2024 NCCU Deep Learning Course Final Project

Spectrogram-Based Deep Learning for Audio Denoising

112753120 謝皓雲 113753114 林祐祥

ABSTRACT

This project investigates the application of non-recurrent models, specifically CNNs and U-Nets, for audio denoising tasks. By converting audio into spectrograms, the problem is reframed as an image-processing challenge. Simplified architectures are designed to balance training efficiency and effectiveness. Results show that U-Nets slightly outperform CNNs in enhancing audio intelligibility.

---

1. INTRODUCTION

Audio denoising is an essential preprocessing task in many speech-based applications, including telecommunications, voice-controlled systems, and assistive hearing devices. Traditional models for audio processing often leverage the sequential nature of audio data using recurrent architectures like RNNs and LSTMs. These models, while effective, suffer from limitations in scalability and parallelization. This project seeks to explore the capability of non-recurrent models like CNNs and U-Nets in addressing these challenges by treating audio spectrograms as image-like data.

In this project, we want to examine the possibility of using non-recurrent models like CNN and U-Net to denoise speech audio by converting audio data to **spectrogram**—a visual representation of the audio signal across frequency and time. Spectrograms preserve the time-frequency structure of audio while transforming the problem into an image-processing task, which non-recurrent models handle efficiently. Figure 1. is an example of how a spectrogram looks like.

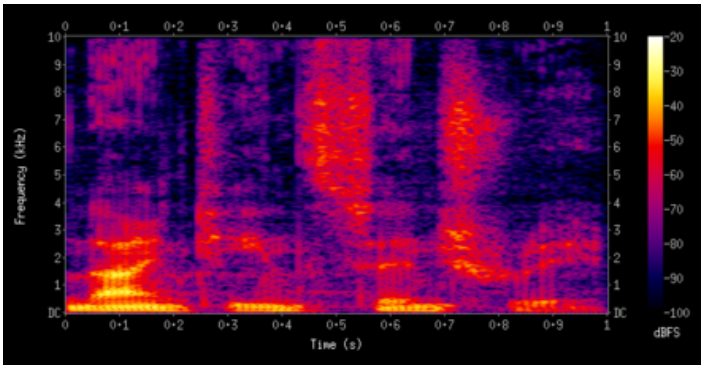

Figure 1. Example of how spectrograms look like. (source: Wikipedia)

2. RELATED WORK

Recent advancements in deep learning have revolutionized audio processing. Among the most notable contributions:

1. **Wave-U-Net for Speech Source Separation (Stoller et al., 2018)**:

   This model extends the U-Net architecture by introducing multi-scale features, allowing it to handle diverse frequency components in speech signals. Wave-U-Net demonstrated its potential in separating multiple sound sources within a single audio stream, making it a precursor for tasks like denoising.

2. **Residual Attention U-Net for Speech Denoising (Takahashi et al., 2020)**:

   This approach integrates attention mechanisms into the U-Net framework, focusing on relevant time-frequency regions in the spectrogram. The attention modules enhanced denoising performance by dynamically prioritizing significant features.

3. **CNN-Based Spectrogram Processing**:

   Several studies explored CNNs for audio tasks due to their ability to extract local spatial patterns. While basic CNNs are less sophisticated than U-Nets, their simplicity allows for efficient training and deployment. For example, Zhao et al. (2019) applied CNNs to spectrograms for environmental sound classification, highlighting their effectiveness in audio tasks.

Inspired by these works, this project adopts a simplified U-Net architecture to optimize the trade-off between performance and computational cost. By reducing architectural complexity, we aim to achieve competitive denoising performance while minimizing training time.

3. DATASET

We utilized the Kaggle noise cancellation dataset (by Srujan Nagamalla), which contains paired clean and noisy WAV audio samples:

- **Training Set**: 11,572 clean/noisy pairs.
- **Testing Set**: 824 clean/noisy pairs.

This dataset is well-suited for evaluating audio denoising models as it provides a diverse range of noise patterns and clean speech signals.

The dataset can be accessed via this link:

https://www.kaggle.com/datasets/srujannagamally/noise-cancellation?select=clean_trainset_wav

4. METHODOLOGY

4.1 Pipeline

1. **Conversion to Spectrograms**:
   - Audio signals are transformed into spectrograms using the Short-Time Fourier Transform (STFT), separating the magnitude and phase components.
2. **Model Processing**:
   - The magnitude spectrogram is processed by CNN or U-Net models to generate a denoised version.
3. **Reconstruction to Audio**:
   - The denoised magnitude spectrogram is combined with the original phase component and converted back to audio using the inverse STFT.

Figure 2. is the basic pipeline of our project, note that in which the phase spectrum would be kept unmodified as it is the key to convert denoised image back to audio.

