Audio Denoising

Introduction

Audio denoising is an important task in signal processing that involves removing unwanted noise from audio signals while preserving desired information. There are several traditional methods for audio denoising, including the use of Wiener, median and adaptive filters. These methods rely on statistical properties of the noise and the signal to estimate and remove the noise from the audio signal.

In recent years, deep learning-based approaches have shown promising results in audio denoising tasks. In this project, we implemented a supervised learning approach for audio denoising using a U-Net architecture[1]. The U-Net architecture is a convolutional neural network that has been widely used in image segmentation tasks but has also demonstrated strong performance in audio signal processing tasks.

Other state-of-the-art methods for audio denoising include generative adversarial networks (GANs), autoencoders, and deep neural networks (DNNs), STFT-SEANet[3] and least-squares autoregressive interpolation (LSA-C)[2]. These methods leverage the power of deep learning to effectively remove noise from audio signals. However, our proposed approach, which utilizes a U-Net architecture with a supervised learning approach, offers a distinct advantage over these other methods, as it is able to provide efficient and effective denoising performance.

Overall, this project explores the potential of using a supervised learning approach with a U-Net architecture for audio denoising and provides subjective and objective evaluation for the denosing effect. By leveraging the power of deep learning, our proposed approach offers a promising solution to the challenge of audio denoising.

Implementation

U-Net is a popular convolutional neural network (CNN) architecture that was initially developed for medical image segmentation tasks but has since been widely used in other applications. The U-Net architecture consists of an encoder-decoder structure with skip connections that enable the network to effectively learn the features of an image while preserving its spatial resolution.

As shown in figure 1, The encoder component of the U-Net architecture consists of a series of convolutional layers that extract high-level features from the input image. The decoder component, on the other hand, consists of a series of transposed convolutional layers that reconstruct the output image based on the learned features. The skip connections between the encoder and decoder components help to preserve the spatial information of the input image and allow the network to learn more accurate representations of the image features. The input to the neural network is the 2-channel (magnitude, phase) STFT representation of the signal, which is processed by the network in smaller windowed segments. An additional 10 channels of frequency-positional embeddings [4] are also added to the input data to help the network learn positional information with regard to the STFT frequency bins.

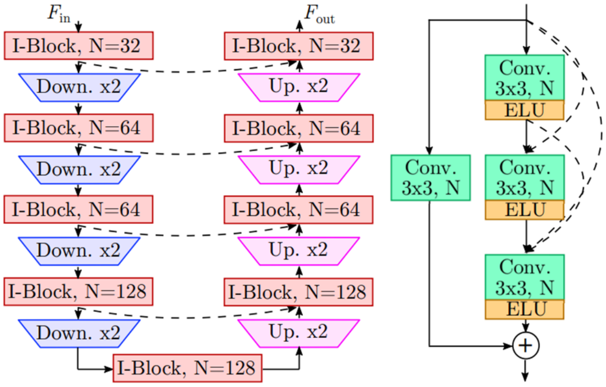

Figure 1: U-net and I-Block structure

During training the network processes input through the Stage-1 U-Net and SAM (supervised attention mechanism), creating higher-level features that are then concatenated to the input for the second U-Net to learn from.

The training loss for the full model is the sum of L1 losses (absolute difference with targets) across frequency bins of both the SAM and the Stage-2 UNet outputs, in order to minimize error in the SAM output and give the Stage-2 U-Net more meaningful concatenated feature representations. In this project we compared the full model's performance to an additional smaller model, which includes only one UNet preceded and followed by convolution layers. This smaller model has 6.9 million learnable parameters, while the full model has 13.8 million learnable parameters

