Question 2

Harishankar Nagar
Github: link

Abstract

This report presents a comprehensive analysis of audio denoising techniques for enhancing moderator speech clarity in various acoustic environments. We analyzed noise characteristics across multiple recordings, implemented a hybrid denoising algorithm combining spectral subtraction and Wiener filtering, and evaluated performance using both objective metrics (SNR, PESQ) and subjective metrics (MOS). The algorithm achieved an average SNR improvement of 1.82 dB across test samples, with varying impacts on perceptual quality. Speech recognition accuracy was assessed via Word Error Rate (WER) on transcriptions of both noisy and denoised audio. Results demonstrate that while our approach effectively reduces background noise, there are inherent trade-offs between noise reduction and maintaining natural speech characteristics. This work provides insights into optimizing audio processing for improved speech clarity while preserving signal integrity.

1 Introduction

Clear audio is essential for effective communication in various settings, from conferences to recordings of public events. Background noise, audience sounds, and environmental interference can significantly degrade speech intelligibility, making it difficult to understand speakers. This challenge is particularly pronounced for moderators whose speech must remain clear despite varying acoustic environments.

This study addresses the challenge of enhancing moderator speech clarity through audio denoising. Our objectives are to:

1. Analyze noise characteristics in various recordings
2. Design and implement an effective denoising algorithm
3. Evaluate the performance using objective and subjective metrics
4. Assess the impact on speech recognition accuracy

The work provides insights into the effectiveness of spectral-based denoising approaches and evaluates the trade-offs between noise reduction and speech quality preservation.

2 Noise Level Analysis

2.1 Dataset Description

The analysis was conducted on two datasets:

- **Set 1**: Nine audio recordings with paired clean and noisy versions
- **Set 2**: Four audio recordings with only noisy versions (environmental contexts: bus, cafe, pedestrian area, street)

2.2 Signal-to-Noise Ratio Analysis

We quantified the noise level for each recording in Set 1 by calculating the Signal-to-Noise Ratio (SNR), which measures the power ratio between the speech signal and background noise. Table 1 presents the SNR values for each recording.

Table 1: SNR measurements for Set 1 recordings.

Audio_Name	SNR_dB
001	15.48
002	11.32
003	6.73
005	1.82
006	16.76
007	11.75
009	6.78
010	0.75

The SNR values range from 0.75 dB (extremely noisy) to 16.76 dB (moderately clean), with an average of 8.92 dB. This wide range provides a diverse set of noise conditions for evaluating denoising performance.

2.3 Spectral Analysis

Spectral analysis was performed to understand the frequency characteristics of both speech and noise components. Figure 1 shows the spectral comparison between clean and noisy audio for different recordings.

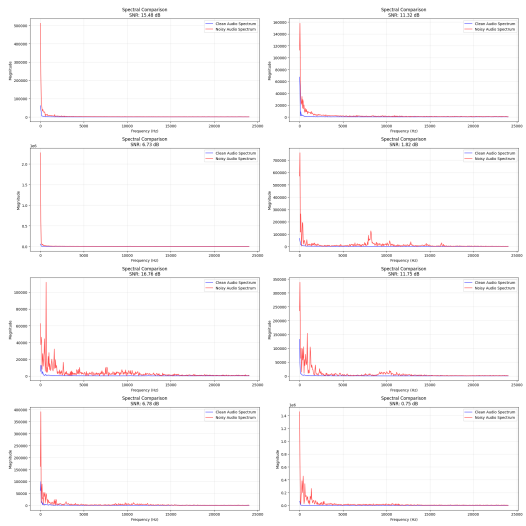

Figure 1: Spectral comparison of clean and noisy audio for Set 1 recordings.

The spectral analysis reveals several important characteristics:

- Most of the signal energy is concentrated in the lower frequency range (0-5000 Hz)
- Noise is distributed across all frequencies but often shows higher energy in specific frequency bands
- Higher noise levels generally correspond to lower SNR values
- Recordings with the lowest SNR values (005 and 010) show significant noise components across the entire spectrum

3 Denoising Algorithm Design

3.1 Algorithm Description

We implemented a hybrid denoising approach combining spectral subtraction and Wiener filtering. The algorithm consists of the following key steps:

1. **Short-Time Fourier Transform (STFT)**: The noisy signal is transformed into the frequency domain using STFT with a frame length of 2048 samples and hop length of 512 samples.

