partial derivatives: Instantaneous Frequency (IF) across time and Group Delay (GD) across frequency. However, existing open-source models lack effective mechanisms to supervise these critical phase derivatives and fail to fully leverage the M/S representation for spatial reconstruction, which leads to audible artifacts, such as transient smearing and an inaccurate stereo image, limiting their use in professional applications.

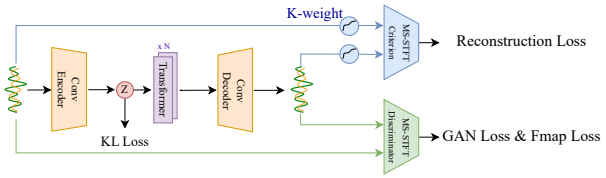

Fig. 1. Architecture of ϵar-VAE

To address these shortcomings, we introduce ϵar-VAE, an open-source VAE model optimized for high-fidelity music reconstruction. Our model incorporates a K-weighting perceptual filter, which we demonstrate is psychoacoustically better suited for music signals than A-weighting. To ensure phase coherence, we propose novel loss functions that implicitly optimize the phase performance by supervising its derivatives (IF & GD). Additionally, we apply the reconstruction loss with a new Mid/Side/Left/Right (MSLR) weighting scheme to maximize the preservation of both spatial and spectral details. Through these targeted designs, ϵar-VAE achieves state-of-the-art reconstruction performance across multiple objective evaluations, setting a new benchmark for open-source high-fidelity audio VAEs.

We summarize our contributions as follows:

- We analyse and integrate K-weighting filter into the VAE training pipeline, aligning the reconstruction objective with psychoacoustics of music perception, in contrast to the commonly A-weighting.
- We propose novel phase-aware loss functions that supervise phase derivatives to implicitly model critical phase differences, thereby enhancing transient clarity and phase coherence.
- We introduce a novel supervision strategy that separately constrains magnitude and phase, employing all MSLR components for magnitude reconstruction while using only LR to ensure phase coherence.

2. RELATED WORK

While traditional generative VAEs[3] utilize a Kullback-Leibler (KL) divergence loss to enforce a continuous Gaussian prior for generation, the reconstruction task prioritizes the model's ability to faithfully compress and restore signals, which have shifted towards discrete quantized representations. This approach, pioneered by VQ-VAE [4] and now standard in neural audio codecs like EnCodec [5] and DAC [6], excels at achieving high compression ratios. To enhance the perceptual quality of the decoded audio, these frameworks often incorporate a powerful adversarial component, leveraging discriminators from vocoders like MelGAN [7] and HiFi-GAN [8].

Despite their success, the inherent information loss from quantizing remains a fundamental limitation, as subtle details crucial for reconstruction can be discarded at the bottleneck.

In contrast, the approach revisiting continuous latent representations offers a potentially higher-fidelity pathway for reconstruction. The VAE model from Stable-Audio-Open (SAO) [1] stands as a prominent example of this approach, employing a VAE-GAN framework with adversarial loss and a down-weighted KL divergence to learn a continuous representation at a high compression rate. This model provides a powerful, open-source baseline for the community.

3. ϵar-VAE

Our model, inspired by the VAE architecture of SAO, is a partially convolutional VAE complemented by transformer-based bottleneck layers, trained with a composite adversarial objective. As shown in figure 1, the overall architecture consists of a traditional VAE generator and a powerful time-frequency domain discriminator. The generator encodes the input waveform into a latent representation and then decodes it back into a waveform, while the discriminator distinguishes audios between real and reconstructed version, thereby guiding the generator to produce higher-fidelity output.

3.1. Generator

Our generator employs an encoder-decoder architecture featuring several key designs optimized for music reconstruction. The encoder utilizes a series of strided convolutional blocks with the SnakeBeta activation function [9], which outperforms alternatives such as ELU in our experiment. The decoder is designed asymmetrically: it mirrors the encoder's convolutional structure using transposed convolutions for upsampling but also incorporates a powerful transformer module with RoPE position embeddings [10] on the decoding path. This asymmetric design delegates local feature extraction to the efficient encoder, while the decoder's transformers model global dependencies, yielding superior performance over symmetric architectures like Mimi [11]. We selected transposed convolutions rather than upsampling-plus-convolution because the former preserves greater signal energy and perceptual loudness, which is more significant than slightly higher high-frequency clarity offered by the latter. Finally, all convolutional layers are weight-normalized [1] for training stability.

