A MACHINE LEARNING APPROACH TO NOISE SUPPRESSION AND TIMBRE ENHANCEMENT OF EARLY 20TH CENTURY MUSIC

Jiaxi You Department of EEE, The University of Manchester, Manchester, UK
Patrick Gaydecki Department of EEE, The University of Manchester, Manchester, UK
Claire Mitchell Division of Neuroscience & Experimental Psychology, The University of Manchester, Manchester, UK

1 INTRODUCTION

This research focuses on the restoration of historical recordings produced before 1925, which is often termed the “Acoustic Era” [1]. The recording process relied solely on mechanical devices without microphones or electrical amplifiers. As shown in Figure 1, the horn was the primary device for capturing sound [2]. The wide end of the horn faced the instruments and performers, while the narrow end was connected to a diaphragm. The sound waves were gathered and funnelled towards the thin diaphragm and the energy from these sound waves caused the diaphragm to vibrate in response to the acoustic pressure change. A stylus was attached to the diaphragm, which is a pointed tool that translates the vibrations of the diaphragm into physical motion [3]. The stylus moved correspondingly while the diaphragm vibrated, etching the sound waves into a blank, rotating wax or shellac medium.

Figure 1: An American Studio of Victor, an American Subsidiary of Gramophone Co. [2]

Achieving an appropriate volume and balance between different sound sources was challenging, and the mechanical nature of the process imposed significant limitations on both the dynamic range and frequency response. This often resulted in recordings that lacked fidelity and clarity, with the frequency range typically restricted to between about 250 Hz and 2500 Hz. Additionally, the recording studios of the early twentieth century were far from ideal environments for producing high-quality sound, leading to recordings that were marred by various forms of ambient noise. With the aim of reviving the original sound from that period, this research restores digitized old recordings in two phases: noise removal and bandwidth enhancement.

During the Acoustic Era, several companies emerged as major players in the industry, playing crucial roles in the development and distribution of recorded music [2-5]. These included the Columbia Phonograph Company, founded in 1887; Edison Records, founded in 1888; the Berliner Gramophone Company, founded in 1895; and the Victor Talking Machine Company, founded in 1901. Most of the recording data used in this research originates from these four companies. This paper focuses entirely on solo piano recordings.

2 AUDIO DENOISING

2.1 Proposed Model Architecture

Inspired by the original U-Net convolutional neural network structure, initially developed for medical image segmentation [6], the audio denoising tasks can be approached as a segmentation problem: segmenting/separating the denoised audio signal from the noise. While U-Net has been successfully applied to denoising tasks [7-10], most studies focus on speech processing, which lacks the complex harmonic structures of music. Some research has addressed music denoising or source separation [11], but often with datasets featuring less dominant noise compared to those from the Acoustic Era.

The architecture of U-Net is characterised by a symmetric U-shaped structure, which consists of a contracting path and an expansive path, as shown in Figure 2. Similar to a typical CNN, the contracting path involves repeated application of convolutional and max-pooling layers to capture context and reduce spatial dimensions. The expansive path involves upsampling and convolutional layers to reconstruct the spatial dimensions and achieve precise localisation. Skip connections are used between corresponding layers of the contracting and expansive paths to combine high-resolution features from the contracting path with the upsampled output, enhancing the network's accuracy.

