BNMusic: Blending Environmental Noises into Personalized Music

Anonymous Author(s)
Affiliation
Address
email

Abstract

While being disturbed by environmental noises, the acoustic masking technique is a conventional way to reduce the annoyance in audio engineering that seeks to cover up the noises with other dominant yet less intrusive sounds. However, misalignment between the dominant sound and the noise—such as mismatched downbeats—often requires an excessive volume increase to achieve effective masking. Motivated by recent advances in cross-modal generation, in this work, we introduce an alternative method to acoustic masking, aiming to reduce the noticeability of environmental noises by blending them into personalized music generated based on user-provided text prompts. Following the paradigm of music generation using mel-spectrogram representations, we propose a ***B**lending **N**oises into Personalized **M**usic (BNMusic)* framework with two key stages. The first stage synthesizes a complete piece of music in a mel-spectrogram representation that encapsulates the musical essence of the noise. In the second stage, we adaptively amplify the generated music segment to further reduce noise perception and enhance the blending effectiveness, while preserving auditory quality. Our experiments with comprehensive evaluations on MusicBench, EPIC-SOUNDS, and ESC-50 demonstrate the effectiveness of our framework, highlighting the ability to blend environmental noise with rhythmically aligned, adaptively amplified, and enjoyable music segments, minimizing the noticeability of the noise, thereby improving overall acoustic experiences.

1 Introduction

In public environments like subway trains, passengers are often exposed to persistent noise. While active noise cancellation (ANC) [5] is effective for personal use, it's not practical in group settings. Equipping everyone with ANC headphones is unrealistic, and ANC systems struggle with high-frequency noise. We propose a new approach: instead of eliminating noise for individuals, we aim to blend the environmental noise with designed music in a way that reduces its perceptual salience for a group. This shift from suppression to harmonious masking enables scalable auditory enhancement in shared environments, improving comfort without requiring personal devices. This approach could also be applied in other noise-prone settings, such as elevators or household appliances, where blending the noise with aligned music could improve the auditory experience.

Inspired by psychoacoustic principles of audio masking, where one sound reduces the perception of another, our task focuses on introducing background sounds to mask unwanted noise. Traditionally, white noise or unrelated music has been used for this purpose, but these methods often fall short. White noise struggles with non-stationary sounds and, when amplified, can become irritating. Similarly, unrelated music requires high volume to mask noise effectively, which can cause discomfort. Instead of full masking, our approach aims for perceptual blending—generating music rhythmically and spectrally aligned with the noise. This partial masking reduces annoyance without overwhelming the listener, integrating residual noise components into the music.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Effective blending requires the generated music to mask the noise perceptually, even at low volumes. To achieve this, the music must align with the noise in terms of rhythm and structure, integrating naturally without relying on high loudness. While recent advances in music generation using mel-spectrogram representations [2, 12, 8, 6, 14, 15, 16, 7] show progress, most models are trained on clean, structured inputs and struggle with noisy, unstructured data. We build on this emerging paradigm by generating music in the frequency domain, aligning its structure and rhythmic patterns with the surrounding noise for seamless auditory blending, thereby reducing the listener's awareness of the noise.

We propose a novel method, ***B**lending **N**oises into Personalized **M**usic (BNMusic)*, which uses adaptive loudness-amplified music to blend with ambient noise. The approach consists of two interconnected stages. In Stage 1, we apply a two-step outpainting and inpainting process on the noise mel-spectrogram to generate music that aligns rhythmically and spectrally with the noise's high-energy regions. By conditioning on these regions, the model generates music that effectively targets the perceptually prominent components of the noise. In Stage 2, we adaptively amplify the music's loudness, leveraging the alignment from Stage 1. Since the music already inherits the frequency distribution of the noise's high-energy region, only modest amplification is needed to achieve effective masking, without excessively increasing the overall volume. This two-stage process works together to mask the most salient noise components, which significantly reduces the noise's perceptual presence, leaving the less intrusive low-energy components with minimal impact on the listener's experience. Thus, our design ensures perceptually effective blending with minimal gain, preserving musical coherence while suppressing unwanted noise.