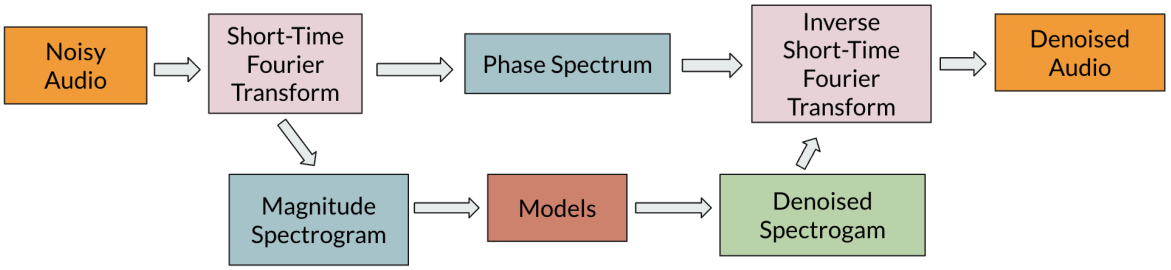

Figure 2. Pipeline of this project.

4.2 Evaluation Metric

We use the **Short-Time Objective Intelligibility (STOI)** metric, which quantifies intelligibility by calculating the correlation between clean and denoised speech in overlapping short-time frames. STOI scores range from 0 to 1, with higher values indicating better intelligibility. STOI can be calculated as below:

$$\mathrm{STOI} = \frac{1}{K}\sum_{k=1}^{K} \rho\left(\mathbf{x}_k, \hat{\mathbf{x}}_k\right)$$

4.3 Experiment Design

Two models were trained:

1. **Basic CNN**: A simple architecture designed to test the baseline effectiveness of CNNs in processing spectrograms.

2. **Simplified U-Net**: A lightweight U-Net design that reduces the number of parameters while retaining its hierarchical feature extraction capability.

Their architectures are shown in Figure 3.

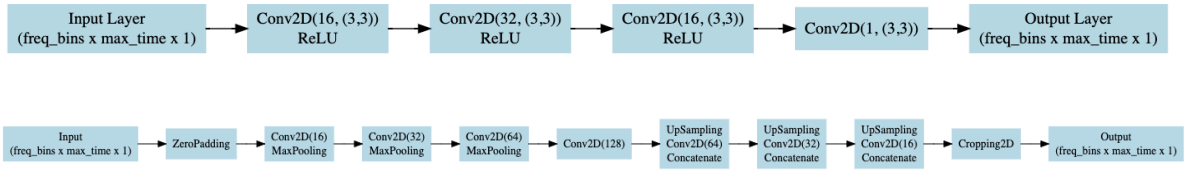

Figure 3. Model architectures.

5. RESULTS

5.1 Spectrogram and Audio Comparisons

- **Average STOI Scores**:
    - CNN: 0.91
    - U-Net: 0.9257

Both models successfully reduced noise, particularly high-frequency components. The U-Net slightly outperformed the CNN, demonstrating better preservation of low-frequency speech signals critical for intelligibility.

5.2 Waveform Comparison

Waveform analyses revealed that both models effectively suppressed background noise while preserving the overall structure of the clean speech signal. As what can be seen in Figure 4., green lines (stands for denoised data) successfully reduced noises, which resulted in the exposure of purple lines (noisy signals); and are closer to orange lines (ground truth).

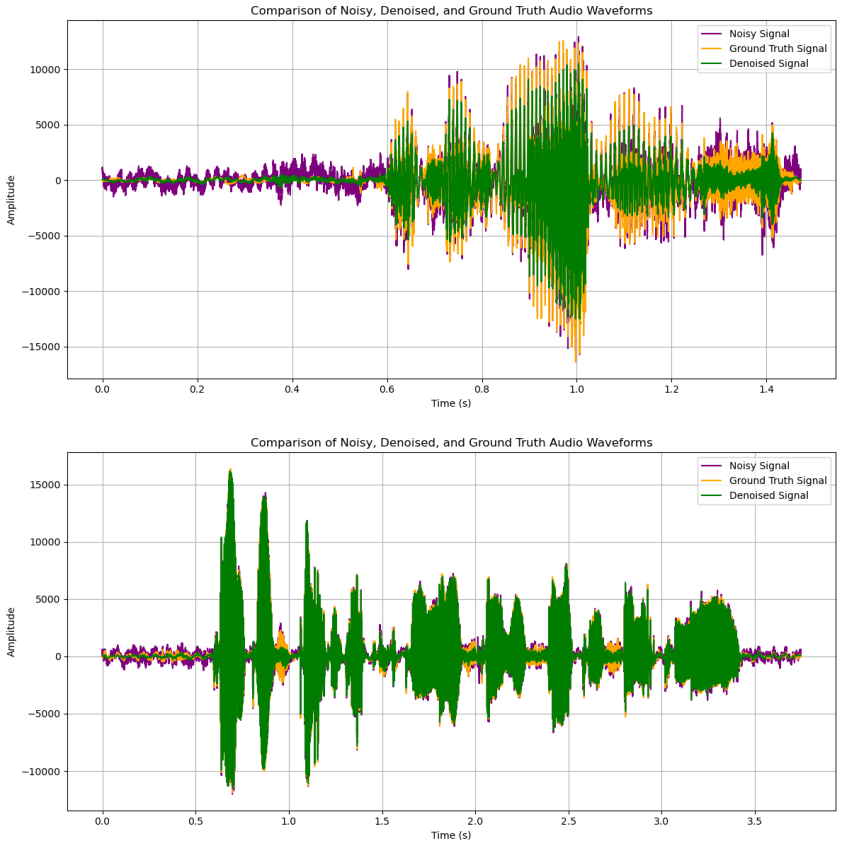

Figure 4. Waveform comparison on two of the test audio.

