KANUNET-ECA: A NOVEL APPROACH FOR DENOSING CONCERT MUSIC RECORDINGS

Shijie Zhang

School of Computer Science and Technology, Fudan University, Shanghai, China

ABSTRACT

During concerts, people often use their phones to spontaneously record memorable moments, but these recordings are frequently accompanied by noise such as cheering and applause, which diminishes the playback experience. In this paper, we introduce a novel task specifically designed for denoising music in concert environments, a challenge that has been largely overlooked in previous research. To support this task, we have created a new concert denoising dataset, which includes songs performed in various major languages at concerts, accompanied by noise segments such as cheering and applause. Building on this, we propose KANUnet with ECA, a method that combines the traditional UNet network with the recently proposed KAN network to process the time-frequency representation of spectrograms and remove background noises. Extensive experiments demonstrate that our method not only outperforms previous models in denoising performance but also attempts to restore disrupted musical structures during the denoising process.

Index Terms— music denoising, deep learning,music reconstruction

1. INTRODUCTION

With the widespread use of smartphones and economic development, concerts have become a popular way for people to relieve work stress. Audience members often use their phones to record exciting moments of the concerts anytime and anywhere. However, recordings made in the audience area often include cheers and applause, affecting the quality of the music.

Audio denoising and enhancement are crucial tasks in the field of signal processing [1], with a long history of development[2, 3, 4, 5, 6]. However, most of these methods are primarily aimed at speech denoising and dereverberation [7, 8]. Speech denoising and enhancement techniques are designed for the relatively stable frequencies of speech signals, unlike the complex spectra and dynamic variations of music. Focusing on the current developments in music denoising, Gfeller et al. [9] utilize a CNN neural network based on STFT spectral representations to denoise historical music recordings, creating a training dataset by artificially mixing noise from silent segments of old records with clean music. Moliner and Välimäki [10] introduce a deep neural network architecture featuring two layers of U-Net, processing audio's spectral representations and training with real noise data to simultaneously remove noises from old analog recordings. While these models [9, 10] effectively filter out common electronic noises from historical recordings, most people have less exposure to classical music. Instead, they often record popular music, such as live concert performances. Therefore, we propose a concert music denoising task.

Currently, several music datasets [11, 12] have been proposed for tasks such as music transcription and source separation, among other music information retrieval tasks. For example, the MUSDB18 [11] dataset contains 150 English songs and four sources for each song, primarily used for source separation tasks. However, the sources of these songs are not necessarily from concerts, which is not entirely suitable for our task. To accomplish the concert denoising task, we have collected recordings of songs performed by singers from various mainstream languages at concerts worldwide. We have extensively collected noise segments such as applause and cheering sounds from concerts to simulate the realistic background noise of a concert. Using these data, we created pairs of clean music and noisy music for subsequent training.Additionally,we have also collected recordings from live concerts for use in subsequent testing.

U-Net [13] was originally developed for medical image segmentation tasks and is now widely applied to music related tasks [14, 15, 16]. Wave-U-Net [14] integrates the U-Net architecture to operate directly in the time domain, allowing it to effectively handle the complexities of music information and capture both local and contextual information. Recently, compared to traditional MLP [4] networks, Kolmogorov-Arnold Networks (KANs) [17]have gained attention due to their ability to learn non-linearities, and it makes sense to effectively leverage KAN to bridge the gap between the network's physical attributes and empirical performance. Int this paper, we aim to integrate the advanced KAN network into the backbone of U-Net through a convolutional KAN mixed architectural style to accomplish music denosing and ehancement. We employ a multilayered deep encoder-decoder architecture with skip connections.The main contributions of this paper are as follows:

1. We introduce a concert music denoising task, motivated by the fact that people often record live concert perfor-

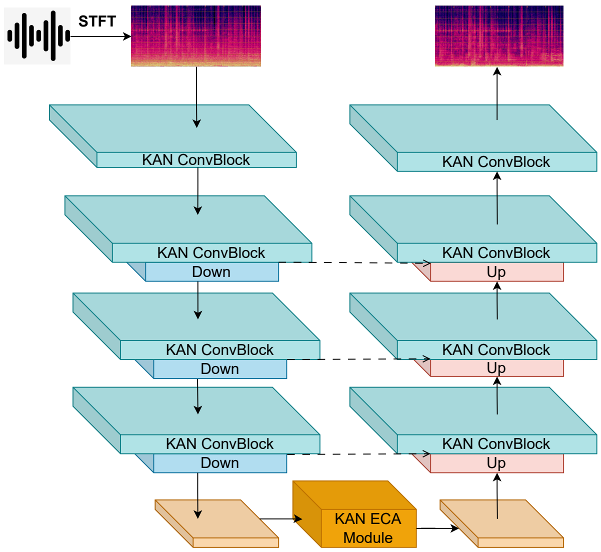


Fig. 1: The framework of the KANUnet with ECA system.

mances, which contain significant background noise.