3.2. Discriminator

For adversarial training, we employ a Multi-Resolution STFT Discriminator (MR-STFTD) as MSD, inspired by Encodec [5]. This approach assesses the signal across various STFT resolutions, enabling it to detect a wide range of artifacts from coarse spectral errors to fine-grained phase inconsistencies. Notably, we omit the Multi-Period Discriminator (MPD). Our experiments show that MPD introduces spatial positioning artifacts in the stereo field, which we attribute to its fixed periodic analysis being ill-suited for the complex, inconstant rhythms of music. In contrast, a single MSD provides robust supervision without introducing such spatial distortions.

While some works like BigVGAN ([9]) apply the complementary MSD in Constant-Q Transform (CQT) representation, we find this approach degrades the VAE's ability to represent music. As shown in figure 2, CQT overemphasizes low-to-mid frequency melodic features, leading to the lack of high-frequency harmonics, whereas the STFT provides a more balanced and suitable frequency response for our model.

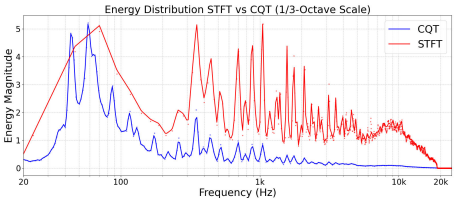

Fig. 2. Energy-based frequency response comparison between CQT and STFT of a musical excerpt.

3.3. Loss functions

Multi-Scale Log-Magnitude Loss To guide the generator's spectral amplitude reconstruction, we adopt the multiscale STFT loss formulation from EnCodec [5]. This loss, denoted as $\mathcal{L}_{\text{stft}-\text{mag}}$, computes the L1 distance between the logarithmic magnitudes of the predicted and target spectrograms over a set of different STFT resolutions, effectively capturing both coarse harmonic structures and fine temporal details.

Feature-Map Loss We also employ the feature-matching loss from EnCodec [5] to enhance perceptual quality. This loss, $\mathcal{L}_{\text{fmap}}$, minimizes the L1 distance between the intermediate feature maps extracted from the discriminator's layers for the ground-truth and generated audio.

Adversarial Loss We employ the least-squares GAN objective from BigVGAN [9] for adversarial training, defining the generator loss $\mathcal{L}_{\text{Adv}}(G)$ and the discriminator loss $\mathcal{L}_{\text{Adv}}(D)$.

KL Loss We regularize the latent space using a Kullback-Leibler (KL) divergence loss, which aligns the posterior distribution $q(z|x)$ with a standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to promote a continuous and well-structured representation.

Correlation Loss Inspired by music production metrics, our *Correlation Loss* directly penalizes phase deviations between the ground truth spectrogram $\mathbf{S}$ and its reconstruction $\hat{\mathbf{S}}$ is defined as:

$$\mathcal{L}_{\text{corr}} = 1 - \sum \text{Re}\left(\frac{\hat{S}S}{\left|\hat{S}\right| |S| + \varepsilon}\right), \tag{1}$$

where the term within the summation normalizes the cross-power spectrum, simplifying to the cosine of the phase difference between the signals, $\cos(\phi_{\hat{S}} - \phi_S)$. Minimizing this loss thus encourages perfect phase coherence.

Phase Loss Phase instability often introduces "electrical buzz" artifacts. To mitigate this, our *Phase Loss* constrains the phase's first-order partial derivatives: Instantaneous Frequency (IF) and Group Delay (GD). These derivatives are computed via finite differences on the phase $\phi = \arg(S)$ and $\hat{\phi} = \arg(\hat{S})$:

$$\begin{cases} \text{IF}(\phi)_t = \phi_{t+1} - \phi_t, \\ \text{GD}(\phi)_f = -(\phi_{f+1} - \phi_f), \end{cases} \tag{2}$$

where the subscripts t and f denote the time and frequency dimensions, respectively. All phase differences are computed modulo 2π to resolve the discontinuity at the $\pm\pi$ boundary. The total loss penalizes the $L1$ norm between the ground-truth and estimated derivatives, to optimize phase coherence in a more stable and perceptually relevant manner than direct phase supervision.