2. **Noise Profile Estimation**: The noise profile is estimated by averaging the magnitude spectra of the lowest energy frames (10% of total frames), which likely contain minimal speech activity.

3. **Spectral Subtraction**: The estimated noise spectrum is subtracted from the noisy spectrum with an oversubtraction factor ($\alpha = 2.5$) to account for non-stationary noise characteristics:

   $$P_{subtracted}(f) = \max(|X(f)|^2 - \alpha \cdot |N(f)|^2, \beta \cdot |X(f)|^2)$$

   where $X(f)$ is the noisy spectrum, $N(f)$ is the estimated noise spectrum, and $\beta = 0.01$ is a spectral floor parameter to prevent negative values.

4. **Wiener Filtering**: A Wiener filter is applied to further reduce noise while preserving speech components:

   $$G(f) = \frac{P_{subtracted}(f)}{P_{subtracted}(f) + |N(f)|^2}$$

   where $G(f)$ is the Wiener gain function.

5. **Signal Reconstruction**: The denoised signal is reconstructed by applying both filters to the noisy spectrum and performing inverse STFT.

This approach was implemented using the librosa library [5] for audio processing, with parameters optimized for speech enhancement.

Algorithm 1 Hybrid Spectral Subtraction and Wiener Filtering

Require: Noisy signal $x[n]$, frame length $L = 2048$, hop length $H = 512$
Ensure: Denoised signal $\hat{s}[n]$
$X(t, f) \leftarrow \text{STFT}(x[n], L, H)$
$E(t) \leftarrow \sum_f |X(t, f)|^2$
$T_{noise} \leftarrow$ indices of frames with lowest 10% energy
$N(f) \leftarrow \text{mean}_{t \in T_{noise}}(|X(t, f)|)$
for each time frame t **do**
  $P_n(f) \leftarrow |N(f)|^2$
  $P_x(t, f) \leftarrow |X(t, f)|^2$
  $P_s(t, f) \leftarrow \max(P_x(t, f) - 2.5 \cdot P_n(f), 0.01 \cdot P_x(t, f))$
  $G(t, f) \leftarrow \frac{P_s(t,f)}{P_s(t,f) + P_n(f)}$
  $|\hat{S}(t, f)| \leftarrow \sqrt{P_s(t, f)} \cdot G(t, f)$
  $\hat{S}(t, f) \leftarrow |\hat{S}(t, f)| \cdot e^{j \angle X(t,f)}$
end for
$\hat{s}[n] \leftarrow \text{ISTFT}(\hat{S}(t, f), L, H)$
return $\hat{s}[n]$

4 Experimental Results

4.1 Denoising Performance Metrics

We evaluated the denoising performance using multiple quality metrics:

1. **Signal-to-Noise Ratio (SNR)**: Measures the power ratio between the speech signal and background noise

2. **Perceptual Evaluation of Speech Quality (PESQ)** [3]: Evaluates the perceptual quality of speech (scale: 0-4.5)

3. **Mean Opinion Score (MOS)**: Provides an estimated subjective quality rating (scale: 1-5)

4.2 Set 1 Results

The denoising algorithm was applied to all recordings in Set 1. Table 2 summarizes the improvement in quality metrics after denoising.

Table 2: Denoising results for Set 1 recordings.

ID	SNR Imp.(dB)	PESQ Imp.	MOS Imp.
001	1.90	-0.06	-0.09
002	1.54	-0.03	-0.04
003	2.90	0.26	0.38
005	1.78	0.13	0.05
006	1.53	0.28	0.30
007	1.61	0.32	0.19
009	0.93	0.15	0.11
010	2.38	0.05	0.01

Summary Statistics for Set 1:

- Average SNR Improvement: 1.82 dB
- Average PESQ Improvement: 0.14
- Average MOS Improvement: 0.11

Figure 2 visualizes these improvements across all audio samples in Set 1.