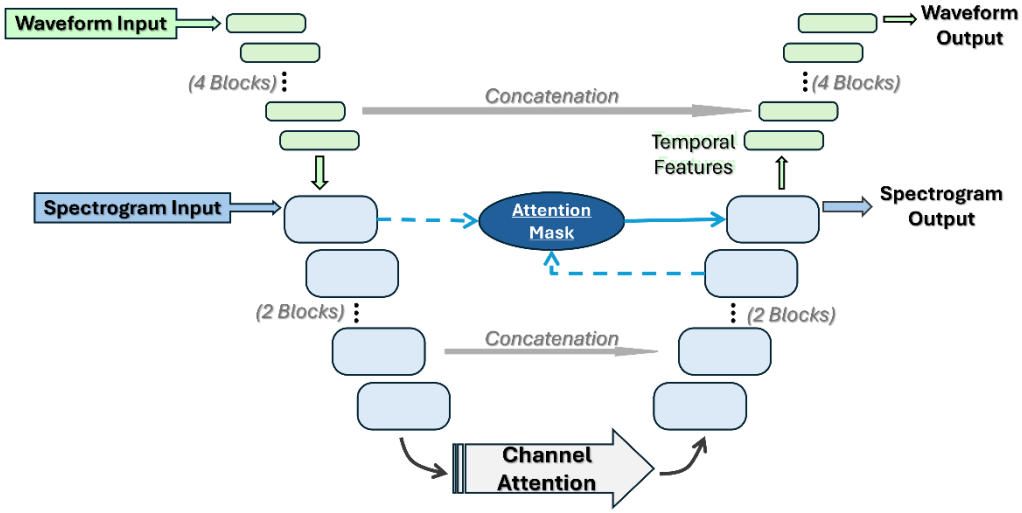

Figure 2: The Proposed Double Input U-Net Structure

Downsampling

This research employs representations from both the spectral domain and temporal domains for better denoising performance with fewer artefacts. Initially, the raw waveform is processed through encoder blockers for early feature extraction and downsampling. This preprocessing ensures that the waveform representations are dimensionally aligned with the spectral representations. The two representations are then passed to the same encoder, serving as two channels within the same U-Net structure. The approach provides the network with comprehensive information about the audio signal, enabling the model to learn the relationship between the spectral and the temporal domains.

Each waveform encoder block comprises a one-dimensional convolutional layer for feature extraction, an activation layer to introduce one-linearity to the model, and a max pooling layer to halve the sample size. Initially, the waveform input tensor has a shape of $(32767, 1)$ and after passing through eight feature extraction blocks, the feature tensor reshapes to $(256, 256)$. This tensor is combined with the spectrogram input to create a two-channel input, $(256, 256, 2)$, for further feature extraction. Each blue encoder block in Figure 2 consists of two 2D convolutional layers, two activation layers, and a max pooling with a pool size of $(2, 2)$. The encoders progressively reduce the data dimensions while increasing the feature depth. Ultimately, the bottleneck layer has a shape of $(8, 8, 512)$, representing

the most compact and abstracted form of the input audio data. This layer serves as the bridge between the encoder and the decoder.

2.1.1 Channel-Attention

Inspired by the channel-attention for speech enhancement of multichannel recordings [12], a channel-attention mask is employed at the bottleneck layer to improve feature representation by emphasising the most important channels of the compact audio data. As illustrated in Figure 3, it yields a refined output tensor that highlights the crucial channels and enhances the overall model performance.

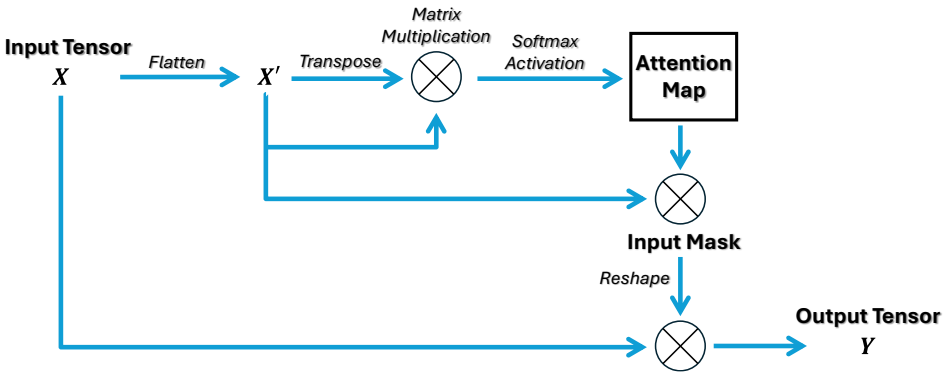

Figure 3: Channel Attention at Bottleneck

2.1.2 Upsampling

Symmetrically upsampling consists of six two-channel decoders and eight temporal signal decoders to expand the compact tensor back to $(256, 256)$ spectrogram output and $(32767, 1)$ waveform output. The decoders of both stages involve nearest neighbour upsampling, concatenation, convolution and activation; the only difference is that a self-attention mechanism is applied to the two-channel decoders to enhance important spatial locations, as shown in Figure 4. In U-Net architecture, the output from the upsampling is concatenated with corresponding feature maps from the downsampling path to integrate high-resolution spatial information with deeper feature information. Before concatenating, the self-attention gate facilitates better integration of features from different stages of the network, thus improving the quality of the upsampled features.

