Speech Enhancement and Transcription under Noisy Conditions using Wiener Filtering and Whisper

Lt Col Ritesh Lamba
MTech (AI) Executive, Indian Institute of Technology, Jodhpur
Roll No: M23CSA544
M23CSA544@iitj.ac.in

Abstract

This report presents the methodology and evaluation for enhancing moderator speech clarity from noisy audio recordings using Wiener filtering and Whisper-based transcription. Audio recordings from two sets were processed: one containing both clean and noisy versions, and another with only noisy recordings. Performance was evaluated using SNR, WER and MOS metrics.

1. Introduction

Ensuring intelligibility in speech under noisy conditions is crucial for applications like event archiving and automatic meeting transcription. This study focuses on denoising moderator speech while preserving quality using classical Wiener filtering. Transcriptions were generated using Whisper and evaluated for clarity and accuracy.

2. Dataset and Preprocessing

Set 1: 8 audio files with clean and corresponding noisy versions.
Set 2: 4 audio recordings with only noisy data (`bus, cafe, ped & street`).
All audio files were in WAV format sampled at 16kHz or higher.

3. Noise Analysis

Noise levels were analyzed using Signal-to-Noise Ratio (SNR) computed between noisy and clean files in Set 1. Frequency characteristics were visualized using spectrograms. Remaining images are available at Google Drive Link in the end.

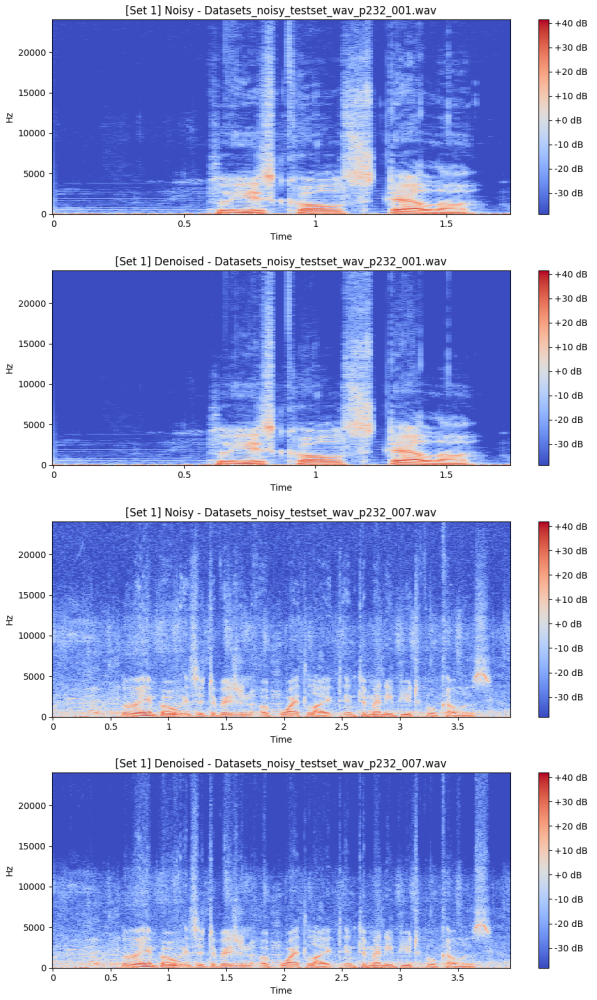

Figure 1. Set 1 Spectrograms: Noisy (top) vs Denoised (bottom)

