I speak Conductor

Aditya Jabade
Department of Applied Mathematics
aj3324@gmail.com

Swapnil Banerjee
Department of Electrical Engineering
sb5041@columbia.edu

Bharvi Acharya
Department of Electrical Engineering
ba2766@columbia.edu

Rishita Yadav
Department of Electrical Engineering
ry2501@columbia.edu

***Abstract*—This project explores audio-signal processing techniques such as Fourier Transform-based high-pass filter, Weiner filter, and audio-enhancement to improve the quality and intelligibility of audio signals, particularly NYC MTA conductor announcements. Mechanical noise was characterized using spectrograms of the moving subway, and a manually designed high-pass filter was implemented to mitigate its effects. A Wiener filter was utilized to suppress additional non-mechanical noise. Audio amplification and de-reverberation techniques were tested to enhance the quality of noise-filtered audio outputs. Performance was evaluated subjectively by listening to the audio outputs of the cascaded noise filter and audio-enhancement system. It was found that while the mechanical noise is mitigated considerably by the high-pass filter, it results in a slight degradation of the conductor's speech. The non-mechanical noise remains persistent even after applying the Wiener filter and audio enhancement techniques.**

***Index Terms*—NYC MTA, unintelligible audio, noise filtering, noise cancellation, speech enhancement, dereverberation**

I. Introduction

Anyone who has traveled in the New York City subway can confirm that a persistent issue impacting rider experience is the unintelligible nature of many conductor announcements. Poor audio quality, background noise, inconsistent volume, and unclear speech make it difficult for passengers to understand critical information, such as service changes, delays, or emergency instructions. This project proposes a solution to enhance the clarity and intelligibility of MTA conductor announcements by leveraging noise-cancellation techniques and audio enhancement, ensuring that passengers receive clear and accurate information. Although the current scope is limited to post-processing a pre-recorded audio snippet, real-time implementation of this algorithmic workflow should lead to a safer, more informed, and less stressful experience for MTA commuters. This project has been partially motivated by the fictional character of Lily Aldrin from the American sitcom 'How I Met Your Mother' who claims to speak conductor language.

II. Technical Approach

To implement the solution proposed in Section I, a systematic procedure was followed. This process involved formulating a model for the problem and the corresponding solution methodology, outlining the key assumptions made and their implications for the project's scope, creating an appropriate dataset, and conducting experiments to analyze the results based on the proposed methodology.

A. Modeling

The problem described in the Section I is modeled in Fig.1. The conductor's speech at the source (speaker) is infused with undesired audio signals such as the mechanical noise of the moving subway, passenger chatter, and other unidentified sources. These disturbances result in unintelligible announcements at the output (the listener's ear) Fig. 2 describes the

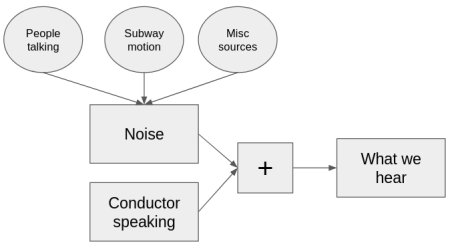

Fig. 1. Block-diagram of the problem

proposed solution with 2 blocks A and B at the output. Block A filters out the noise contaminating the conductor's speech signal, while Block B further enhances the filtered signal by eliminating any residual echo or reverberation. This cascaded system aims to perfectly reproduce the conductor's speech through the speakers in a specific subway coach. [1] [2]

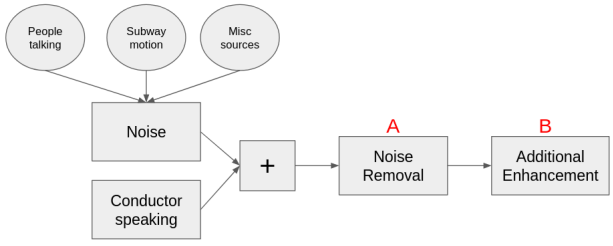

Fig. 2. Block-diagram of the proposed solution

B. Assumptions

The following points describe the assumptions made in the process of solving the problem modeled in Section II-A and its resulting implications:

1) The speaker in the listener's coach is regarded as the source of the conductor's speech signal. Consequently, any signal contamination occurring during the transmission from the conductor's coach to the speaker in the listener's coach is beyond this project's scope.
2) An iPhone recording represents the listener's perspective. Thus, any inconsistencies between the output signal heard by the listener directly and the iPhone recording are also beyond the scope.

