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 nonlinear data. By leveraging generative modeling, we seek to create a universal approach to improve speech signals affected by various distortions.

This project is using the paper titled ”Speech Signal Improvement Using Causal Generative Diffusion Models”[1] published by IEEE in 2023, link to paper.

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.

4. Proposed Solution

Our proposed solution involves developing a speech enhancement system based on diffusion models. We will train the model using a diverse dataset containing clean speech samples corrupted by different types of distortions, including background noise, reverberation, transmission errors, and codec artifacts.

- Diffusion Process: This core method progresses from clean to corrupted speech during training, gradually adding Gaussian noise. At inference, a reverse process removes corruption, generating an estimate of clean speech. A trained deep neural network approximates the necessary score function.
- Network Architecture: A modified version of NCSN++ is utilized for score estimation. This encoder-decoder architecture employs 2D convolutions, with adjustments for causality such as modified padding and cumulative group normalization.
- Automatic Gain Control (AGC): AGC is applied before and after enhancement to normalize the spectrogram of corrupted speech and maximize loudness. It tracks maximum values per magnitude frame over frequency bins in a causal manner, aided by speech activity detection methods and a causal compressor.

4.1. Dataset

We plan to use publicly available speech datasets such as the VCTK dataset, QUT corpus as the environmental noise and take room impulse responses from the DNS challenge for reverberation. These datasets contain clean speech recordings as well as artificially generated noise samples, which will allow us to simulate various types and levels of distortions. VCTK, QUT, DNS.

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*

5. Evaluation metrics

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

1. DNSMOS (Differential Mean Opinion Score)
2. SIG (Speech Quality)
3. OVRL (Overall Quality)

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.