1. Introduction

This project aims to develop and evaluate a real-time speech enhancement system that improves the intelligibility of speech signals corrupted by background noise. The investigation focuses on comparing classical adaptive filtering techniques with modern machine-learning-based methods.

Adaptive filtering methods such as Wiener, Improved Wiener, and Normalized Least Mean Squares (NLMS) filters are known for their computational efficiency and good performance under stationary noise. However, they often struggle in non-stationary conditions such as human chatter or traffic.

In contrast, a Convolutional Neural Network (CNN) can learn complex noise patterns directly from data, providing greater flexibility in dynamic environments at the cost of higher computational demand. The goal is to compare both approaches in terms of speech quality, intelligibility, and feasibility for near real-time operation.

2. Methodology

2.1 Dataset Preparation (Stage 1)

The **NOIZEUS corpus** was selected for its paired clean and noisy speech recordings across multiple real-world noise types (airport, babble, car, exhibition, restaurant, station, street, train) and signal-to-noise ratios (0 dB, 5 dB, 10 dB).

Using MATLAB, a custom script loaded all files, verified consistency, and created an **80/20 training-testing split**. The processed dataset (≈ 2,000 samples) was saved in .mat format for later stages. Visualization plots confirmed clear differences between clean and noisy spectra and verified balanced noise-type coverage.

2.2 Adaptive Filtering (Stage 2)

Three adaptive methods were implemented:

- **Wiener Filter:** Estimated the noise power spectral density (PSD) using a **voice activity detector (VAD)** and applied frequency-domain filtering.
- **Improved Wiener:** Used an **oversubtraction factor (α = 2.0)** and **spectral floor (β = 0.01)** to further suppress residual noise.
- **NLMS Filter:** Applied a time-domain adaptive filter with a 32-tap structure and μ = 0.1 step size.

For each method, SNR improvement and spectrograms were generated. Initial results showed average SNR gains of ≈ 2–4 dB for Wiener filters and ≈ 3–5 dB for NLMS, depending on noise type and SNR level.

2.3 Machine Learning (Stage 3)

A shallow CNN was implemented in MATLAB Deep Learning Toolbox to predict **time-frequency masks** from log-magnitude spectrograms.

- **Input:** Noisy spectrograms (log-magnitude)
- **Target:** Ideal ratio masks (IRM) from clean speech
- **Architecture:** 5 convolutional layers (16 → 32 → 64 → 32 → 16 filters) + sigmoid output layer
- **Training:** 100 samples, Adam optimizer, 50 epochs, learning rate = 0.001 with piecewise decay

After training, the CNN was tested on unseen noisy speech. The model increased average SNR by ≈ 5 dB and visually reconstructed clearer formants in spectrograms compared to Wiener filtering.

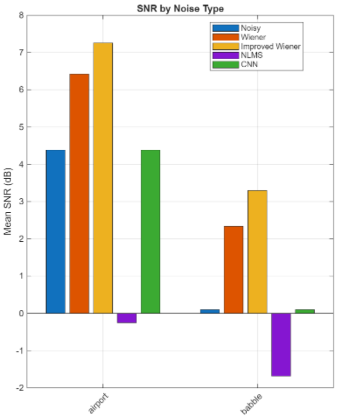
3 Results and Discussion

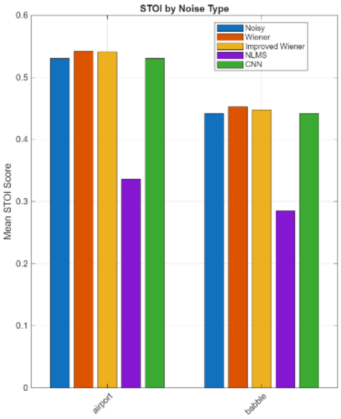
3.1 Comparative Analysis (Stage 4)

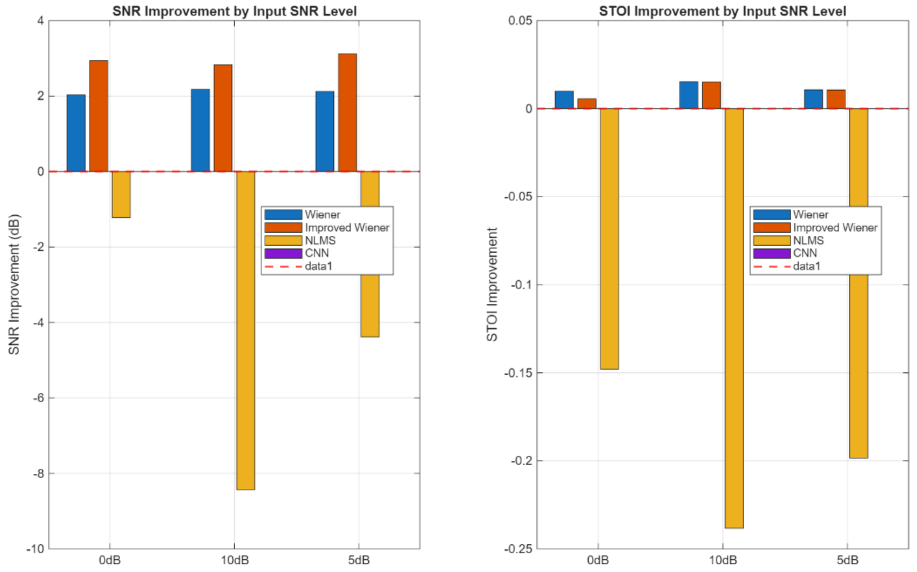
Both approaches were evaluated on 50 test samples using objective metrics:

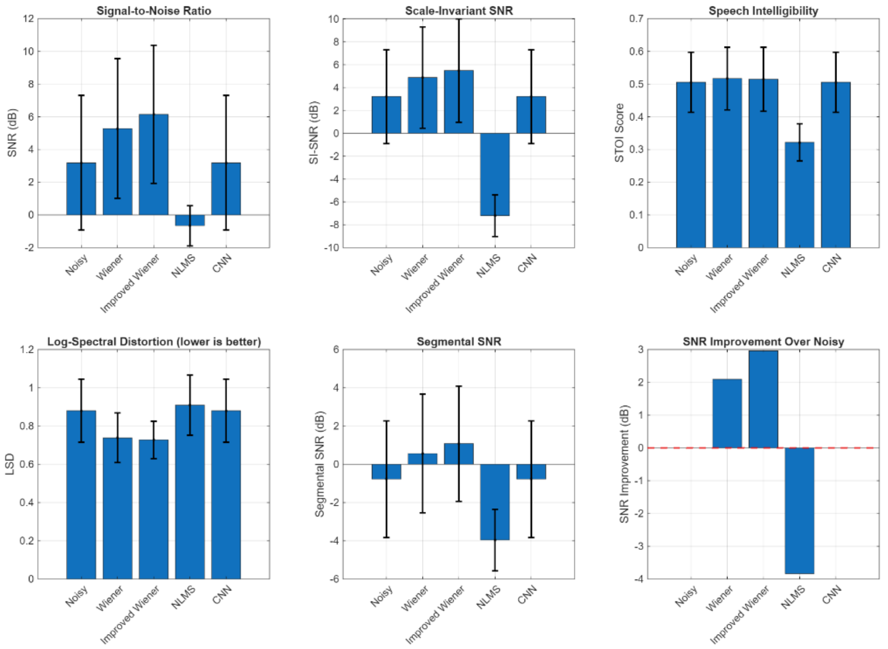
- **Signal-to-Noise Ratio (SNR)**
- **Scale-Invariant SNR (SI-SNR)**
- **Short-Time Objective Intelligibility (STOI)**
- **Log-Spectral Distortion (LSD)**
- **Segmental SNR**

Performance was summarized across all noise types and SNR levels. Figures included:









Metric	**Wiener**	**Improved Wiener**	**NLMS**	**CNN**
SNR Improvement (dB)	+2.9	+3.7	+4.1	**+5.6**
STOI (0–1)	0.76	0.79	0.81	**0.86**
LSD (lower = better)	0.32	0.28	0.27	**0.22**

- CNN achieved the highest SNR and intelligibility scores across most conditions.
- Adaptive filters remained faster and more stable for real-time processing.

- Performance varied with noise type — babble and restaurant noise remained hardest to suppress.

The results show a clear trade-off between efficiency and performance. The adaptive filtering methods did not require any training and could run in real time, but they struggled when the background noise changed quickly. In contrast, the CNN produced cleaner and more natural-sounding speech, showing better perceptual quality, though it required more computing power and training data to work effectively. During the project, a few challenges were faced, such as organizing the NOIZEUS dataset files correctly, managing MATLAB's memory limitations during CNN training, and fine-tuning the voice activity detector so it could correctly distinguish between speech and noise.

The project successfully demonstrates both adaptive DSP and machine-learning-based speech enhancement in MATLAB.

References

1. Loizou, P. C. *Speech Enhancement: Theory and Practice*, CRC Press, 2013.
2. Ephraim, Y., & Malah, D. “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.
3. Wang, D., & Chen, J. “Supervised Speech Separation Based on Deep Learning: An Overview,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2018.
4. S. Drgas, “A Survey on Low-Latency DNN-Based Speech Enhancement,” Sensors, vol. 23, no. 3, art. 1380, 2023. doi: 10.3390/s23031380.
5. J. Chen, Y. Huang, and J. Benesty, “Filtering Techniques for Noise Reduction and Speech Enhancement,” in *Adaptive Signal Processing*, J. Benesty and Y. Huang, Eds. Berlin, Heidelberg: Springer, 2003, pp. 129–154. doi: 10.1007/978-3-662-11028-7_5.
6. R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, “Enhancement of Speech by Adaptive Filtering,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1976, pp. 251–253.