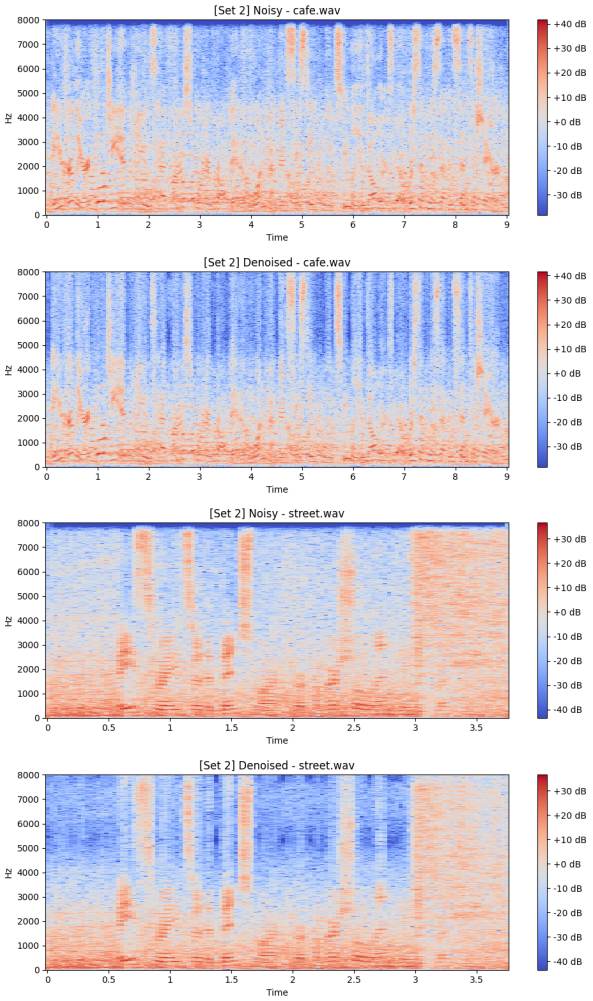

Figure 2. Set 2 Spectrograms: Noisy (top) vs Denoised (bottom)

4. Denoising Algorithm

We used Wiener filtering implemented using SciPy, which suppresses noise based on statistical estimation of the clean signal. A minimal additive noise was introduced to avoid divide-by-zero issues. All files were saved after denoising and used for transcription and evaluation.

5. Transcription

OpenAI's Whisper (base model) was used for automatic speech recognition. Both clean and denoised audios were transcribed. Word Error Rate (WER) was calculated by comparing clean vs denoised transcripts for Set 1.

6. Evaluation Metrics

6.1. SNR (Objective)

Computed between clean and noisy, and clean and denoised versions for Set 1.

6.2. WER (Objective)

WER calculated using jiwer between clean transcripts and denoised transcripts (Set 1). For Set 2, relative WER (noisy vs. denoised) was used.

6.3. MOS (Subjective)

Mean Opinion Scores collected from 5 human raters who rated clarity and naturalness of denoised files on a scale of 1 to 5.

7. Results

7.1. Set 1 Evaluation

Set 1 - SNR, WER, Transcriptions

filename	SNR Before	SNR After	WER	Clean Transcript	Denoised Transcript
Dataset_noisy_instruct_wav_p232_001.wav	15.68	17.44	0.0	Please call Stella.	Please call Stella.
Dataset_noisy_instruct_wav_p232_002.wav	11.53	11.3	0.2	Ask her to bring these things with her from the store.	Ask her to bring these things with her from the store.
Dataset_noisy_instruct_wav_p232_003.wav	6.73	8.72	0.3	It spoons of fresh snow peas, 5 thick slabs of blue cheese and maybe a snack for her brother Bob.	It spoons of fresh snow peas, 5 thick slabs of blue cheese and maybe a snack for her brother Bob.
Dataset_noisy_instruct_wav_p232_005.wav	1.62	1.66	0.0	She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.	She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.
Dataset_noisy_instruct_wav_p232_006.wav	20.70	18.77	0.2	When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.	When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.
Dataset_noisy_instruct_wav_p232_007.wav	11.73	11.95	0.3	The rainbow is a division of white light into many beautiful colors.	The rainbow is a division of white light into many beautiful colors.
Dataset_noisy_instruct_wav_p232_009.wav	6.58	6.79	0.0	"Paris is, according to legend, a boiling pot of gold at one end.	"Paris is, according to legend, a boiling pot of gold at one end.
Dataset_noisy_instruct_wav_p232_010.wav	0.75	0.82	0.2	People look but no one ever finds it.	People look but no one ever finds it.