To evaluate our approach, we conducted objective and subjective assessments using EPIC-SOUNDS [3] and ESC-50 [11] noise sources, covering a wide range of environmental sounds. Our method outperformed other baselines on MusicBench [8]. In summary, we introduce a novel task of *noise blending with music*, propose a method to construct and adaptively amplify music to blend with noise, and demonstrate through experiments that our method effectively minimizes noise perception while preserving auditory coherence.

2 *BNMusic* framework: Blending Noises into personalized Music

In this section, we present the details of our BNMusic framework, designed to blend noise A_{Noise} with adaptive amplified music A_{Music} generated from A_{Noise} and text condition C_{text}. Our method extends the existing model's application without additional training.

Problem statement. We formalize noise blending as an alternative to traditional masking, which often relies on high volume. Given a noise segment A_{Noise} and a user prompt C_{text}, the goal is to generate a music segment A_{Music} that, when played with the noise, reduces its perceptual salience and integrates residual components into the musical texture, enhancing the overall auditory experience. As shown in Fig. 1, the masked noise exhibits a regular rhythm, enabling alignment with music through our two-stage BNMusic framework.

Pre-processing. The input noise $A_{\text{Noise}} \in \mathbb{R}^{t \times f_s}$ is first converted to a mel-spectrogram $\mathbf{S}_{\text{Noise}} = \text{Mel}|\text{STFT}(A_{\text{Noise}})| \in \mathbb{R}^{W \times H}$, where STFT is the Short-Time Fourier Transform and Mel denotes mel-filtering. This transforms the 1D audio signal into a 2D representation, which is then mapped to grayscale pixel intensities $\mathbf{x}_{\text{Noise}} \in [0, 255]^{W \times H \times 1}$, with lower values indicating louder regions. To highlight high-energy areas, a binary mask $\mathbf{M} \in \{0, 1\}^{W \times H}$ is applied, producing the masked spectrogram $\widetilde{\mathbf{x}}_{\text{Noise}} = \mathbf{x}_{\text{Noise}} \odot \mathbf{M}$.

Stage 1: Noise-aligned music synthesis. We use a two-step outpainting and inpainting process to preserve the rhythmic essence of the input noise while blending it into music. The mask $\mathbf{M}$ isolates the core noise region in the image $\widetilde{\mathbf{x}}_{\text{Noise}}$, dividing it into two parts. During the outpainting stage, the model generates music to fill the space surrounding the core, allowing the noise to diffuse outward. The masked mel-spectrogram $\widetilde{\mathbf{x}}_{\text{Noise}}$ and the text prompt C_{text} are encoded into latent representations and passed through the modified LDM model from Riffusion [2], which is fine-tuned for music generation. Given the noisy latent $\mathbf{z}_t$ at timestep t, the model predicts the added noise $\boldsymbol{\epsilon}_\theta$ using a U-Net conditioned on the corrupted mel-spectrogram $\widetilde{\mathbf{x}}_{\text{Noise}}$ and the prompt C_{text}. The posterior distribution of the previous latent state $\mathbf{z}_{t-1}$ is computed as:

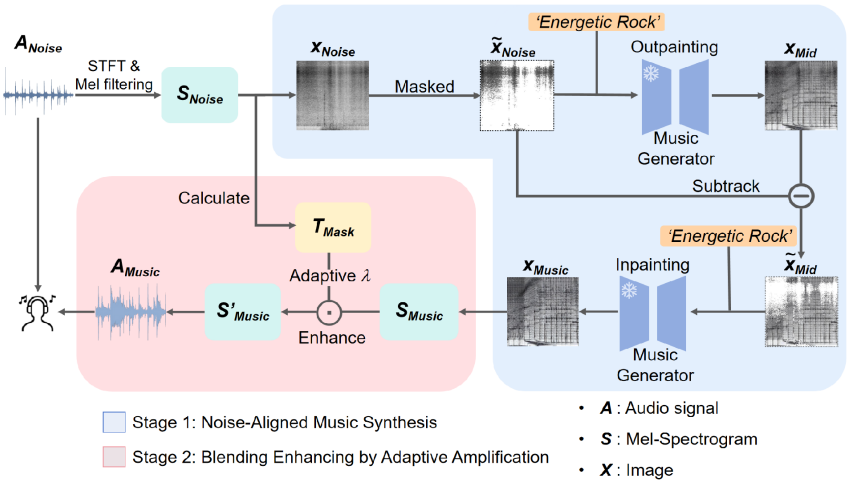

Figure 1: **Overall pipeline of our proposed BNMusic framework to achieve noise blending with frozen music generators.** The two stages of our approach are marked with different background colors. In Stage 1, our approach generates music that aligns with the noise, and in Stage 2 we adaptive amplify the music signal to reach the most ideal and reasonable blending with the noise.