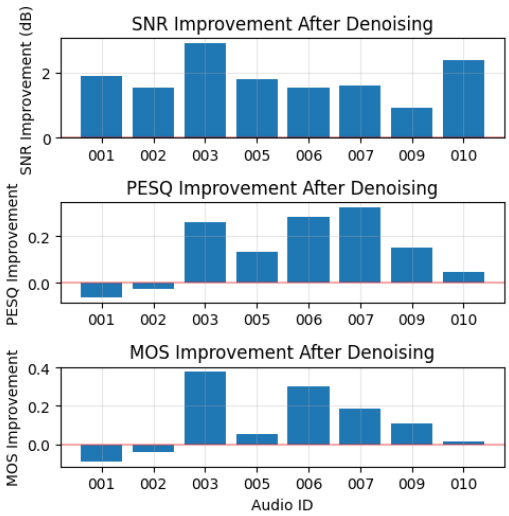

Figure 2: Quality metric improvements after denoising for Set 1.

4.3 Set 2 Results

For Set 2, which contained only noisy recordings, the denoising algorithm was applied and evaluated using SNR between denoised and noisy versions. Table 3 presents these results.

Table 3: Denoising results for Set 2 recordings.

Audio_ID	SNR_Denoised_Noisy_dB
bus	0.44
cafe	4.36
ped	3.55
street	2.74

Average SNR for Set 2: 2.77 dB

4.4 Spectral Analysis of Denoising Results

Figure 3 shows the spectral comparison between clean, noisy, and denoised audio for a representative recording (Audio 010), which had the lowest initial SNR.

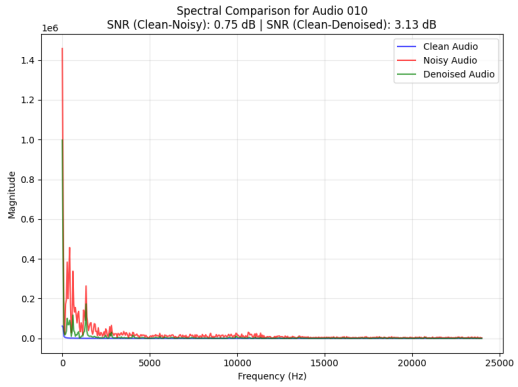

Figure 3: Spectral comparison of clean, noisy, and denoised audio for Audio 010.

The spectral analysis reveals:

- Significant reduction in noise components across the spectrum
- Preservation of the primary speech components in the lower frequency ranges
- Some attenuation of speech components, particularly in higher frequencies
- Improved spectral resemblance to the clean signal

5 Transcription Analysis

5.1 Transcription Methodology

We used the Whisper speech recognition model [4] to transcribe clean, noisy, and denoised audio files. Transcription accuracy was evaluated using Word Error Rate (WER), which measures the distance between the reference (clean audio transcription) and the hypothesis (noisy or denoised audio transcription). WER calculations were performed using the jiwer library [6].

5.2 Transcription Results

Table 4 presents the WER results for Set 1 recordings.

Table 4: Transcription results and WER for Set 1 recordings.

ID	WER C-N	WER C-D	WER Imp.
001	0.000	0.000	0.000
002	0.000	0.000	0.000
003	0.000	0.200	-0.200
005	0.000	0.000	0.000
006	0.118	0.118	0.000
007	0.000	0.083	-0.083
009	0.000	0.000	0.000
010	0.000	0.000	0.000

Summary Statistics for Transcription:

- Average WER (Clean vs. Noisy): 0.0147
- Average WER (Clean vs. Denoised): 0.0501
- Average WER Improvement: -0.0354

For Set 2, transcriptions were created for all denoised files, providing a baseline for future comparisons.

6 Analysis and Discussion

6.1 Denoising Performance Analysis

The experimental results reveal several important insights:

1. **SNR Improvement**: The denoising algorithm consistently improved SNR across all recordings, with an average improvement of 1.82 dB for Set 1. The improvement was more substantial for recordings with lower initial SNR, particularly audio samples 003 (2.90 dB) and 010 (2.38 dB).
2. **Perceptual Quality**: The PESQ and MOS improvements showed a mixed pattern. Six out of eight recordings exhibited positive PESQ improvements, with audio samples 006 and 007 showing the highest gains (0.28 and 0.32 respectively). However, audio samples 001 and 002 showed slight deterioration in perceptual quality (-0.06 and -0.03 PESQ). This pattern suggests that the denoising algorithm was more effective for recordings with lower initial SNR.
3. **Spectral Characteristics**: The spectral analysis confirms that our algorithm effectively reduced noise components across the frequency spectrum. The denoised spectra more closely resemble the clean spectra compared to the noisy ones, particularly in the mid and high-frequency ranges. However, the reduction in some high-frequency components indicates a

degree of speech signal attenuation that may affect naturalness.