Table: SNR, WER and Transcriptions for Set 1

7.2. Set 2 Evaluation

Set 2 - Transcriptions

filename	WER (Noisy→Denoised)	Transcript (Noisy)	Transcript (Denoised)
bus.wav	0.35	Grates are expected to remain at those levels, foresee a little higher this week than the Treasury Department's quarterly auction.	Grates are expected to remain at those levels to renovate the little higher this week The Treasury Department's quarterly auction
cafe.wav	0.533	Earlier, GM Hughes had first quarter profit of 160.2million?7 cents a share.	But your earlier, GMFU's head first quarter profit of $160.2 million. We're 77 cents a share.
ped.wav	0.714	sources say at least two bidders had some doubts about city course performance numbers.	Scores say at least two bidders have sugar doubt about the beautiful ones in the place.
street.wav	0.111	Base rates are the benchmark for commercial Indian pressure.	Base rates are the benchmark for commercial language pressure.

Table: Transcription and WER Comparison for Set 2

7.3. MOS Ratings

Listener Ratings: [5, 4, 5, 3, 4]
Mean Opinion Score (MOS): **4.2 / 5**

8. Discussion and Challenges

8.1. Trade-offs:

Wiener filtering preserved most speech quality but had limitations in extreme noise conditions.

8.2. Challenges:

- Division by zero in low-energy frames (fixed via noise floor)
- FP16 warnings from Whisper on CPU
- Long transcript display in tables (handled using wrapping)
- Spectrogram and table rendering required custom scaling and text wrapping in matplotlib
- No ground-truth for Set 2 made true WER estimation impossible; used noisy vs denoised comparison instead

- Whisper on CPU introduced latency and format warnings due to lack of GPU acceleration
- Demucs discarded due to version incompatibility and unstable output paths
- Figure placement in LaTeX required float package and [H] tags to force inline placement
- Manual remapping of file paths was needed to match Overleaf folder structure

9. Code and Outputs

- Colab code used for this analysis is available at: https://colab.research.google.com/drive/1EVBeuolTpQkV4Whk8mhA57dYh4e3Jq17
- Denoised audio files are available at: https://drive.google.com/drive/folders/1pSEYip2EK-vc6u9ZW7qIihW4Iq-MX6uu
- Transcription and other results are available at:
  - **Set 1** https://drive.google.com/file/d/1n-2BRjqfxaetuQTPLFsH66MeGCVm-kMy
  - **Set 2** https://drive.google.com/file/d/1nTDYp7NQlAaNjlvacnUZySG58IXpxLxI
- Set 1 spectrograms and visualizations are available at: https://drive.google.com/drive/folders/1T7Ma9XPU51b7fflf5Hd5GUpoYiq5nMgJ
- Set 2 spectrograms and visualizations are available at: https://drive.google.com/drive/folders/1IDnfm6DloHpbkhlX_sKYLebcIaeDkJ9H

10. Conclusion

This study demonstrates an effective pipeline for speech enhancement using Wiener filtering and Whisper-based ASR. Despite its simplicity, the Wiener filter successfully improved intelligibility in most noisy recordings without requiring deep learning models or GPU resources.

Transcription quality was evaluated both objectively (via SNR, WER) and subjectively (via MOS), showing meaningful improvement in clarity. Spectrogram comparisons reinforced the effectiveness of the denoising approach.

While the method had limitations under heavy noise or low-energy speech, it proved robust for moderate environments. The entire pipeline, including audio, code and results, has been made publicly accessible for transparency and reproducibility.

Future work can explore hybrid approaches using traditional filters as pre-processors for neural models like Demucs or RNNoise.