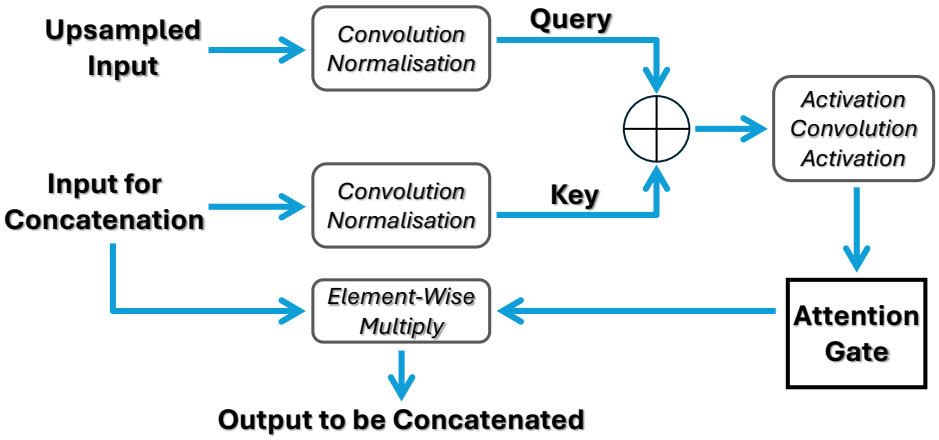

Figure 4: Self-Attention Mechanism for the Two-Channel Upsampling

2.2 Data Preparation

For a fully-supervised model, pairing every old recording with a high-fidelity, noise-free modern recording is not feasible. Instead, noise is artificially added to model recordings. Considering the

bandwidth limitations of historical recordings and computational efficiency, the training dataset uses a sampling rate of $11025\ Hz$, which sufficiently captures signals more than $25000\ Hz$ [2]. The target noisy audio pieces, Y, can be viewed as a mixture of the clean, narrow bandwidth signal, X, and the noise, N:

$$Y = X + \alpha * N \quad (1)$$

Where α is a signal-to-ratio (SNR) scaling factor, with a value between 0.7 and 1.2. To ensure the quality of the training database, which is crucial to the model's performance, the dataset includes noise excerpts that represent a range of degradations. A large collection of digitised gramophone records and noise-free music audio are publicly accessible in the Internet Archive [13]. The noise database includes white noise, ambient noise from the recording environment, low-frequency rumble from the turntable, and clicks and thumps from irregularities in the storage medium, which is sourced from historical recordings in noise-only segments to comprehensively cover the characteristics and types of noise of the era. In addition to the clear piano pieces, recordings featuring individual keynotes have also been manually collected. This approach aims to provide the model with a better understanding of the piano's harmonic structure. The dataset comprises 165 minutes of noise from old recordings, approximately 4.2 hours of high-quality piano audio and 12.3 hours of noisy pieces generated with $Equation\ (1)$.

2.3 Training

The training dataset consists of $15{,}000$ frames, amounting to 12.3 hours. From this, 10% is reserved for validation. The model is trained using the Adam optimiser with a learning rate of 0.0001. The Huber loss function is employed as it provides a balance between Mean Absolute Error (MAE) and Mean Squared Error (MSE). The training is conducted with a batch size of 10 and over 100 epochs. The entire training process required approximately 6 hours, utilizing $4.5\ GB$ of GPU RAM with Google Colab L4 GPU.

2.4 Denoised Results

2.4.1 Second Rhapsody – Franz Liszt, Publication date: 1919

Figure 5 presents s a frame of the piano piece along with its corresponding spectrogram. The green dotted arrows in the waveform highlight the muffled signals, typically indicative of ambient noise from the recording environment or low-frequency rumble generated by the turntable. Figure 6 shows the result from the proposed model. Both the waveform and spectrogram demonstrate a significant reduction in noise while preserving the integrity of the original piano signal.