$$p(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \widetilde{\mathbf{x}}_{\text{Noise}}, C_{\text{text}}) = \mathcal{N}\left(\mathbf{z}_{t-1}; \mu(\mathbf{z}_t, \boldsymbol{\epsilon}_\theta), \sigma_t^2 \mathbf{I}\right)$$

The reverse process proceeds until $t = 0$, yielding the final latent $\hat{\mathbf{z}}_0$. This latent is passed through the decoder D to reconstruct the mel-spectrogram, including the masked region $\mathbf{x}_{\text{Mid}} = D(\hat{\mathbf{z}}_0)$. The outpainted region of $\mathbf{x}_{\text{Mid}}$ aligns rhythmic patterns with the surrounding noise. We invert the mask $\mathbf{M}$ and inpaint the higher-energy components of the noise: $\widetilde{\mathbf{x}}_{\text{Mid}} = \mathbf{x}_{\text{Mid}} \odot (\mathbf{1} - \mathbf{M})$.

After a second inpainting, we obtain the final musical content $\mathbf{x}_{\text{Music}}$, integrating rhythmic patterns while eliminating distractions.

Stage 2: Blending enhancement by adaptive amplification. To improve blending, we amplify the generated music to enhance its masking effect on the core noise region. First, we compute the noise's spectrogram $\hat{\mathbf{S}}_{\text{Noise}}$ using the Short-Time Fourier Transform (STFT). Based on prior research [10], we derive the threshold matrix:

$$\mathbf{T}_{\text{Mask}} = \text{Mel}|10^{\frac{20 \cdot \log_{10}(\hat{\mathbf{S}}_{\text{Noise}}) + 21}{20}}|$$

The minimum signal-to-mask ratio (SMR) of 21 dB is used to set the masking thresholds. The amplification factor λ is optimized to maximize auditory masking while maintaining acceptable music loudness. The optimization function is:

$$\lambda^* = \arg\min_{\lambda} \{\text{SUM}(\alpha \cdot \mathbf{S}'_{\text{Music}}) + \text{SUM}(\max[(\mathbf{T}_{\text{Mask}} - \mathbf{S}'_{\text{Music}}) \odot \mathbf{M}, \mathbf{0}])\}$$

This ensures the music amplifies the core area's masking while minimizing global amplification. The amplified mel-spectrogram $\mathbf{S}'_{\text{Music}}$ is converted back to audio:

$$A_{\text{Music}} = \text{ISTFT}(\text{Griffin-Lim}(\text{Mel}^{-1}(\mathbf{S}'_{\text{Music}})))$$

3 Experiment

3.1 Experiment setup

Dataset. Our dataset comprises noise clips, real music clips, and text prompts. We source 1,000 segments from **EPIC-SOUNDS**[3] (58 human actions, 140 objects) and 300 from **ESC-50**[11] (50

Table 1: **Subjective and objective evaluation results.** We report subjective scores for overall quality (OVL) and perceived noise level (PER) on adaptively amplified samples, along with objective metrics (FAD and KL) evaluated on both adaptive amplified and direct output audio samples. BNMusic achieves the highest subjective scores and consistently superior objective metrics across both settings.