C. Data

The team recorded audio snippets while traveling in NYC subways through various routes from Columbia University to Penn Station. Additionally, a couple of recordings available on YouTube were extracted and incorporated into the dataset to create a more comprehensive and diverse set of audio samples for evaluation. To ensure variability, recordings were made from different locations within a coach, at different times of the day, and across various routes, capturing various acoustic conditions.

III. Experiments

Noise filtering and audio enhancement techniques were designed and tested for each of the respective blocks A and B detailed in the Section II-A.

A. Noise Filtering

The two primary sources of noise identified in the Section II-A were passenger chatter and the motion of the subway. It was speculated that owing to the nature of these sources they would exhibit distinctly different frequency characteristics, with passenger chatter dominating the mid-frequency range, while mechanical vibrations resulting from the subway motion would be more prominent in the lower frequencies. Thus separate methods were chosen to deal with the two sources.

1) Subway Motion (Mechanical-Noise): It was anticipated that this noise would predominantly occur in the low-frequency range. To test this hypothesis, audio recordings of the moving subway were collected and analyzed using a spectrogram [8] [5] [4]. The spectrograms in Fig. 3 and 4 illustrate the frequency composition of the mechanical noise. The dark bands (representing maximal amplitude) are observed primarily in the low-frequency region, confirming the hypothesis. Furthermore, the spectrograms in Fig. 5 and Fig. 6, which include audio segments of people speaking in the moving subway, exhibit additional (mildly) dominant components in the mid-frequency range. Even then, the low-frequency dark bands, indicative of mechanical (subway) noise, remain evident.

Thus, a high-pass filter was deemed suitable to eliminate mechanical noise. The threshold frequency was manually set

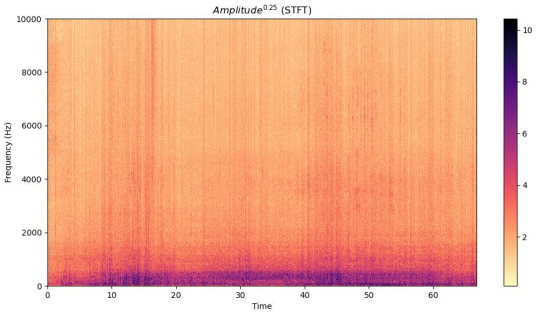

Fig. 3. 72th St

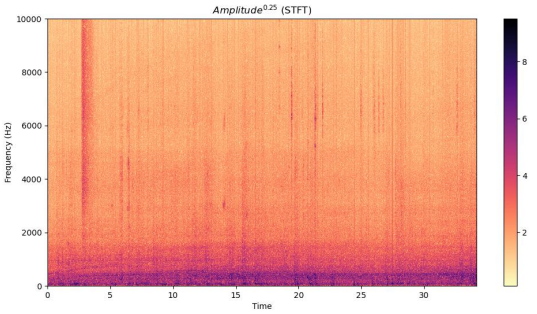

Fig. 4. Sala-Thai (74th St)