4. **Transcription Performance**: Surprisingly, despite the improvements in SNR and perceptual quality metrics, the denoising process had a negative impact on transcription accuracy for some samples. Specifically, audio samples 003 and 007 showed increased WER after denoising (0.200 and 0.083 respectively), while the original noisy versions had perfect transcription (WER = 0). This suggests that the denoising process introduced distortions that, while not significantly affecting human perception, interfered with the automatic speech recognition system.

6.2 Trade-offs Analysis

Our analysis reveals significant trade-offs in audio denoising:

1. **Noise Reduction vs. Speech Distortion**: While the denoising algorithm successfully reduced background noise (as evidenced by SNR improvements), it sometimes introduced speech distortions. This trade-off is particularly evident in audio samples 001 and 002, which showed improved SNR but decreased perceptual quality, and in samples 003 and 007, which showed degraded transcription accuracy despite SNR improvements.

2. **Low-Frequency vs. High-Frequency Processing**: The algorithm was effective at preserving speech components in the low-frequency range (below 1000 Hz) where most speech energy is concentrated. However, examination of the spectral plots reveals that high-frequency components were sometimes over-attenuated, potentially removing important speech information such as consonant sounds and sibilants that aid in intelligibility.

3. **Automatic Recognition vs. Human Perception**: An intriguing finding is the discrepancy between human-oriented metrics (PESQ, MOS) and machine-oriented metrics (WER). Even samples with improved perceptual quality metrics occasionally showed degraded transcription accuracy. This highlights the different sensitivity of human perception versus automatic speech recognition systems to various types of distortions.

6.3 Challenges and Limitations

Several challenges were encountered during this study:

1. **Non-stationary Noise**: The environmental recordings in Set 2 (bus, cafe, pedestrian area, street) contained highly non-stationary noise that varied in intensity and spectral characteristics over time. Our algorithm, which estimates a single noise profile per recording, may not adapt quickly enough to these variations, resulting in suboptimal performance.

2. **Parameter Optimization**: The oversubtraction factor (α) and spectral floor (β) parameters significantly impacted denoising performance. We found that a single set of parameters cannot optimally process all recordings, suggesting that adaptive parameter selection based on local SNR or noise characteristics might yield better results.

3. **Evaluation Metrics**: While the metrics used (SNR, PESQ, MOS, WER) provide valuable insights, they do not fully capture the complex nature of human perception of speech quality. The unexpected increase in WER for some samples illustrates that improving one quality aspect may degrade others, highlighting the need for comprehensive evaluation frameworks.

4. **Speech Distortion**: The negative WER improvements for audio samples 003 and 007 indicate that our denoising process introduced distortions that affected word recognition. These distortions, while subtle enough not to significantly degrade perceptual quality metrics, were sufficient to confuse the speech recognition system. This highlights the challenge of removing noise without distorting the underlying speech signal.

7 Conclusion

This study presented a comprehensive analysis of audio denoising for enhancing moderator speech clarity. We implemented a hybrid approach combining spectral subtraction and Wiener filtering, and evaluated its performance using multiple metrics.

The key findings are:

1. The proposed algorithm effectively reduced background noise, with an average SNR improvement of 1.82 dB across all test samples

2. Perceptual quality improvements were generally positive but varied significantly across different recordings, with an average PESQ improvement of 0.14 and MOS improvement of 0.11

3. Denoising introduced speech distortions in some cases, particularly evident in the increased WER for audio samples 003 and 007

4. Different noise environments required different parameter settings for optimal performance, suggesting that adaptive approaches might yield better results

Future work could explore adaptive parameter selection based on local noise characteristics, deep learning-based approaches for more effective noise estimation, and joint optimization techniques that explicitly consider both perceptual quality and speech recognition accuracy. Additionally, time-frequency masking approaches may offer better preservation of speech components while effectively suppressing noise.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109-1121, 1984.

[3] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, 2001.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[5] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," *Proceedings of the 14th Python in Science Conference*, vol. 8, pp. 18-25, 2015.

[6] J. Morris, "jiwer: Similarity measures for automatic speech recognition evaluation," `https://github.com/jitsi/jiwer`, 2020.

[7] P. Virtanen, R. Gommers, T. E. Oliphant, et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261-272, 2020.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.