SPEECH ENHANCEMENT USING GENERATIVE DIFFUSION MODELS

Prabha Sharma - M22AIE224, Prateek Singhal - M22AIE215, Harsh Parashar - M22AIE210
Department of Computer Science, IIT-Jodhpur

1. Introduction

Clear audio and natural-sounding speech are crucial for effective voice communication. However, factors like background noise, room acoustics, transmission errors, and limited bandwidth can degrade speech signals. Traditional speech enhancement methods focus on specific distortions, which may not be effective when multiple distortions are present. This project aims to develop an automated speech enhancement system using diffusion models, known for handling missing and non-linear data. By leveraging generative modeling, we seek to create a universal approach to improve speech signals affected by various distortions.

The code for the project can be found on GitHub(https://github.com/parasharharsh16/Speech-Enhancement-using-Diffusion) and Training experiment's dashboard can be accessed via Wandb(Link in README)

2. Problem Statement

The problem we seek to address is the enhancement of speech signals in noisy and degraded environments. Existing speech enhancement techniques typically rely on heuristic approaches or single-task models, which may not generalize well to diverse types of distortions. Our goal is to overcome these limitations by developing a robust and adaptable speech enhancement system.

3. Literature Review

Jean-Marie Lemercier[2] and their co-authors presented an analysis of using generative model vs discriminative approaches for speech restoration and observed that the generative approach performs globally better than its discriminative counterpart on all tasks.

Recent research in speech enhancement has shown increasing interest in generative modeling approaches, such as diffusion models. These models have demonstrated effectiveness in scenarios with missing and non-linear data. For example, the work by S. Welker, J. Richter, and T. Gerkmann[4] showcased the use of score-based generative models for the STFT domain. Similarly, a paper titled "Speech enhancement and dereverberation with diffusion-based generative models"[3] referred to an approach of using mixed (noise + speech) as input to the Diffusion model instead of pure Gaussian noise, which enabled them to generate high-quality audio.

3.1. Dataset

We have used the publicly available speech datasets such as the VCTK dataset and environmental noise for reverberation. VCTK contains clean speech recordings and we created noisy samples by combining different power of environmental noise with speech recordings. Which allowed us to simulate various types and levels of distortions. VCTK, Environmental Noise.

Audio Files (wav)	Txt Labels
44.2k	44.2k

Table 1: *Audio Files and Text Labels*

Recording Setup	Details
Microphone	Omni-directional microphone (DPA 4035)
Sampling Frequency	96 kHz
Bit Depth	24 bits
Recording Environment	Hemi-anechoic chamber, University of Edinburgh
Pre-processing	Conversion to 16 bits, downsampling to 48 kHz

Table 2: *Recording Setup for Speech Data*

4. Network Architecture

U-Net is a highly effective convolutional neural network (CNN) for speech enhancement, renowned for its robust capability to process and improve audio signals. Its unique architecture is adapted from its original use in medical image segmentation to enhance audio quality by reconstructing clean speech from noisy environments.

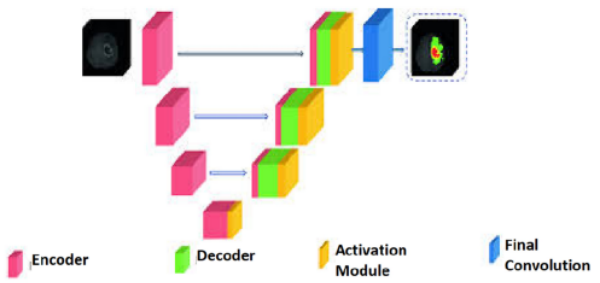

Figure 1: *U-Net Architecture*

4.1. Core Components:

1. **Encoder:** Captures essential speech characteristics by compressing the audio signal, using a series of convolutional and pooling layers to reduce noise while extracting relevant features.
2. **Bottleneck:** Processes the most abstract features at the lowest resolution, focusing on capturing the overall sound structure essential for effective noise reduction.
3. **Decoder:** Reconstructs the enhanced speech by progressively restoring detail and clarity through up-sampling and convolutional layers, utilizing crucial skip connections from the encoder to maintain high-quality output.

4.2. Key Feature:

1. **Skip Connections:** Skip connections are vital, as they reintegrate detailed, high-resolution features from the encoder directly to the decoder, ensuring that the reconstructed speech is clear and precise.