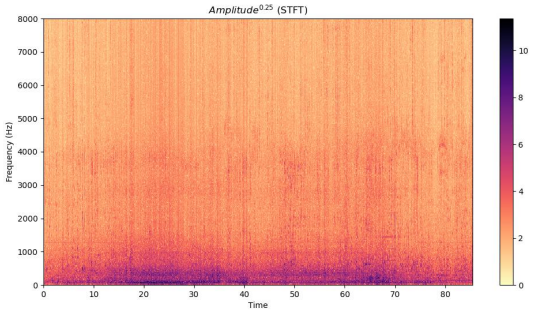

Fig. 5. ColumbiaUniversity-I

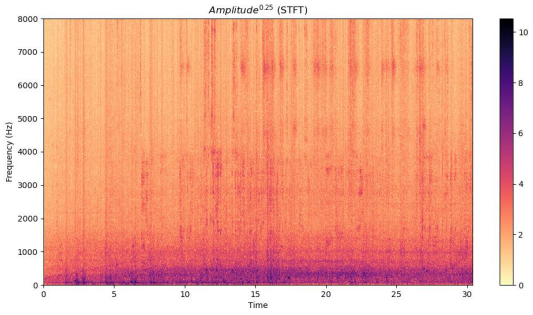

Fig. 6. ColumbiaUniversity-II

at 600 Hz, examining the spectrograms of audio recordin from the dataset. A fourth-order digital high-pass Butterwor filter, commonly used in audio processing, was designed usir SciPy's signal processing toolbox. Audio recordings from th dataset containing the conductor's voice were processed analyze the effectiveness of the filtering technique. Figs. and 10 depict the spectrograms of two such audio recordin from the dataset. Figs. 9 and 10 respectively denote th amplitude comparisons of the original and filtered signa The diminished amplitude in the filtered signal signifies th removal of mechanical noise, i.e. frequency components belo 600 Hz. A detailed discussion of the results is presented Section IV.

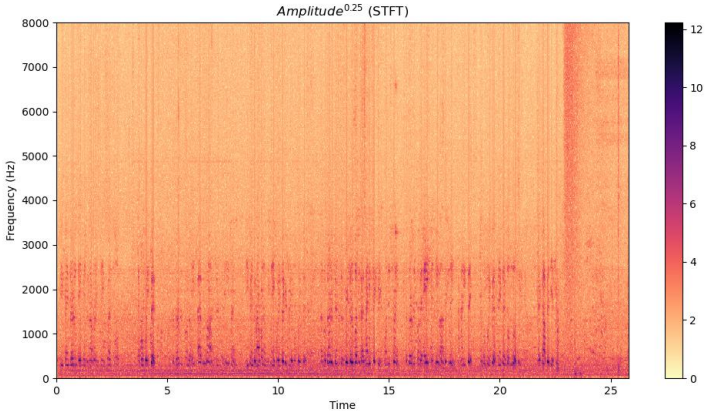

Fig. 7. Spectrogram: Penn Station-I

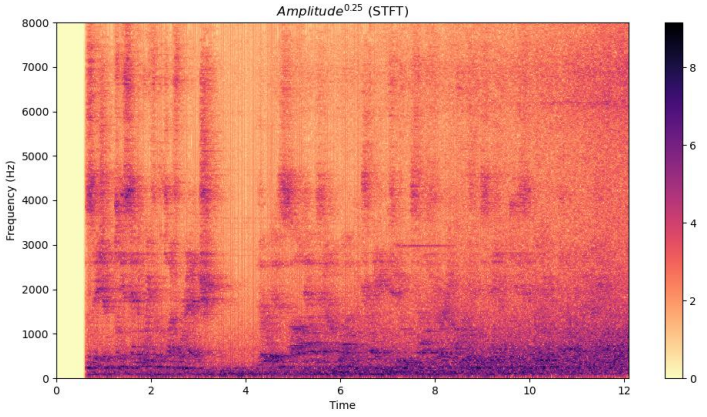

Fig. 8. Spectrogram: YouTube-I

2) People chatter and other miscellaneous sources (Non-Mechanical-Noise): Unlike mechanical noise described in Section III-A1, characterizing the non-mechanical noise consisting primarily of passenger chatter purely based on its frequency components is challenging due to its variability, requiring additional temporal and contextual analysis for accurate characterization. Furthermore, the same could be used to