K-Weighting Curve K-weighting, originating from ITU-R BS.1770 [12] loudness measurement standards, which is widely applied in music production, is designed to approximate the frequency-dependent sensitivity of the human ear in 3. This cascaded filter

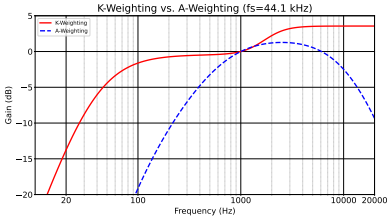

Fig. 3. K-weighting vs. A-weighting Curve

accentuates mid and high frequency bands where human hearing is most sensitive, while attenuating lower frequencies. By pre-filtering the signal with $H_{\mathrm{K}}(z)$, we ensure that reconstruction losses are evaluated in a perceptually relevant domain.

M-S-L-R Split Inspired from the music mixing process and SAO [1], we also apply the stereophonic split in two different representations, Left-Right (S_L, S_R) and Mid-Side (S_M, S_S), where mid is defined as $(S_L + S_R)/2$, side $(S_L - S_R)/2$. For different supervision attributes, we take certain combinations of the generator and the discriminator separately:

Total Loss The complete training objective consists of two parts: one for the generator $\mathcal{L}_G$ and one for the discriminator $\mathcal{L}_D$. These are optimized in alternating steps:

$$\begin{cases} \mathcal{L}_G = \lambda_{\mathrm{stft-mag}}\mathcal{L}_{\mathrm{stft-mag}} + \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{phase}}\mathcal{L}_{\mathrm{phase}} \\ \qquad + \lambda_{\mathrm{fmap}}\mathcal{L}_{\mathrm{fmap}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{Adv}}(G;D) + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}, \\ \mathcal{L}_D = \mathcal{L}_{\mathrm{Adv}}(D;G), \end{cases} \tag{3}$$

where $\lambda_{\{\cdot\}}$ are hyperparameters that control the relative importance of each loss component.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

Our model is trained on a combination of large-scale public datasets and a high-quality, proprietary in-house dataset. The training process is conducted in two stages: pre-training and continue-training. First, we use a diverse mix of public data including FSD50K [13], FMA [14], and DISCO-10M [15]. Subsequently, for the continue-training stage, the model is trained on our in-house dataset of approximately 10,000 hours of professionally produced music.

4.2. Data Pipeline

To ensure data quality, we designed a two-stage filtering pipeline, applied progressively. All datasets undergo Stage 1 for format standardization, while only our in-house data is subjected to the full pipeline.

Stage 1: Format and Loudness Standardization All the files are formatted to 44.1 kHz stereo, with files natively sampled below this rate being discarded. Next, we filter based on perceived loudness. Using the LUFS-I metric [16], we retain only tracks with an integrated loudness between -22 and -5 LUFS, removing acoustically extreme examples.

Stage 2: True Peak Filtering For our in-house data, we filter true peak levels less than +1dB to handle signal clipping. Different from the any-clip-rejection strategy from Encodec [6], this lenient criterion is deliberately chosen to accommodate the moderate, intentional clipping common in modern music mastering.

4.3. Training Details

The ϵar-VAE achieves a 1024x compression rate via a 5-layer convolutional encoder/decoder structure with strides $[2, 4, 4, 4, 8]$ and a 128-dimensional latent space. The decoder is augmented with two transformer layers using RoPE. The multiscale discriminator uses STFT window sizes of $[2048, 1024, 512, 256, 128]$. The full model contains 141M parameters.

We train all models on 8 A100 GPUs using the AdamW optimizer with a learning rate of 3×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. The loss weights are set as follows: $\lambda_{\mathrm{stft-mag}} = 50$, $\lambda_{\mathrm{corr}} = 10$, $\lambda_{\mathrm{phase}} = 10$, $\lambda_{\mathrm{fmap}} = 20$, $\lambda_{\mathrm{adv}} = 1$, and $\lambda_{\mathrm{KL}} = 10^{-6}$ for reconstruction.

We train the model for a total of 2M steps over three distinct phases:

Phase 1: Warm-up (10k steps) We train only the generator using the STFT and KL losses, accompanied by a linear learning rate warm-up.

Phase 2: Pre-train (1M steps) On the public datasets, we activate all loss components and train the generator and discriminator alternately. The learning rate is halved at 200k, 400k, and 600k steps.

Phase 3: Continue-train (1M steps) We continue training on our in-house dataset using the same full loss configuration and alternate training scheme as in Phase 2, with a constant learning rate.

4.4. Results

4.4.1. Novel Evaluation Metrics

To evaluate phase accuracy, we introduce two custom metrics, Individual Channel Phase Coherence and Cross Channel Phase Coherence.