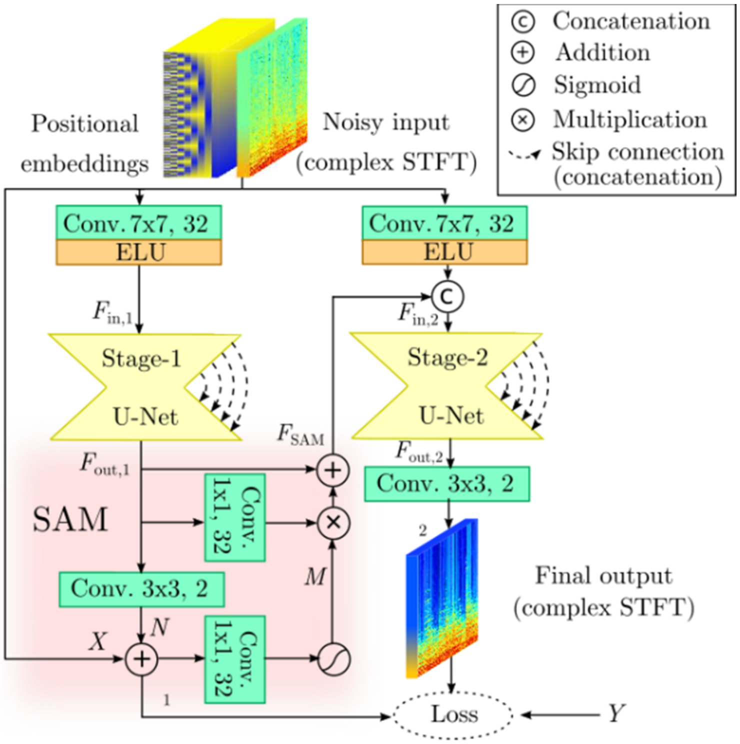

Figure 2: Two-stage denoising model with the SAM module

In our project, we attempted to denoise wildlife recordings by removing environmental noise from them. For example, in a recording of bird sounds, one could potentially hear also be human's talk, noise from vehicles, wind, etc... Our target then was to get rid of such noise and retain only the clean chirping.

The dataset used for this task is the ESC-50 dataset[5], which is a labeled collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification. The dataset consists of 5-second-long recordings organized into 50 semantical classes (with 40 examples per class). For this task- we split the dataset into two disjoint sets, which we called "animal" classes, and "noise" classes. In the data preprocessing, we used the following Eq.(1) to create noisy, corrupted audio samples.

$$X = \beta(Y + \alpha N) \tag{1}$$

Where variable Y represents the clean animal sound, N is added noise, and X the resulting corrupted sample. The two additional parameters α, β represent signal-to-noise-ratio and mean level scaling factor respectively. These allow adjusting the volumes of clean and noisy components randomly to imitate a real-life situation, in which levels vary and may be dynamic throughout the same recording.

For the task of data augmentation, we truncated the noise classes into small segments(2048 samples), windowed them (with Hamming window), and then combined them randomly with clean animal samples to get corrupted audio. This was a mean to give us flexibility in data, as well effectively prevent the model from over-fitting to our data. Our prepared dataset consists of one animal class (rooster), and one noise class (plane), with 90 audio samples in the training set and 30 samples in the validation set. When splitting the training data we made sure to use different plane and rooster samples for each of the two sets. We trained both models across 50 epochs, employing the adam optimizer with parameters lr=0.0005, betas = (0.7,0.9).

Result analysis

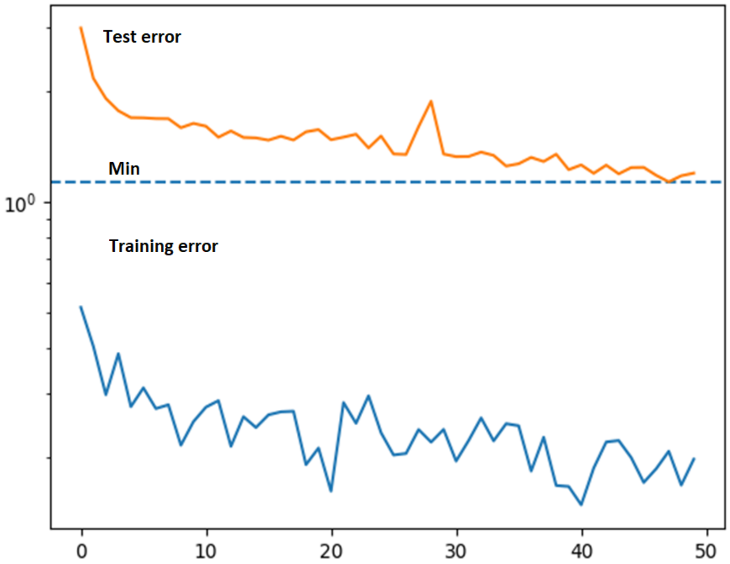

Figure 3: Smaller model error

We notice that the complete model performs better than the smaller model, as it achieves a lower minimum error on the test data set. We also compared the performances of the two models by comparing the signal