Advantages of Using U-Net for Speech Enhancement:

1. **Effective Feature Integration:** Detail and Context: U-Net's architecture, with its encoder-decoder structure and skip connections, effectively captures detailed and contextual features of audio signals. This enables precise differentiation between speech components and noise, essential in noisy environments.
2. **High-Quality Reconstruction:** Skip Connections: These connections facilitate the use of detailed features from earlier layers in the decoding process, enhancing the clarity and intelligibility of the reconstructed speech.
3. **Handling Temporal Dependencies:** Adaptation to Audio: Though originally designed for spatial data, U-Net can be adapted to process temporal features in audio, making it suitable for sequential data like speech.
4. **Flexibility and Adaptability:** Customizable Architecture: The network can be tailored in depth, filter counts, and configurations to meet specific speech signal characteristics, allowing effective handling of different noise types.
5. **Improved Performance in Noisy Conditions:** Robustness to Noise: U-Net is adept at distinguishing speech from noise, proving beneficial in applications like mobile communications and hearing aids where background noise is prevalent. In summary, U-Net's ability to merge contextual and detailed information through its unique architecture makes it highly effective for speech enhancement tasks, particularly in environments with variable noise characteristics.

5. Results

In addition to the subjective evaluation we used the below metrics to evaluate the performance of the models.

1. Evaluations Loss and Similarity

Test Loss	Similarity Score
0.0215	0.9980

Table 3: *Test Loss and Similarity Score*

2. Wave Spectograms Our analysis involved evaluating wave spectrograms in three specific scenarios: clean audio, noisy audio, and the predicted (clean) audio generated by our model. The spectrogram of the clean audio served as a reference, showcasing the frequency components of the original signal. In contrast, the spectrogram of the noisy audio revealed the alteration of the frequency spectrum caused by environmental noise or interference. The spectrogram of the predicted (clean) audio clearly demonstrated the model's ability to effectively remove noise, as it successfully reconstructed the clean signal from the noisy input. These observations provide valuable insights into how the model improves audio quality by removing noise and restoring signal clarity.

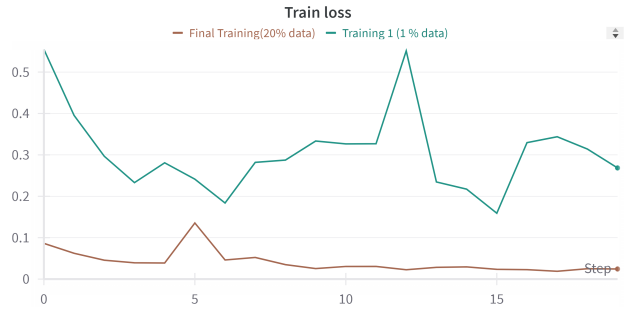

Figure 2: *Model Training Loss/Epochs Plot*

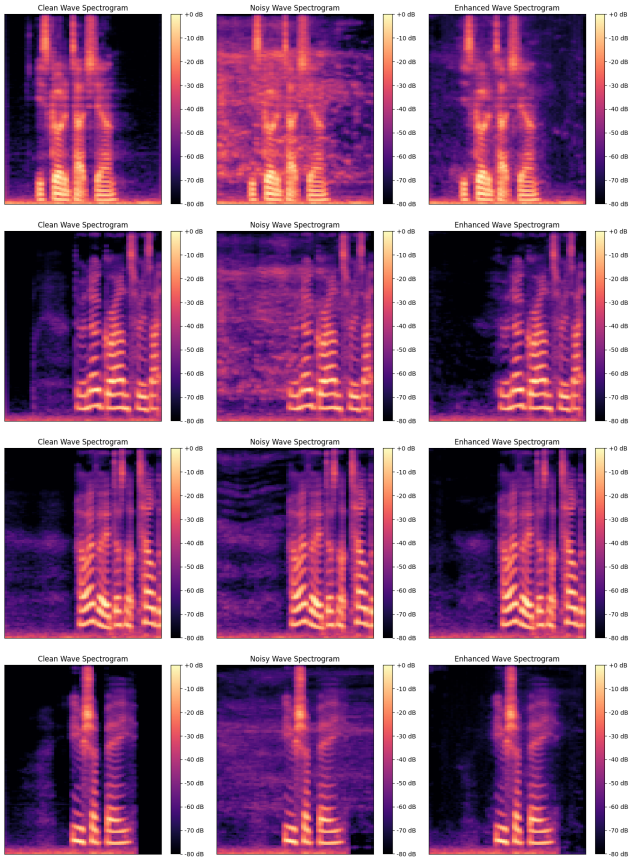

Figure 3: *Spectogram Results*

6. References

(a) Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Tal Peer, Timo Gerkmann, "Speech Signal Improvement Using Causal Generative Diffusion Models," *IEEE*, 2023.

(b) J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023.

(c) J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv preprint*, 2022.

(d) S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," *Interspeech*, 2022.