Individual Channel Phase Coherence (ICPC) ICPC quantifies the stability of phase errors within each channel. It is derived by first computing the phase error $\Delta\phi(f,t)$ at each time-frequency bin, with magnitude-based weighting applied to mitigate the influence of noise. For each time frame, a coherence score $\mathcal{C}_t$ is calculated as the mean resultant length of these weighted phase error phasors. The final ICPC score is the energy-weighted average of these per-frame scores, where the energy E_t for each frame is the sum of its corresponding weights.

$$\mathrm{ICPC} = \frac{\sum_t \mathcal{C}_t \cdot E_t}{\sum_t E_t + \epsilon}. \tag{4}$$

Cross Channel Phase Coherence (CCPC) CCPC extends this concept to stereo signals by measuring the preservation of the Inter-channel Phase Difference (IPD). Its calculation follows the same principle as ICPC, but is based on the error in the IPD instead of the single-channel phase error. The final score is similarly the energy-weighted average across time, where the frame energy $E_{\mathrm{inter},t}$ is the sum of the corresponding inter-channel weights.

$$\mathrm{CCPC} = \frac{\sum_t \mathcal{C}_{\mathrm{inter},t} \cdot E_{\mathrm{inter},t}}{\sum_t E_{\mathrm{inter},t} + \epsilon}. \tag{5}$$

4.4.2. Evaluation Setting

We evaluate performance using several objective metrics, primarily sourced from the auraloss library [17]: Multi-Scale STFT (MS-STFT) distance, Multi-Scale Mel (MS-Mel) distance, and SI-SDR. The multiscale metrics were configured with FFT sizes of $[4096, 2048, 1024, 512, 256, 128]$ and a hop size of one-quarter the window size. Additionally, we measure the True Peak loudness

Table 1. Results on MuChin and In-house validation split (side-by-side comparison).

Model	Channels/ Rate (Hz)	Latent Rate	MuChin						In-house validation					
			Mel dist↓	STFT dist↓	ICPC↑	CCPC↑	SI-SDR↑	dbTP dist↓	Mel dist↓	STFT dist↓	ICPC↑	CCPC↑	SI-SDR↑	dbTP dist↓
DAC	1/44.1k	86Hz	0.71	1.33	94.25%	90.69%	6.14	0.29	0.67	1.21	94.49%	90.47%	6.68	0.35
Encodec	2/48k	50Hz	0.84	1.57	89.89%	89.34%	3.64	0.10	0.80	1.49	90.07%	89.42%	3.99	0.14
AGC	2/48k	100Hz	0.71	1.46	94.70%	94.96%	7.58	0.27	0.65	1.39	94.16%	94.66%	8.21	0.09
SAO	2/44.1k	21.5Hz	0.75	1.64	90.70%	91.41%	4.62	0.29	0.64	1.34	90.37%	91.12%	5.23	0.36
ϵar-VAE	2/44.1k	43Hz	**0.55**	**1.17**	**96.78%**	**96.81%**	**9.99**	**0.05**	**0.55**	**1.12**	**96.66%**	**96.52%**	**11.00**	**0.05**

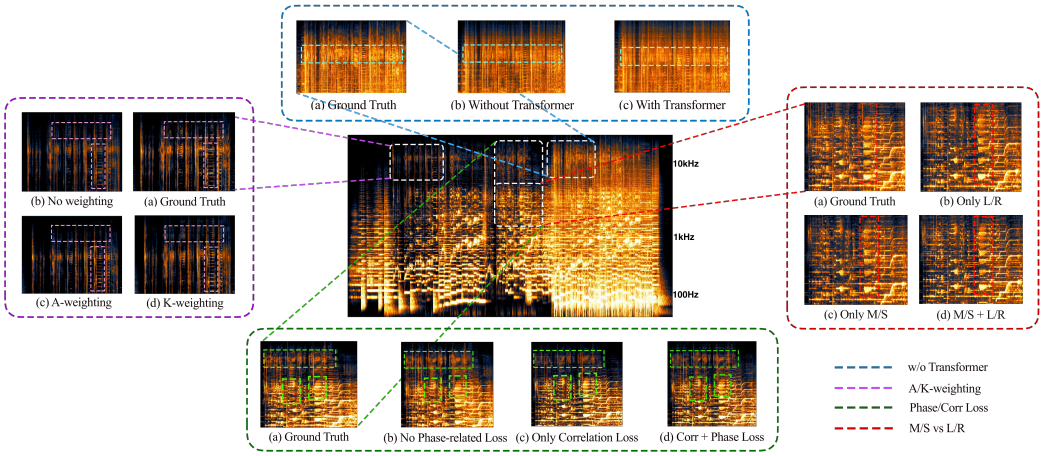

Fig. 4. All Ablation Study

difference (dbTP) using FFmpeg [18] and the designed metrics in 4.4.1.