However, from purely spectrograms, both models basically learnt to directly remove high-pitched signals that are the most likely to be the noise (refer to Figure 5.). Despite the good results in terms of subjective hearing and waveform comparison, the spectrograms of denoised data did not look very similar to the ground truth. This could be due to the loss function choice, as we used STOI instead of image based metrics such as SSIM or PSNR.

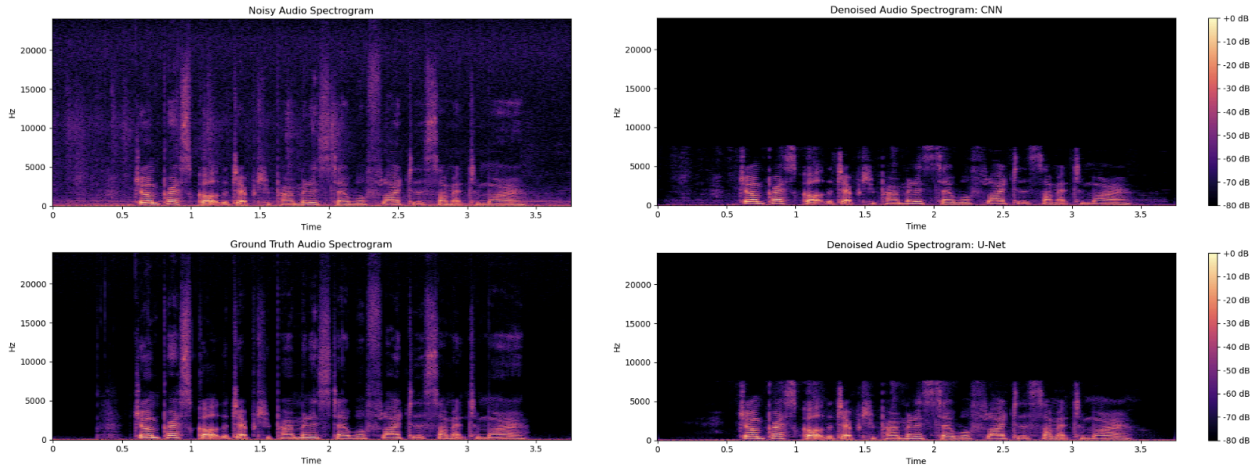

Figure 5. Spectrograms comparison on one of the test audio.

If one directly removed high-pitch noises manually using STFT and a sound filter, the average STOI would be 0.75, which, in comparison to our deep learning models (with the STOI around 0.91), is significantly lower. That means our models still perform better than non-machine-learning, non-deep-learning methods.

6. DISSUSSION

6.1 Key Observations

- The models performed well in removing high-frequency noise, aligning with the intuition that such components often constitute background noise. However, low-frequency noise or speech distortions proved more challenging.
- While STOI is a suitable metric for assessing intelligibility, its limitations in reflecting perceptual similarity between spectrograms were evident. For instance, the denoised spectrograms diverged significantly from the ground truth, despite high STOI scores.

6.2 Potential Improvements

1. **Loss Function Optimization**:
   - Using metrics like **Peak Signal-to-Noise Ratio (PSNR)** or **Structural Similarity Index (SSIM)** as loss functions could encourage better spectrogram reconstruction. However, these metrics may not directly correlate with intelligibility, making them suboptimal for audio tasks.
2. **Phase Reconstruction**:
   - Incorporating phase information during denoising could enhance audio fidelity, as the current models rely on the unprocessed phase component, limiting the quality of the reconstructed audio.
3. **Attention Mechanisms**:
   - Adding attention layers, inspired by Takahashi et al. (2020), could help models focus on important regions of the spectrogram, improving denoising performance.

6.3 Implications

The simplified U-Net demonstrated an excellent balance between efficiency and performance, making it suitable for real-time applications. Its ability to outperform the CNN highlights the importance of hierarchical feature extraction in handling complex audio signals.

7. CONCLUSION

1. Both CNN and U-Net models effectively denoised audio, with the U-Net showing a slight advantage in intelligibility.

2. STOI, while effective, may not fully capture perceptual quality, necessitating further exploration of alternative evaluation metrics.

3. Future work could explore advanced architectural features, such as attention mechanisms and phase-aware processing, to further enhance denoising performance.

8. REFERENCES

1. D. Stoller, S. Ewert, and S. Dixon, “Wave-U-Net: A multi-scale neural network for end-to-end audio source separation,” ISMIR, Sept. 2018.

2. J. Takahashi, S. Kohmura, and T. Togawa, “Speech Denoising with Residual Attention U-Net,” Proc. 8th IIAE Int. Conf. Ind. Appl. Eng., 2020.

3. Zhao et al., “Environmental Sound Classification with CNN-based Spectrogram Analysis,” IEEE Trans. Audio, 2019.