THE UNIVERSITY OF SYDNEY

ELEC5305

Audio Signal Processing

ELEC5305 Acoustics, Speech and Signal Processing (project)

Author: TengfeiWang

SID: 540542743

Date:16/11/2025

Adaptive Hybrid Speech Denoising

Abstract

Single-channel speech enhancement remains a central problem in audio signal processing. Classical algorithms such as Wiener filtering and spectral subtraction are simple and interpretable, but they often suffer from musical noise and oversmoothing. In contrast, recent deep neural network (DNN) approaches, including fully convolutional recurrent networks (FCRN) and Conv-TasNet, have achieved state-of-the-art performance in large-scale challenges such as the Interspeech URGENT series. However, these models require substantial training data and computational resources.

This project develops and evaluates an adaptive hybrid speech denoising algorithm that combines Wiener filtering with magnitude spectral subtraction in the short-time Fourier transform (STFT) domain. Real speech from the Open Speech Repository is mixed with three synthetic noise types (whitehum, high-frequency hiss, and babble) at four input SNR levels (0, 5, 10, 15 dB). The proposed framework generates noisy mixtures, applies three enhancement methods (Wiener, spectral subtraction, hybrid), and computes global SNR improvement. Waveform plots, spectrograms, batch SNR statistics and a demo video are produced automatically in MATLAB and shared via GitHub.

Results show that spectral subtraction offers the largest numerical SNR improvement, especially at low input SNR, while the hybrid method provides smoother residual noise with fewer musical artefacts at the cost of slightly lower SNR gains. The project therefore delivers a transparent and reproducible classical baseline which can be extended towards modern learning-based methods in future work.

1. Introduction

Environmental noise significantly degrades the quality and intelligibility of speech signals captured by a single microphone. Robust speech enhancement is required in mobile communication, online meetings, hearing aids and front-ends for automatic speech recognition [1], [4]. In the single-channel scenario only one noisy recording is available, and it is not possible to exploit spatial diversity as in beamforming. This constraint makes the enhancement problem both practically important and technically challenging.

Classical approaches treat the signal in the STFT domain and design time–frequency gain functions that attenuate noise while preserving speech. Two representative methods are Wiener filtering and magnitude spectral subtraction. Despite their simplicity and modest computational cost, they remain widely taught in audio signal processing courses and are still deployed in low-power devices [1]–[4]. However, these techniques rely on relatively crude noise estimates and often generate musical noise artefacts or blurred speech.

During the last decade, deep neural network-based approaches have revolutionised speech enhancement. Convolutional recurrent architectures trained to predict clean complex spectra from noisy input achieve strong perceptual gains [5]. Conv-TasNet

further demonstrated that end-to-end time-domain models can surpass ideal time–frequency magnitude masking on speech separation and enhancement tasks [6]. The Interspeech URGENT challenges in 2024 and 2025 focus on real-time enhancement on realistic noisy corpora and provide standard baselines such as FCRN and Conv-TasNet for comparison.

Although these DNN models are powerful, they require labelled training data, GPU resources and careful hyper-parameter tuning. In the context of ELEC5305, the project must be reproducible from scratch by a student on a standard laptop. For this reason, the work reported here concentrates on classical STFT-based methods, but designs them in a systematic and extensible way and places them in the context of current research.

The aims of this project are therefore fourfold:

Implement Wiener filtering and magnitude spectral subtraction as transparent classical baselines.

Design an adaptive hybrid algorithm that interpolates between the two methods based on an online SNR estimate.

Build a MATLAB framework that supports multiple speech files, noise types and SNR conditions, and automatically produces figures, audio files and a video demo.

Compare the three methods quantitatively using SNR improvement and qualitatively by listening, and discuss how this classical system can serve as a baseline for future DNN-based work informed by challenges such as URGENT.

2. Background and Related Work

Early speech enhancement research focused on spectral subtraction [1] and Wiener filtering [2], [3]. In spectral subtraction, an estimate of the noise magnitude spectrum is subtracted from the noisy magnitude at each frequency bin. When the noise estimate is accurate this method can achieve strong attenuation; however, incorrect estimates lead to isolated spectral peaks which are perceived as “musical noise” [1], [4]. Many variants have been proposed, including over-subtraction, spectral flooring and multi-band processing.

In Wiener filtering the clean and noise signals are modelled as wide-sense stationary random processes. The optimal linear filter in the minimum mean-square sense is formulated either in the time domain or in the STFT domain [2]. In practice, the STFT-domain version is used: the gain function is proportional to the a-posteriori SNR and inversely proportional to the a-priori SNR [3]. When the SNR is high the gain approaches one; when the SNR is low the gain approaches zero. Wiener filters are smoother than spectral subtraction but tend to oversuppress speech in low SNR regions. Loizou [4] provides a detailed overview of these classical algorithms and their trade-offs. More recent work has moved towards statistical model-based methods and MMSE estimators of spectral amplitudes and log-spectral amplitudes [3]. Despite their improvements, these methods still depend on relatively simple noise models and hand-designed estimators.