2. We collect and construct a multi-language, multi-style music dataset, which includes a large number of real concert recordings and their corresponding noise segments.

3. We propose KANUnet with ECA, a KAN-Unet backbone specifically designed for denoising concert music.

2. METHODS

Fig.1 outlines the overall process of the model. The initial step involves applying the Short-Time Fourier Transform (STFT) to the noisy audio signal: $y = \text{STFT}(x)$. For a given waveform x, the STFT operation transforms the waveform into a complex spectrogram $y \in Y^{C\times F\times T}$, where F and T represent the frequency and time dimensions, respectively, The model takes y as its input and is composed of three main components: an encoder, a bottleneck with Efficient Channel Attention (ECA), and a decoder. The encoder reduces the spatial dimensions of the input spectrogram while increasing feature depth through downsampling KAN ConvBlocks. The bottleneck applies channel-wise attention via a KAN ECA module to enhance key features and suppress noise. The decoder restores the original spatial dimensions by upsampling and combining feature maps with skip connections for accurate reconstruction. The output is a denoised spectrogram, which can be converted back to the time domain using iSTFT.

2.1. KANConv Block

In an attempt to address the challenges of poor parameter efficiency and reduced interpretability that are intrinsic to MLP, Liu et al. [17] proposed the Kolmogorov-Arnold Network (KAN), eliminating the reliance on linear weight matrices by employing learnable activation functions on the edges and parameterized activation functions as weights. This design significantly reduces the complexity and the number of parameters required for precise modeling and KAN can achieve comparable or superior performance with smaller model sizes. Additionally, the structure enhances model interpretability without compromising performance, making them suitable for a wide range of applications. A K-layer KAN can be characterized as a nesting of multiple KAN layers:

$$KAN(\mathbf{Z}) = (\Phi_{K-1} \circ \Phi_{K-2} \circ \cdots \circ \Phi_1 \circ \Phi_0)\mathbf{Z} \qquad (1)$$

Where the i-th layer of the entire KAN network is denoted by Φ_i.

The KANConvBlock (Fig.2a) comprises three key components: a KAN convolutional layer, a batch normalization (BN) layer, and a ReLU activation function.. KAN ConvBlock is chosen to enhance feature extraction while maintaining computational efficiency. This KANConvBlock structure can be repeated across L layers. The output of each KAN convolutional block for a spectrogram representation y can be expressed as:

$$\mathbf{y}_\ell = \text{Relu}\left(\text{Batchnorm}\left(\text{KANConv}(\mathbf{y}_{\ell-1})\right)\right) \qquad (2)$$

where ℓ represents the ℓ-th KANConv layer.

2.2. KANUnet with ECA

We use a structure similar to the U-Net [13]. For downsampling, we combine max pooling with a KAN Convolution (KANConv) block. Unlike regular convolution, KANConv[18] is designed to capture more complex patterns by leveraging the Kolmogorov-Arnold [17] representation, which allows for better feature extraction with fewer parameters. This results in more efficient learning and improved model performance. The KAN ECA Module (Fig.2b) starts with a KAN ConvBlock to extract music features, followed by an Efficient Channel Attention(ECA) [19] block to enhance channel attention, enabling effective noise reduction while preserving the musical signal. In the upsampling path, we use bilinear interpolation and then concatenate the upsampled feature maps with the corresponding ones from the downsampling path. These skip connections preserve important spatial information, ensuring that fine details are retained.

2.3. Loss Function

We utilize a weighted sum of three losses: the magnitude loss, the complex spectrogram loss, and the multi-resolution loss.

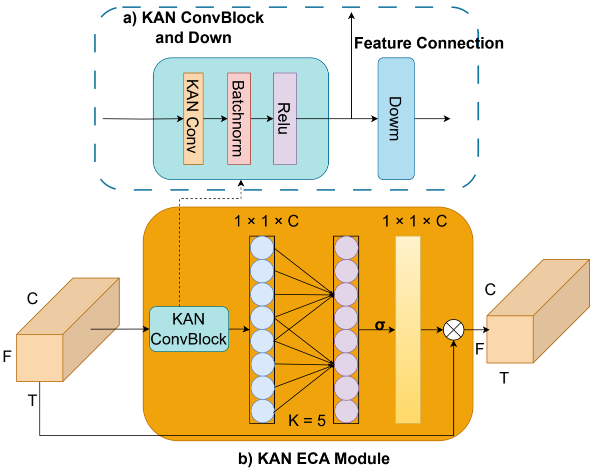


Fig. 2: (a) The overall of KAN ConvBlock and Down (b) KAN ECA Module

The magnitude loss, denoted as $\mathcal{L}_{\text{mag}}$, is defined as the Mean Absolute Error (MAE) between the estimated magnitude $\hat{M}$ and the clean magnitude M. The complex spectrogram loss, $\mathcal{L}_{\text{ri}}$, is the MAE between the estimated complex spectrogram $\hat{S}$ and the clean complex spectrogram S:

$$\mathcal{L}_{\text{mag}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{M}_i - M_i\right\|, \quad \mathcal{L}_{\text{ri}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{S}_i - S_i\right\| \tag{3}$$