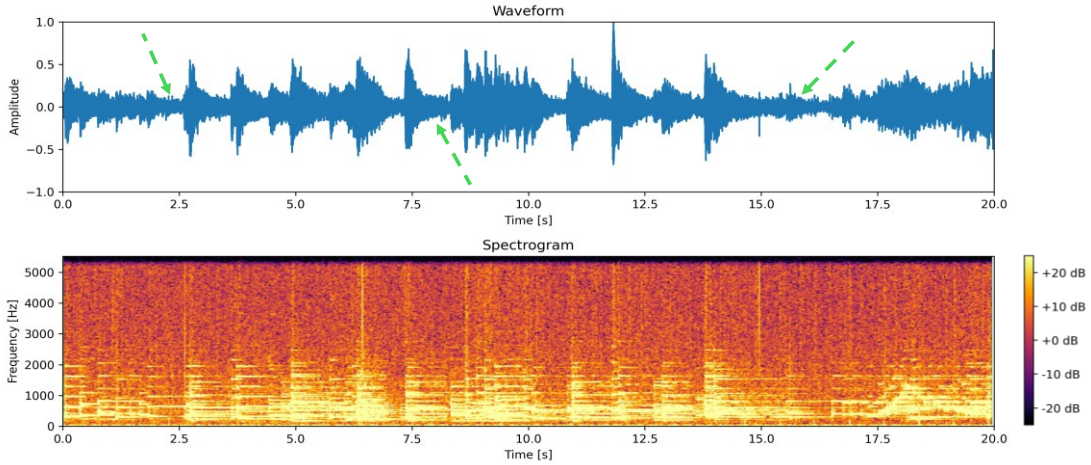

Figure 5: Noisy Piece from Second Rhapsody.

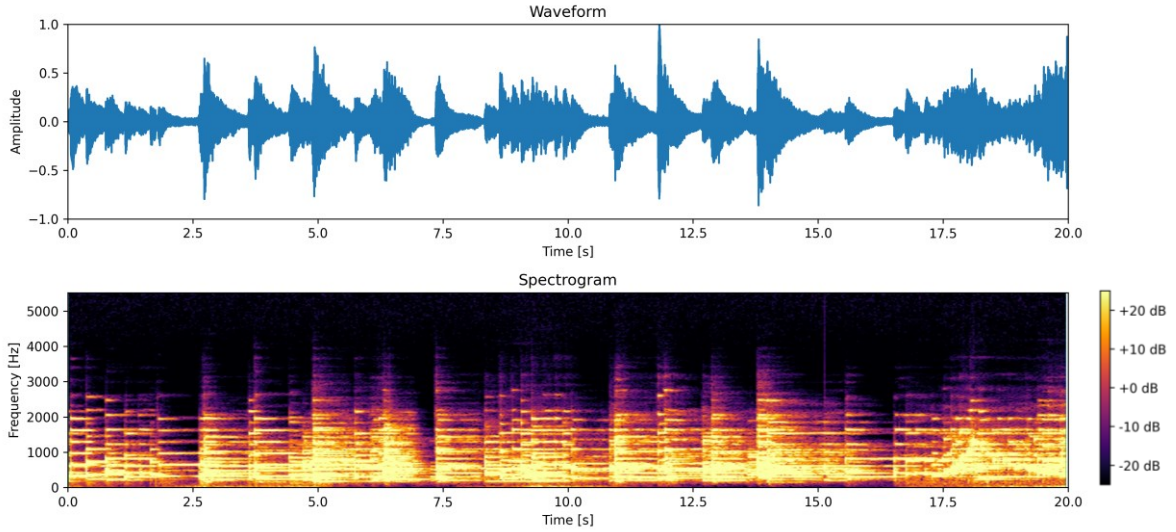

Figure 6: The Denoised piece of Figure 5.

2.4.2 Unknown Piano Piece from the Early 1920s

Figure 7 illustrates the audio suffering from considerable white noise, with the green arrows indicating the presence of clicks and pops/thumps. Figure 8 shows the denoised result, where all clicks have been removed and noise energy significantly reduced.

In the research, the spectrogram output was exclusively used for reconstruction and the denoised results consistently demonstrate that the proposed model is highly effective at denoising various types of noise. This model proves to be a versatile and generalised denoising method, capable of removing unwanted noise while preserving the integrity of the original audio signal, ensuring that the desired signal remains unaltered. The model demonstrated strong performance while requiring relatively simple training and minimal computational resources. Compared to other denoising methods, it operates efficiently with a smaller GPU size, making it a practical and resource-efficient solution.