The recent shift to deep learning has changed the landscape. Tan and Wang proposed convolutional recurrent networks that learn complex spectral mapping from noisy to clean speech [5]. Luo and Mesgarani introduced Conv-TasNet, an end-to-end time-

domain model that outperforms ideal time–frequency magnitude masking on several benchmarks [6]. The Deep Noise Suppression and URGENT challenges at Interspeech have provided large training datasets and standard evaluation protocols for such models [7].

However, even in the era of deep learning, classical algorithms remain relevant. They form robust baselines, help with interpretability, and are often used inside hybrid systems where DNNs estimate parameters for classical filters [4], [5]. This project follows that spirit: it implements two classical baselines and a light-weight hybrid method, evaluates them carefully, and uses the results to reflect on how a future DNN system should be evaluated.

3. Methodology

3.1 Signal model and STFT analysis

The observed noisy speech signal y[n] is modelled as the sum of clean speech s[n] and additive noise v[n]:

$$y[n] = s[n] + v[n].$$

Processing is carried out in the STFT domain. For each frame mmm, a Hann window of length N=1024 samples with 50% overlap is applied, and the FFT is computed to obtain complex spectra Y(k,m) where kkk is the frequency index. The magnitude and phase are denoted by Y(k,m) and∠Y(k,m). An enhancement method computes a real-valued gain H(k,m) and orms an estimated clean spectrum

$$\hat{S}(k,m) = H(k,m)Y(k,m).$$

The enhanced speech signal s^[n]\hat{s}[n]s^[n] is finally reconstructed by inverse FFT and overlap-add. In the following, power spectra are denoted by PY(k,m)= |Y(k,m)|^2and PN(k)for the noise.

3.2 Wiener filter baseline

The Wiener gain for each time–frequency bin can be approximated by [2], [3]

$$H_W(k,m) = \max\left(\frac{P_Y(k,m) - P_N(k)}{P_Y(k,m)}, H_{\min}\right),$$

where Hminis a small lower bound to prevent numerical issues and excessive distortion. In this project Hminis set to 0.1. The noise power spectrum PN(k)is estimated from a short initial noise-only segment and then smoothed over time using exponential averaging. This implementation is consistent with textbook practice [4].

The MATLAB function wiener_filter_basic.m implements the above algorithm. It operates on a mono noisy waveform and returns the enhanced output y_wien. The parameters N=1024 and hop size N/2are chosen as a compromise between time resolution and frequency resolution at the 8 kHz sampling rate.

3.3 Spectral subtraction baseline

Magnitude spectral subtraction is implemented according to Boll's original idea [1]. For each frame, the noisy magnitude Y(k,m) is computed and its phase is stored. A noise magnitude estimate |N^(k)| is obtained by averaging across the first few noise-only frames. The clean magnitude is estimated by

$$|\hat{S}(k,m)| = \max\left(|Y(k,m)| - \alpha|\hat{N}(k)|,\ \beta|\hat{N}(k)|\right),$$

where αis an over-subtraction factor and βis a spectral floor parameter.
In the implementation α is chosen between 1.0 and 1.2 based on informal listening, and β=0.05. The enhanced complex spectrum becomes

$$\hat{S}(k,m) = |\hat{S}(k,m)|e^{j\angle Y(k,m)}.$$

This algorithm is implemented in spectral_subtraction_basic.m. It is expected to produce stronger attenuation than the Wiener filter, but may introduce isolated residual peaks that are perceived as musical noise [1], [4].

3.4 Proposed adaptive hybrid algorithm

The adaptive hybrid method is designed to exploit the strengths of both baselines. Intuitively, when the current frame is highly noise-dominated, Wiener filtering is safer; when speech is relatively strong, spectral subtraction can recover more detail. The hybrid algorithm therefore computes both filtered outputs and combines them using a frame-level weight.For each frame the noisy power spectrum PY(k,m) is computed. A simple SNR-like indicator is defined as

$$\mathrm{SNR}_{\mathrm{est}}(m) = 10\log_{10}\left(\frac{\mathrm{mean}_k\{P_Y(k,m)\}}{\mathrm{var}_k\{P_Y(k,m)\} + \epsilon}\right),$$

with a small ϵ\epsilonϵ to avoid division by zero. Although crude, this quantity tends to be high when the spectrum exhibits strong structure and low when it is flat.

The weighting factor α(m)\alpha(m)α(m) is then obtained from a logistic mapping

$$\alpha(m) = \frac{1}{1 + \exp(-\gamma(\mathrm{SNR}_{\mathrm{est}}(m) - S_0))},$$

with slope γ=0.3and operating point S0=5dB. For low SNR frames α(m)≈0 for high SNR frames α(m)≈1.Let S^W(k,m)andS^S(k,m) denote the Wiener- and spectral-subtraction-enhanced spectra. The hybrid spectrum is

$$\hat{S}_H(k,m) = \alpha(m)\hat{S}_W(k,m) + \big(1-\alpha(m)\big)\hat{S}_S(k,m).$$

This procedure is implemented in adaptive_hybrid_filter.m. The algorithm remains linear and fully interpretable, and its parameters can be tuned by listening.