Additionally, we compute the multi-resolution complex spectrogram MAE [20] between the reference wave x and the reconstructed wave $\hat{x}$, using $S = 5$ multi-resolution STFTs with window sizes of [4096, 2048, 1024, 512, 256] and a fixed hop size of 147:

$$\mathcal{L}_{\text{multi-resolution}} = \frac{1}{N}\sum_{i=1}^{N}\|x_i - \hat{x}_i\| + \sum_{s=0}^{S-1}\frac{1}{N_s}\sum_{i=1}^{N_s}\left\|Y_i^{(s)} - \hat{Y}_i^{(s)}\right\| \tag{4}$$

where $Y^{(s)}$ and $\hat{Y}^{(s)}$ represent the complex spectrograms of x and $\hat{x}$ at different resolutions.

The final weighted loss function is expressed as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{mag}} + \beta\mathcal{L}_{\text{ri}} + \gamma\mathcal{L}_{\text{multi-resolution}} \tag{5}$$

where α, β, and γ are the weights for the magnitude loss, complex spectrogram loss, and multi-resolution loss, respectively.

3. DATASET

3.1. Clean Concert Music

We develop an extensive dataset designed to replicate the authentic auditory experience of live concert environments. This dataset consists of high-quality recordings of concert music across multiple major languages. Specifically, we curate 1,000 songs each in Chinese and English, alongside over 500 songs in other widely spoken languages. The selection process for these recordings is meticulous, ensuring that the songs represent a wide variety of genres and performance styles.

Each track in this dataset comes with detailed metadata, including not only the song title but also information such as the artist, genre, and recording conditions where available. This metadata enriches the dataset, making it suitable for a variety of tasks beyond music denoising, such as music identification, genre classification, and even content-based music recommendation systems. The inclusion of clean, unadulterated concert music recordings provides a crucial benchmark for evaluating the performance of noise reduction algorithms, particularly in scenarios where high fidelity is paramount.

3.2. Noisy Concert Music

To create a realistic environment for music denoising research, we collect a large number of noise samples, specifically targeting background noise commonly found in live concert settings. These noise samples include various types of applause, cheering, and crowd noise, sourced from various online platforms. The total duration of these noise recordings is approximately eight hours, providing a diverse array of acoustic textures to simulate different crowd dynamics and venue acoustics.

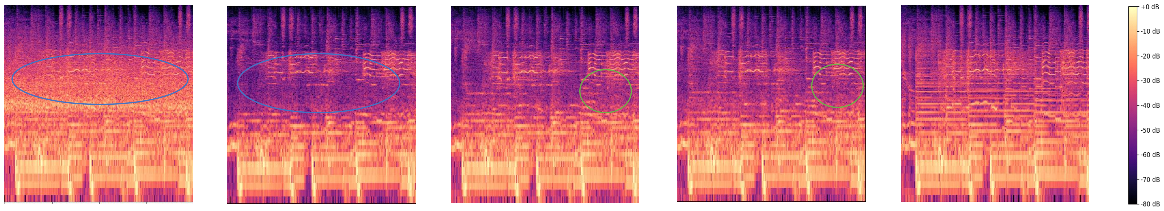
The clean music dataset is systematically segmented into 10-second clips, a length chosen to balance computational efficiency with the ability to capture significant musical content. Each clean clip is then randomly paired with a noise recording segment, with the duration of the noise segments adjusted to match that of the music clips. The noise levels are also modified according to different crowd intensities, thereby simulating the dramatic fluctuations in noise levels typical of concert scenarios.

The resulting noisy clips will be used for training within a self-supervised learning Model. In this approach, the model learns to separate the clean music signal from the noisy input by exploiting the inherent structure of the audio data. Specifically, the self-supervised model is trained to reconstruct the clean version of the music from the noisy input, using the paired clean and noisy clips as ground truth.