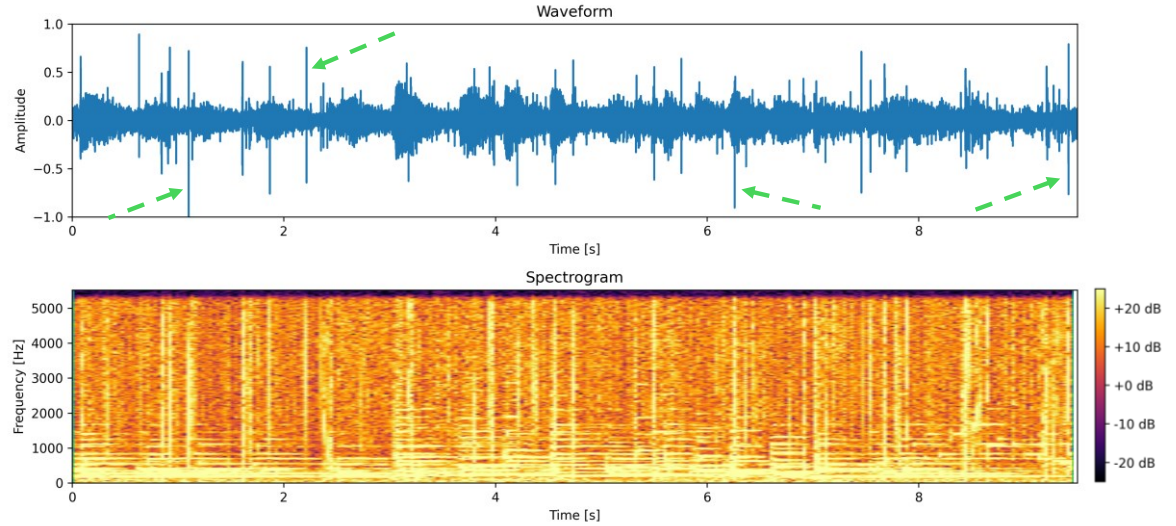

Figure 7: Noisy Piano Piece.

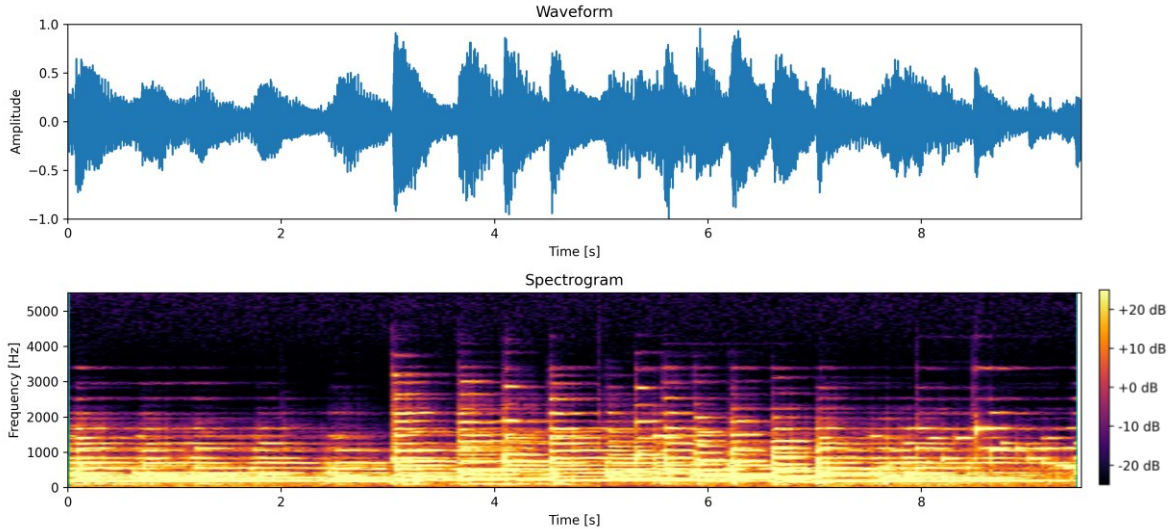

Figure 8: The Denoised Piece of Figure 7.

3 BANDWIDTH ENHANCEMENT

Due to the limitations of the mechanical recording process, the recording captured signals only up to the maximum $4{,}000\ Hz$. As a result, it typically requires enhancement of low-frequency signals and prediction of the missing high-frequency content.

3.1 Linear Approach - Compensating Filters

Applying filters to emphasise different frequency ranges can enhance the audio by making certain elements more prominent. Figure 9 shows Chopin's denoised piano piece *Revolutionary Étude* and Figure 10 is the enhanced result, with filters applied to the low-frequency range, resulting in a stronger bass and better listening experience. However, this linear approach is limited to signals present in the original recordings, and the audio still lacks the brightness of higher frequencies, indicating the underrepresentation of high-frequency signals. For a more balanced and natural listening experience, nonlinear approaches, such as generative models, are required to predict the missing bandwidth.