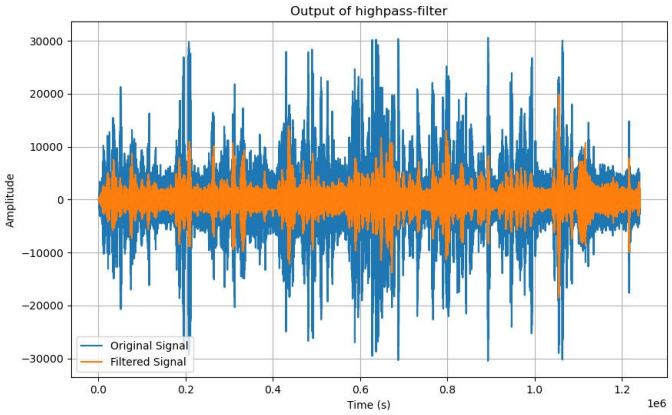

Fig. 9. High-pass filter output (Penn Station-I)

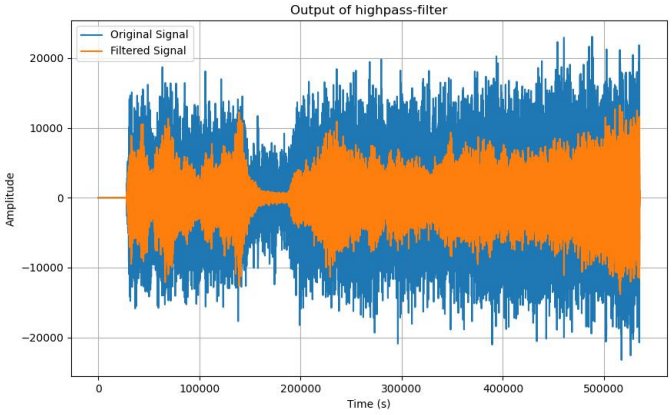

Fig. 10. High-pass filter output (YouTube-I)

identify the conductor's speech if such a characterization was possible. This approach has not been explored in this work.

To reduce such noise, a Wiener filter was applied to the audio signal after it passed through a high-pass filter. Unlike other filters, the Wiener filter is effective in estimating both the noise and signal power spectral densities from a given audio input. Its adaptive nature allows it to minimize the error between the filtered signal and the original signal. For this implementation, the filter was designed with a window size of 11 using SciPy's signal processing toolbox [10]. The window size was determined after a few parameter sweeps with the objective being preserving the 'quality' of conductor-speech. Figs. 11 and 12 depicts the amplitude comparisons of the original and filtered signal for two such audio recordings from the dataset. It must be noted that the original signal in this case is the output of the high-pass filter designed in Section III-A1. The observed marginal reduction of amplitude in the filtered signal signifies the removal of certain frequency components by the wiener filter. A detailed discussion of this is presented in Section IV.

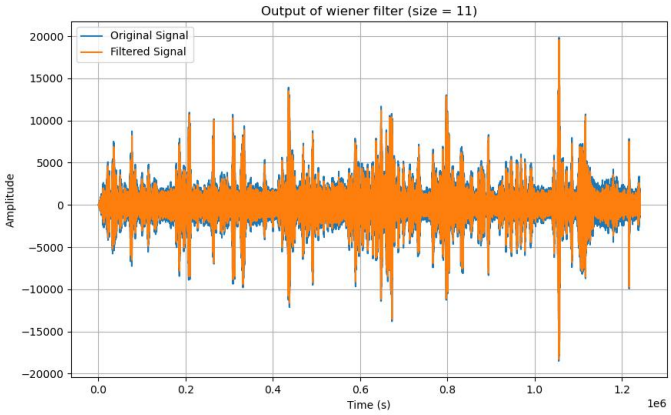

Fig. 11. Wiener filter output (Penn Station-I)