4. EXPERIMENTAL SETUP

4.1. Model and training configurations

In this training, we used 10-second noisy and clean audio clips as training pairs. During each iteration, a 2-second segment is randomly selected from the 10-second clip and fed into the model. To enhance the model's robustness and generalization, we applied data augmentation to the noisy audio,



(a) noisy **(b) Unet** **(c) KANUnet** **(d) KANUnet(ECA)** **(e) clean**

Fig. 3: Comparison of spectrograms for different methods.

with a signal-to-noise ratio (SNR) range of [-15, 5] dB to simulate various noise intensities in concert environments.The model is total trained for 50 epochs with the Adam optimizer with an initial learning rate of 5e-4 and batch size of 2 for each GPU.

The STFT window size is 46ms (1024-point FFT) with a hop size of 11.6ms. Thus, the spectrogram will have 513 frequency bins. We use a window function of the same size as the FFT, which enhances the frequency resolution. Regarding the model architecture, the feature dimension is first increased to 64 through the encoder, followed by three downsampling layers, where the feature dimensions are progressively increased to 128, 256, and 512. Each downsampling block contains two convolutional layers.

4.2. Evaluation metrics

In this experiment, we employed Signal-to-Noise Ratio (SNR)[21] and Signal Distortion Ratio (SDR)[22] as objective metrics for evaluating the effectiveness of music denoising. Additionally, we utilized Perceptual Evaluation of Audio Quality (PEAQ)[23] as a supplementary evaluation criterion.

Table 1: SNR of music denoising

Model	H_noise	M_noise	L_noise
Unet [13]	3.86	5.36	6.98
DCCRN [24]	3.11	4.74	6.31
KANUnet	4.45	6.07	7.47
KANUnet(ECA)	**4.64**	**6.30**	**8.04**

Table 2: SDR of music denoising

Model	H_noise	M_noise	L_noise
Unet[13]	3.20	4.74	6.50
DCCRN[24]	2.42	4.08	5.92
KANUnet	3.81	5.54	7.22
KANUnet(ECA)	**3.99**	**5.74**	**7.62**

5. RESULTS

5.1. Objective Evaluation

We selected 10-second segments from 200 songs, randomly shuffled them, and added three different levels of noise: high-level noise (H noise), mid-level noise (M noise), and low-level noise (L noise). We then evaluated these segments under different noise levels using SNR, SDR, and PEAQ metrics. Tables 1 and 2 present the SNR and SDR performance of different models in the music denoising task. The results show that the KANUnet and KANUnet(ECA) models outperform the Unet [13] model across all noise levels. Particularly under low noise conditions, the KANUnet(ECA) model achieves the highest SNR and SDR values, indicating stronger noise reduction and signal fidelity.

In the spectrograms(Fig.3), the noise reduction performance in the mid-frequency region is clearly observed across different models. The original noisy signal (a) shows a concentration of noise with high energy in the mid-frequency range. The U-Net model (b) reduces some noise, but significant noise artifacts remain in this area. The KANUnet (c) offers improved noise suppression, especially in the mid-frequency region, though faint noise remains. The KANUnet (ECA), with Efficient Channel Attention, further reduces this noise, producing a spectrogram closely matching the clean reference (e).

6. CONCLUSION

In this paper, we introduce a new task of concert music denoising and have collected a dataset consisting of concert songs and noise segments. We also propose a model named KANUnet with ECA for music denoising. Experimental results demonstrate that our model outperforms previous models on the concert denoising dataset we presented. In the future, we will further refine our music denoising methods to better reconstruct the musical structures that are disrupted during the denoising process.

7. REFERENCES

[1] Samuel D Stearns, "Of aldapfive signal processing," 1985.

[2] Michael Berouti, Richard Schwartz, and John Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1979, vol. 4, pp. 208–211.

[3] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[5] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[8] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al., "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.

[9] Beat Gfeller, Dominik Roblek, Marco Tagliasacchi, and Pen Li, "Learning to denoise historical music," in *ISMIR 2020-21st International Society for Music Information Retrieval Conference*, 2020.

[10] Eloi Moliner and Vesa Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 841–845.

[11] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The musdb18 corpus for music separation," 2017.

[12] Vincent Lostanlen and Carmine-Emanuele Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," *arXiv preprint arXiv:1605.06644*, 2016.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[14] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[15] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, pp. 1667, 2019.

[16] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, pp. 2154, 2020.

[17] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[18] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau, "Convolutional kolmogorov-arnold networks," *arXiv preprint arXiv:2406.13155*, 2024.

[19] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11534–11542.

[20] Enric Gusó, Jordi Pons, Santiago Pascual, and Joan Serrà, "On loss functions and evaluation metrics for music source separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 306–310.

[21] Don H Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, pp. 2088, 2006.

[22] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and

Catherine Colomes, “Peaq-the itu standard for objective measurement of perceived audio quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[24] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, “Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement,” *arXiv preprint arXiv:2008.00264*, 2020.