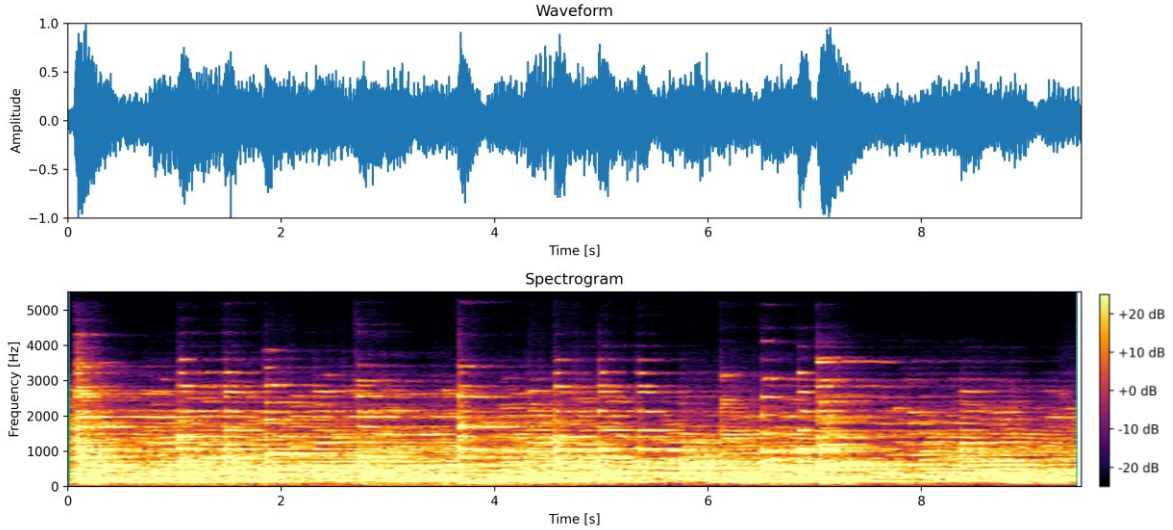

Figure 9: Denoised Piece of *Revolutionary Étude*, Chopin, 1922.

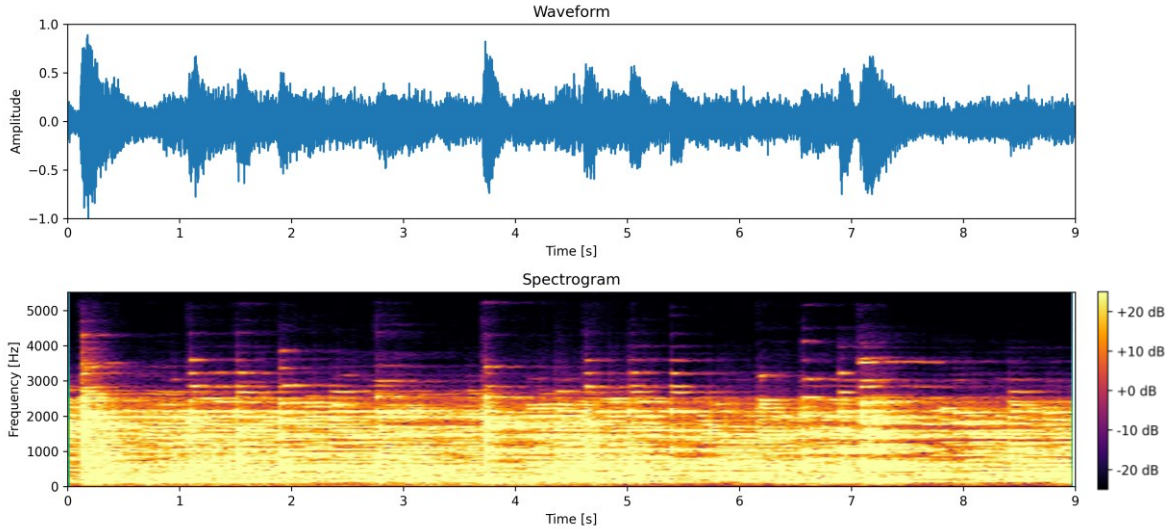

Figure 10: Enhanced Piece of Figure 9.

3.2 Diffusion Models

The diffusion model is a generative model based on statistical and probabilistic principles [14]. Initially developed for image tasks, it primarily works with 2D inputs. It operates by progressively adding Gaussian noise - distinct from audio noise - to the original data and then gradually remapping it to generate new content. These two processes are known as diffusion/forward process and reverse process. This iterative process allows the model to learn complex data distributions, making it particularly effective for tasks such as audio enhancement.