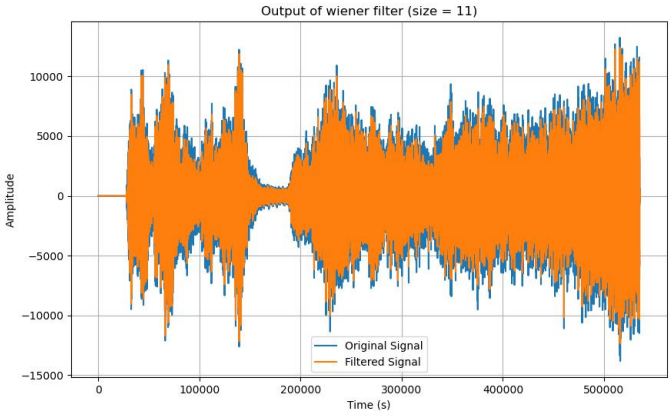

Fig. 12. Wiener filter output (YouTube-I)

Implementation of a Wiener filter was also attempted using AstroML library. However, this approach had several limitations - the noise signal estimated was incorrect and hence resulted in poor filtering. Fig. 13. depicts the power spectral densities (PSDs) of the audio signal as estimated by the AstroML Library [20]. The plot indicates significant noise contributions that interfere with the quality of the signal. Additionally, the noise PSD appears to be constant throughout the frequency spectrum, which is incorrect. The inaccurate noise estimation by the AstroML library resulted in an incorrect filtering attempt as shown in Fig. 14.

Another technique that was implemented for comparison with Wiener Filter was the Savitsky-Golay Filter. [9] A filter of order 4 and window size of 25 was applied on the signal after high-pass. Fig. 15. depicts the the amplitude comparisons of the original and filtered signal. Although the Savitzky-Golay filter showed promising results, the Wiener filter showed better results for the audio files that we used in this analysis.