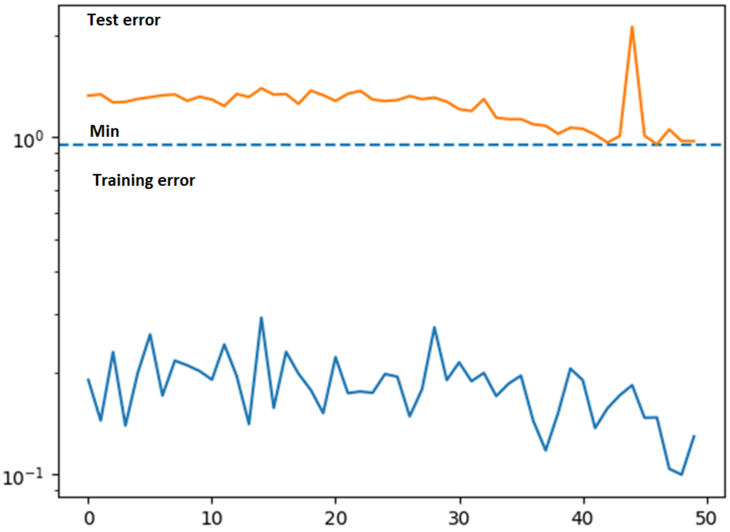

Figure 4: Complete model error

to noise ratio of their outputs. By treating the denoised audio as an estimation of the original clean animal sound, we can subtract the two to evaluate and compare the removed noise components. For this we use Eq.2 to compute segmental SNR across all windowed segments of the clean and corrupted samples, with results for both models shown in Figure 5. We discover that the complete model achieves better SNR than the smaller model for practically every window frame in the two denoised signals.

$$\mathrm{SNR}(k) = \frac{\sum_{f=0}^{F} |S(f,k)|^2}{\sum_{f=0}^{F} |E(f,k)|^2} = \frac{\sum_{f=0}^{F} |S(f,k)|^2}{\sum_{f=0}^{F} |S(f,k) - \hat{S}(f,k)|^2}, \tag{2}$$

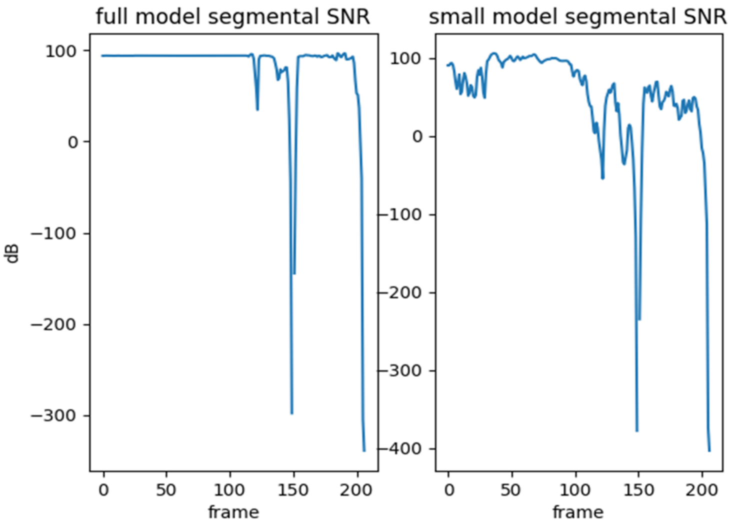

Figure 5: Segmental SNR

Individual contribution

Yifu Liu helped to prepare the dataset, perform data augmentation, and also helped to make the slides and wrote most of the report. Ido Akov implemented the two models using the pytorch library, trained them, evaluated them on the test data and analysed the results.

References

[1] Moliner E, Välimäki V. A two-stage u-net for high-fidelity denoising of historical recordings[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 841-845.

[2] S. J. Godsill and P. J. W. Rayner, Digital Audio Restoration—A Statistical Model Based Approach, Springer, 1998.

[3] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek, "Learning to denoise historical music," in Proc. 21st ISMIR Conf., Montr ́eal, Canada, Oct. 2020, pp. 504–511.

[4] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," in Proc. Interspeech, Oct. 2020, pp. 2487–2491.

[5] Karolpiczak. (n.d.). GitHub - karolpiczak/ESC-50: ESC-50: Dataset for Environmental Sound Classification. GitHub. https://github.com/karolpiczak/ESC-50

[6] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[7] Isik U, Giri R, Phansalkar N, et al. Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss[J]. arXiv preprint arXiv:2008.04470, 2020.