Methods	Subjective Metrics		Objective Metrics			
	Adaptive Amplified		Adaptive Amplified		Direct Outputs	
	OVL↑	PER↑	FAD↓	KL↓	FAD↓	KL↓
Random Music	2.93 ± 0.58	2.63 ± 0.53	**6.84**	2.07	15.41	2.38
MusicGen [1]	2.97 ± 0.34	2.68 ± 0.54	7.08	1.75	10.95	1.85
Riff A2A [2]	2.95 ± 0.60	3.24 ± 0.67	12.82	2.33	13.15	2.25
BNMusic (Ours)	$\mathbf{3.67 \pm 0.55}$	$\mathbf{3.84 \pm 0.63}$	7.98	**1.67**	**7.98**	**1.67**

real-world sounds), covering frequencies from 100 Hz to 10,000 Hz. Music data includes 5,000 five-second clips from **MusicBench**[8] across diverse genres and instrumentation. Additionally, we create 100 text prompts across seven genres (*Pop, EDM, Rock, Hip-hop, Punk, Jazz, Classical*) via ChatGPT[9]. Pairing noise clips with multiple prompts, we generate 14,200 music clips using **Riffusion**[2] and **MusicGen**[1], e.g., 1,000 EPIC-SOUNDS clips × 5 prompts = 5,000 clips; 300 ESC-50 clips × 7 prompts = 2,100 clips.

Baselines. We compare with three baselines: (1) Riffusion audio-to-audio generation, (2) MusicGen melody-conditioned generation, and (3) randomly selected real music from MusicBench. All music clips are overlaid with corresponding noise to simulate realistic auditory conditions for both objective and subjective evaluation.

Implementation details. To ensure pleasant blending without excessive volume, we normalize loudness using Pyln-norm [13] with ITU-R BS.1770-4, setting noise to -18 dB LUFS. Riffusion runs with default settings, processing each sample in 5 seconds on an Nvidia 4090 GPU; preprocessing and amplification take 0.28 seconds. Adaptive amplification uses an overall control parameter $\alpha = 0.14$. Evaluation overlays music and noise, with half of the real music clips paired with noise and the other half serving as ground truth.

3.2 Evaluation

Objective evaluation. We use Fréchet Audio Distance (FAD)[4] and Kullback-Leibler (KL) Divergence to measure similarity between generated and reference audio, with lower values indicating better matching. FAD evaluates feature distributions across batches, while KL is computed pairwise. Scores are calculated on combined noise-music audio against real music, with and without loudness normalization. As shown in Tab. 1, BNMusic achieves the lowest FAD and KL, indicating effective blending and alignment with noise structure. Our direct outputs perform best overall, while Random Music benefits from louder amplification.

Subjective evaluation. We conduct human evaluations on 50 samples, each with five clips: original noise, BNMusic output, and three amplified baselines. Clips are mixed with noise, and listeners rate **OVL** (overall quality) and **PER** (perceived noise) on a 1–5 Likert scale. Results (Tab. 1) show BNMusic provides the most pleasant listening experience and best noise masking, outperforming Riffusion, MusicGen, and real music. Riffusion ranks second, suppressing noise but reducing musicality, while MusicGen and real music offer limited masking.

4 Conclusion and discussion

In conclusion, our *BNMusic* demonstrates superior performance in blending music with environmental noise compared to other methods, effectively reducing the annoyance of the noise while enhancing the overall auditory experience. Through a series of experiments and ablation studies, we show the effectiveness of our approach, as well as the contribution of each key modeling component. By finding an optimal balance between maximizing the pleasantness of the music, controlling its loudness, and aligning it with the noise for more seamless blending, our method ensures that the combined sound provides the best listening environment.

References

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024.

[2] Seth* Forsgren and Hayk* Martiros. Riffusion - Stable diffusion for real-time music generation. 2022.

[3] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[4] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

[5] Sen M Kuo and Dennis R Morgan. Active noise control: A tutorial review. *Proceedings of the IEEE*, 87 (6):943–973, 1999.

[6] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

[7] Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using DDPM inversion. In *Proceedings of the 41st International Conference on Machine Learning*, pages 34603–34629. PMLR, 2024.

[8] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation, 2023.

[9] OpenAI. Chatgpt: Language model for dialogue, 2024. Accessed: 2024-05-21.

[10] T. Painter and A. Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451–515, 2000.

[11] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, 2015.

[12] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion, 2023.

[13] Christian J. Steinmetz and Joshua D. Reiss. pyloudnorm: A simple yet flexible loudness meter in python. In *150th AES Convention*, 2021.

[14] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. Audit: Audio editing by following instructions with latent diffusion models, 2023.

[15] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703, 2024.

[16] Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models, 2024.