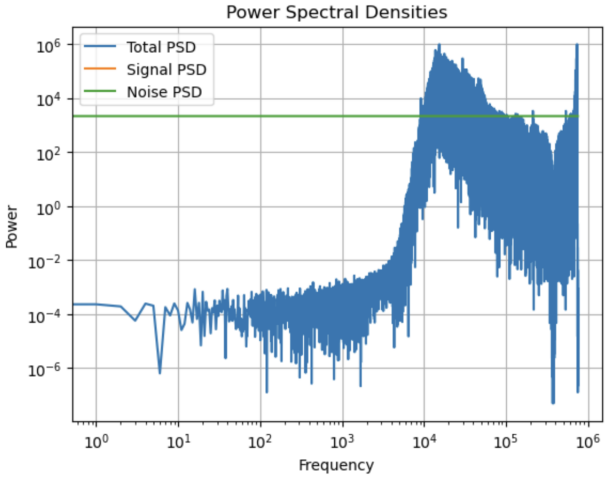

Fig. 13. AstroML Wiener Filter PSD estimation

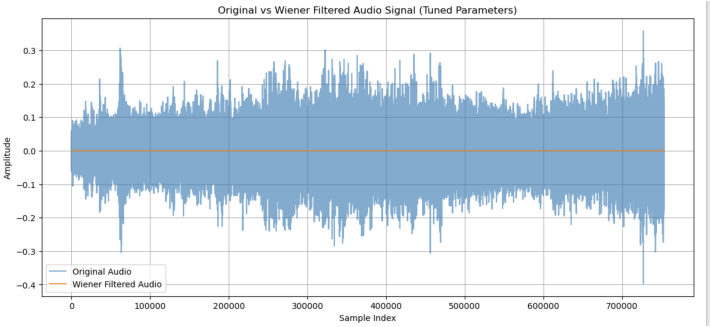

Fig. 14. AstroML Wiener Filter Output

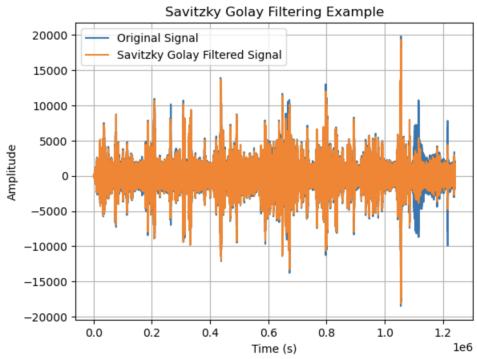

Fig. 15. Savitsky-Golay Filter Output

B. Speech enhancement

It was observed that the audio outputs of the noise-filtering experiments carried out in Sections III-A1 and III-A2 exhibited a noticeable loss in the conductor-speech volume. Furthermore, phenomena such as echo and reverberations remained persistent, as the noise-filtering block was not designed to address these effects. To mitigate these problems, the following experiments were conducted:

1) Audio amplification: A basic audio amplification block was designed to scale the frequency magnitudes of the high-

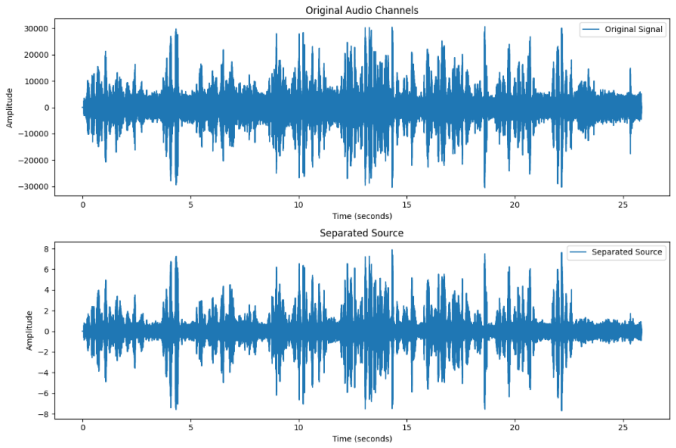

Fig. 16. Amplitude-Time plot of the original signal and a separated source extracted using ICA

pass filtered audio output. While this restored the volume content of the conductor-speech without the mechanical noise, additional undesired components were magnified resulting in sharp blips within the audio segment.

2) Audio dereverberation: Speech quality can be degraded by reverberations depending on the surrounding environment. Even the problem discussed in this project, reverberation hampers the clarity of conductor-speech. To try to mitigate this effect a short exercise was conducted: Inverse STFT was utilized to identify reverberation artifacts, suppress their effect, and consequently generate a dereverberated audio signal. While this approach allowed us to identify and suppress reverberation artifacts to a certain extent, further work is required to fully address the challenges of dereverberation in subway environments.

3) Independent Component Analysis (ICA): Independent Component Analysis (ICA) separates multivariate signals into statistically independent components, often for blind source separation. The FastICA algorithm [16] from the `scikit-learn` library was used to separate the conductor's voice from background noise. Prior to applying ICA, the audio was preprocessed with bandpass filtering [17], [6] to isolate the human speech frequency range (300–3400 Hz). While effective to a certain extent as seen in figure 16, ICA's sensitivity to initialization and its reliance on multiple channels presented challenges for further optimization.

IV. Discussion

A. Results

- **Mechanical noise**: The high-pass filter designed in Section III-A1 was able to suppress the mechanical noise of the moving subway considerably. However, it resulted in a slight degradation of the conductor-speech quality. A plausible explanation for this is that the conductor-speech has non-trivial frequency components even below the manually set threshold of 600 Hz. The high-pass filter in turn removed these components as well.
- **Non-mechanical noise**: The Wiener filter implemented in Section III-A2 was not able to characterize this noise well even after repeated parameter sweeps on the filter size and hence the output showed incremental improvement over the high-pass filter output. One of the possibilities could be that the non-mechanical noise is non-additive (maybe multiplicative) which the Wiener filter is unable to suppress. However in-depth experiments must be carried out to make a definitive claim.
- **Audio-enhancement**: Elementary experiments for dereverberation and source separation were carried out in Section III-B. Looking at the audio outputs, no definitive claim could be made regarding the efficacy of these methods. An in-depth study is needed to design better audio-enhancement methods to improve the noise-filtered output.