3.2.1 Diffusion Process

This process can be thought of as progressively corrupting the data until it becomes pure noise data x_T. Given the time step t and the original clean data x_0, the process can be represented as:

Where $\overline{\alpha}_t$ is a scalar representing the cumulative effect of the noise schedule up to time t, with t being an integer from 0 to T, typically 1000 or more. Gaussian noise ε is drawn from a standard normal distribution $\mathcal{N}(0, I)$. This process is purely statistical and does not involve any trainable parameters.

$$x_t = \sqrt{\overline{\alpha_t}} * x_0 + \sqrt{1 - \overline{\alpha_t}} * \varepsilon \tag{2}$$

3.2.2 Reverse Process

The goal of the reverse diffusion process is to start from the noisy data x_T, which is almost pure noise and iteratively denoise it, to get a distribution that resembles the original data x_0. It can be represented as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{1-\overline{\alpha_t}}\varepsilon_\theta(x_t, t)\right) + \sigma_t z \tag{3}$$

Where $\sigma_t z$ is the variance that does not influence the training result. The neural network model is trained to predict the noise ε_θ added at each step. By estimating and subtracting the noise, the model gradually reconstructs the original data x_0.

3.2.3 Training

The model is trained on 11.2 hours of high-quality piano data and 10% of which is used for validation. The linear scheduler is used for the diffusion process and there are $T = 1000$ time steps. The training is conducted with a batch size is 2 and over $13{,}500$ training steps. The entire training process takes approximately 17 hours.

3.2.4 Bandwidth Enhanced Result

Figure 11 shows a piano piece with limited bandwidth, where the absence of higher frequencies results in a muffled sound and diminished brightness. Figure 12 illustrates the enhanced result, featuring a more balanced and vibrant sound. The restored high frequencies contribute to increased brightness and clarity, providing a more detailed and dynamic audio experience.

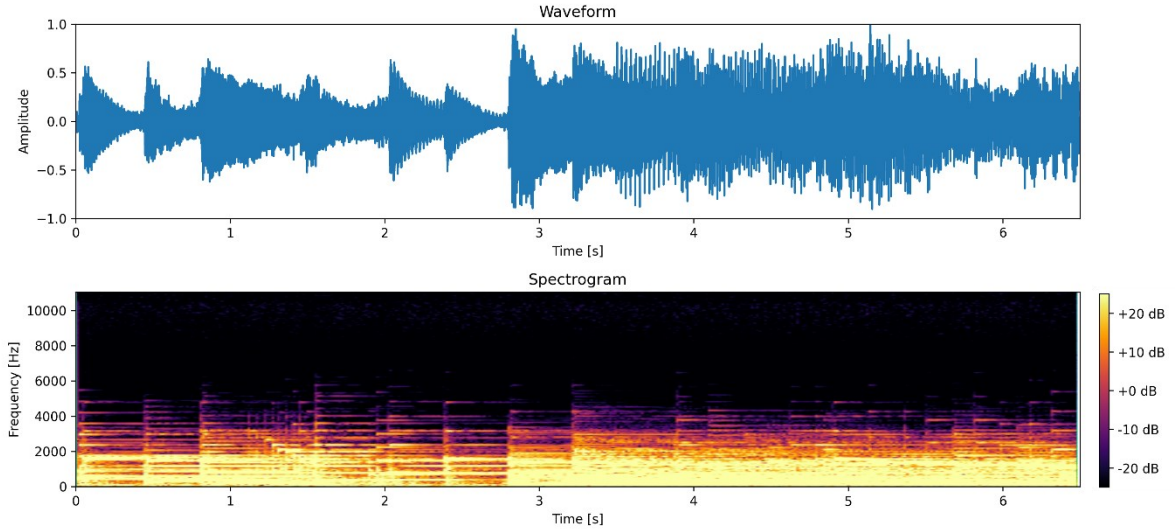

Figure 11: Audio Piece with Limited Bandwidth.