We compare ϵar-VAE against several leading audio reconstruction models: EnCodec, DAC, AudioGen (AGC), and Stable-Audio-Open (SAO). The evaluation is performed on the reconstruction of test sets from the MuChin [19] and our in-house validation datasets. Detailed results are presented in table 1.

4.5. Ablation Study

Our design choices are validated through a series of ablation studies, with qualitative results visualized in figure 4, which compares reconstructed spectrograms from various model configurations against the ground truth.

Impact of Architectural Components The top panel of 4 highlights the role of the transformer layers. Without them, the model fails to reconstruct fine-grained harmonic structures above 10 kHz, confirming that the transformer's self-attention is crucial for modelling long-range frequency dependencies, complementing the convolutional layers' feature extraction.

Impact of Perceptual and Phase-related Losses The left and bottom panels demonstrate the effects of our proposed loss functions. As shown in the left panel, removing the K-weighting pre-filter results in a suboptimal reconstruction, particularly in the critical mid-to-high frequency bands, while the A-weighting curve improperly attenuates high frequencies. The bottom panel illustrates that removing the phase-related losses leads to a loss of clarity and the introduction of audible "current-like" noise. Specifically, the Phase Loss ensures local phase smoothness, while the Correlation Loss contributes to enhances spectral coherence, particularly for polyphonic elements.

Impact of Stereo and Spectral Representation Our ablation results reveal a key principle in stereo supervision. For magnitude, supervision over all four components—left, right, mid, and side—provides a more complete guidance signal for stereo reconstruction. In contrast, phase supervision should be restricted to the pure left/right mode. Incorporating mid/side components into phase losses distorts the physically meaningful Inter-aural Phase Difference (IPD) cues, thereby introducing spatial artifacts.

5. CONCLUSION

In this paper, we present ϵar-VAE, a variational autoencoder that sets a new state-of-the-art for high-fidelity music reconstruction by successfully integrating a structured VAE objective with a powerful adversarial framework. Our ablation studies confirm that the carefully designed components, including novel perceptual and phase-based losses, contribute significantly to its superior performance. While ϵar-VAE provides a robust foundation for reconstruction, its acoustically-focused latent space reflects the limitations for generative downstream tasks, such as mix-style conditioned synthesis. Furthermore, addressing the model's tendency to attenuate subtle spatial effects, such as reverb, by exploring targeted loss functions remains a promising research direction. We believe that tackling these challenges will unlock the next generation of controllable and realistic generative music models, solidifying the role of VAEs in creative audio applications.

6. REFERENCES

[1] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, "Stable audio open," 2024.

[2] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A hierarchical latent vector model for learning long-term structure in music," 2019.

[3] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," 2018.

[5] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," 2022.

[6] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," 2023.

[7] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[8] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[9] Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, "Bigvgan: A universal neural vocoder with large-scale training," 2023.

[10] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu, "Roformer: Enhanced transformer with rotary position embedding," 2023.

[11] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," 2024.

[12] International Telecommunication Union, "Recommendation ITU-R BS.1770-5: Algorithms to measure audio programme loudness and true-peak audio level," Tech. Rep., International Telecommunication Union, 2023.

[13] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[14] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "FMA: A dataset for music analysis," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[15] Luca A. Lanzendörfer, Florian Grötschla, Emil Funke, and Roger Wattenhofer, "Disco-10m: A large-scale music dataset," 2023.

[16] European Broadcasting Union, "EBU R 128: Loudness normalisation and permitted maximum level of audio signals," Recommendation, EBU, Geneva, Switzerland, 2023, Provides practical guidelines for implementing loudness normalization based on ITU-R BS.1770.

[17] Christian J. Steinmetz and Joshua D. Reiss, "auraloss: Audio focused loss functions in PyTorch," in *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020.

[18] Suramya Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, pp. 10, 2006.

[19] Zihao Wang, Shuyu Li, Tao Zhang, Qi Wang, Pengfei Yu, Jinyang Luo, Yan Liu, Ming Xi, and Kejun Zhang, "Muchin: A chinese colloquial description benchmark for evaluating language models in the field of music," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8 2024, pp. 7771–7779.