Acknowledgment

We would like to express our heartfelt gratitude to Prof. John Wright for his invaluable guidance and support throughout this project. His expertise and encouragement were instrumental in shaping our understanding and approach to the challenges we encountered. We are also thankful to the teaching assistants for their patience, insights, and timely feedback, which greatly enhanced our learning experience. Their dedication ensured that we stayed on track and effectively overcame both technical and conceptual hurdles. Lastly, we extend our heartfelt appreciation to our team members for their collaboration, creativity, and dedication. Working together made this journey both rewarding and enjoyable.

References

[1] *Medium*. (n.d.). Medium. https://ujangriswanto08.medium.com/how-independent-component-analysis-can-enhance-speech-separation-90f9d5185a33.
[2] *Medium*. (n.d.-b). Medium. https://ankurdhuriya.medium.com/audio-enhancement-and-denoising-methods-3644f0cad85b.
[3] ChannelName, "Video Title," YouTube, Dec. 2024. Available at: https://www.youtube.com/watch?v=mQzJ_1Rj6qo.
[4] Aalto University, "Spectrogram and the STFT." Available at: https://speechprocessingbook.aalto.fi/Representations/Spectrogram_and_the_STFT.html. Accessed: Dec. 19, 2024.
[5] Ong Zhi Xuan, "Exploring the Short-Time Fourier Transform: Analyzing Time-Varying Audio Signals." Medium. Available at: https://medium.com/@ongzhixuan/exploring-the-short-time-fourier-transform-analyzing-time-varying-audio-signals-98157d1b9a12. Accessed: Dec. 19, 2024.
[6] Carsten Ak, "Independent Component Analysis." GitHub. Available at: https://github.com/akcarsten/Independent_Component_Analysis. Accessed: Dec. 19, 2024.
[7] MathWorks, "LPC Analysis and Synthesis of Speech." Available at: https://www.mathworks.com/help/dsp/ug/lpc-analysis-and-synthesis-of-speech.html. Accessed: Dec. 19, 2024.
[8] Ursinus College, "Module 11: Video 1." CS372 Course Modules. Available at: https://ursinus-cs372-s2023.github.io/Modules/Module11/Video1. Accessed: Dec. 19, 2024.
[9] Medium, "Introduction to the Savitzky-Golay Filter: A Comprehensive Guide Using Python," Available at: https://medium.com/pythoneers/introduction-to-the-savitzky-golay-filter-a-comprehensive-guide-using-python-b2dd07
[10] SciPy Documentation, "Wiener Filter." Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.wiener.html. Accessed: Dec. 19, 2024.
[11] CloudDevs. (2023, December 25). *10 Python Libraries for audio processing*. https://clouddevs.com/python/libraries-for-audio-processing/

[12] Grindlay, G. (2008). *Blind dereverberation of audio signals*. https://www.ee.columbia.edu/g̃rindlay/classes/E4810/dereverb/report_final.pdf
[13] *DDSP: Differentiable Digital Signal Processing*. (2020, January 15). Magenta. https://magenta.tensorflow.org/ddsp
[14] Mpariente. (n.d.). *GitHub - mpariente/pystoi: Python implementation of the Short Term Objective Intelligibility measure*. GitHub. https://github.com/mpariente/pystoi
[15] *Use Stoi to measure intelligibility of noisy speech*. MathWorks. (n.d.). https://www.mathworks.com/help/audio/ref/stoi.html
[16] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[17] S. W. Smith, *Digital signal processing: A practical guide for engineers and scientists*. Elsevier, 2011.
[18] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis—Theory and applications," *International Journal of Imaging Systems and Technology*, vol. 9, no. 4, pp. 195–200, 1999.
[19] A. J. Bell and T. J. Sejnowski, "Information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
[20] AstroML Documentation, "Wiener Filter." Available at: https://www.astroml.org/modules/generated/astroML.filters.wiener_filter.html.