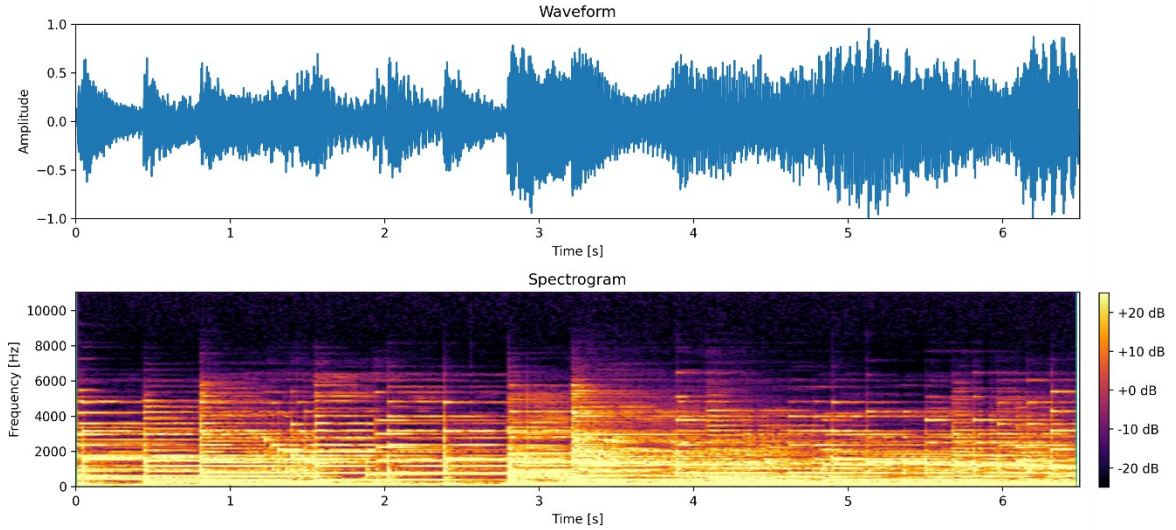

Figure 12: Enhanced result of Figure 11.

4 DISCUSSION AND FUTURE WORK

The diffusion model demonstrates the capability to generate missing audio signals; however, its progressive nature necessitates substantial training resources and extended training time. For severely degraded historical recordings, the model introduces artefacts. Further work will focus on refining the results for such historical audio recordings.

Testing results indicate that the U-Net-based model performs well in piano denoising tasks and shows promise for denoising string instrument recordings, despite being trained exclusively on piano signals. This suggests that the model effectively captures harmonic structures. Additionally, the model was able to denoise audio samples with a sampling rate of $22{,}500$ Hz, even though it was trained only on

a $11{,}025$ Hz sampling rate. These results indicate that the model has broader potential for a variety of denoising tasks.

In the future, integrating the proposed model with diffusion models and further developing it could enable applications such as orchestral denoising or the restoration and enhancement of historical speech recordings.

5 REFERENCES

1. "Acoustical Recording | Articles and Essays | National Jukebox | Digital Collections | Library of Congress", *Library of Congress, Washington, D.C. 20540 USA* [Online]. Available: https://www.loc.gov/collections/national-jukebox/articles-and-essays/acoustical-recording/
2. M. J. Lindquist, "Unit Two, Part One: History of Audio Recording," *mlpp.pressbooks.pub*, 2021 [Online], Available: https://mlpp.pressbooks.pub/audioproduction/chapter/unit-three-part-one-history-of-audio-recording/
3. "The Victor-Victrola Page," *Victor-victrola.com*, 2023 [Online]. Available: http://www.victor-victrola.com/Basics%20of%20the%20Acoustic%20Phonograph.htm
4. J. Sterne, The Audible: Cultural Origins of Sound Reproduction, Durham: Duke University Press, 2003.
5. "Columbia Corporate History: Introduction - Discography of American Historical Recordings, " *adp.library.ucsb.edu* [Online]. Available: https://adp.library.ucsb.edu/index.php/resources/detail/97
6. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", in *Lecture Notes in Computer Science*, 2015, pp. 234-241. doi: 10.1007/978-3-319-24574-4_28
7. A. Jansson, E. J. Humphrey, e. al., "Singing Voice Separation with Deep U-Net Convolutional Networks", in *International Symposium/Conference on Music Information Retrieval*, pp. 745–751, Feb. 2020.
8. D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation", in *19th International Society for Music Information Retrieval Conference*, 2018.
9. R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for Speech Enhancement", in *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, 2019.
10. V. Patkar, T. Parmar, e. al., "Audio Source Separation Using Wave-U-Net with Spectral Loss", in *IEEE International Conference on Communication System, Computing and IT Applications*, 2023.
11. E. Moliner and V. Valimaki, "A Two-Stage U-Net for High-Fidelity Denoising of Historical Recordings", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022.
12. K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, doi: https://doi.org/10.1109/iccv.2015.123.
13. "The Internet Archive", http://archive.org, Accessed: 31st Jul. 2024.
14. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models", in 34th Conference on Neural Information Processing Systems, Canada, 2020.