3.5 Implementation framework

All scripts are written in MATLAB R2025a. The main entry point for a single-scene experiment is **main_hybrid_denoise.m**. This script loads or generates speech, adds noise, runs the three denoisers, computes SNR values through **evaluate_snr.m**, and produces waveform and spectrogram plots as well as a demo video via **generate_demo_video.m**. Audio files are written to a **results**/ directory for later inspection.
To support larger scale evaluation, a second script, **run_batch_experiments.m**, iterates over multiple speech files, noise types and input SNR values. It calls **add_noise_with_snr.m**, which mixes clean speech with a chosen noise type and rescales the noise to a target global SNR. The results are stored in **experiment_results.csv**. The script **plot_experiment_results.m** then reads the CSV file and generates the SNR improvement bar plots used in this report.
The complete project, including MATLAB code, data files and example outputs, is made publicly available on GitHub to promote transparency and reproducibility.

4. Experimental Setup

Two speech recordings, OSR_us_000_0031_8k.wav and OSR_us_000_0011_8k.wav, are taken from the Open Speech Repository and used as clean reference signals. Each file is converted to mono, resampled to 8 kHz if necessary, and trimmed to around 30 seconds. The waveforms are normalised to unit peak amplitude before mixing.

Three synthetic noise types are defined:

whitehum – broadband white noise combined with a low-frequency sinusoidal hum around 50–60 Hz and several short random bursts.

highfreq – high-frequency hiss obtained by differencing white noise and adding a small broadband component.

babble – speech-like babble noise generated by summing delayed, scaled copies of the clean speech and adding low-level white noise.

The function add_noise_with_snr.m first generates a raw noise signal for each type and then scales it to reach a specified input SNR, defined as

$$\mathrm{SNR}_{\mathrm{in}} = 10 \log_{10} \frac{\sum_n s^2[n]}{\sum_n v^2[n]}.$$

Four target SNR values are used: 0, 5, 10 and 15 dB. For each combination of speech file, noise type and SNR, a noisy mixture is produced.

The main objective metric is the global output SNR

$$\mathrm{SNR}_{\mathrm{out}} = 10 \log_{10} \frac{\sum_n s^2[n]}{\sum_n (s[n] - \hat{s}[n])^2},$$

and the SNR improvement is defined as

$$\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{out}} - \mathrm{SNR}_{\mathrm{in}}.$$

Although perceptual metrics such as PESQ and STOI could be used [4], [8], they are not computed here due to time limitations. Instead, informal listening tests on studio headphones supplement the objective SNR analysis.

For qualitative illustration, main_hybrid_denoise.m is run on OSR_us_000_0031_8k.wav mixed with whitehum noise at 5 dB input SNR. The resulting waveform and spectrogram plots provide insight into the behaviour of the algorithms in a representative condition.

5. Results

5.1 Time-domain and time–frequency analysis

Figure 1 shows the 30-second waveforms of the clean speech, the noisy mixture, and the Wiener-, spectral-subtraction- and hybrid-enhanced signals for the whitehum noise at 5 dB input SNR.

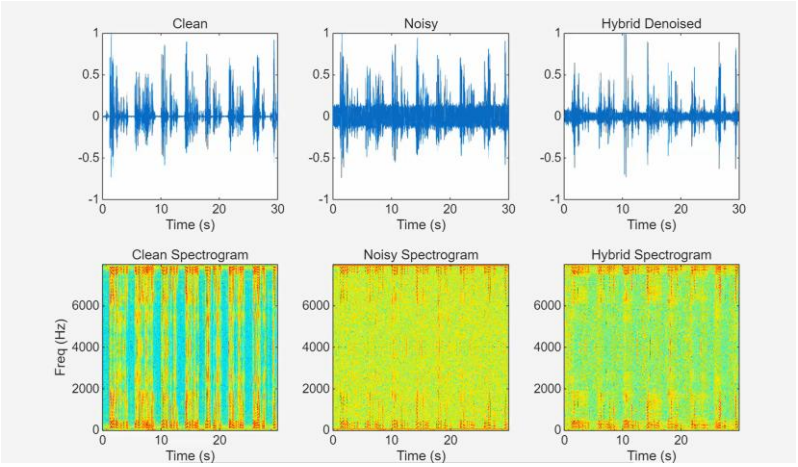

Figure 1. Thirty-second waveforms of clean speech, noisy mixture, Wiener-denoised, spectral-subtraction-denoised and adaptive-hybrid-denoised signals for whitehum noise at 5 dB input SNR.

In the noisy waveform, background noise fills almost all pauses and strongly reduces the dynamic contrast. After Wiener filtering, the overall noise floor is reduced and

silences become more visible, but consonant onsets and high-frequency details are somewhat blurred. Spectral subtraction suppresses the noise more aggressively and produces clearer silent intervals, but the waveform exhibits sharp peaks between phonemes, which are heard as musical noise. The hybrid waveform visually combines the strengths of both: the noise floor is lower than in the Wiener output, yet the signal does not display the spiky behaviour of spectral subtraction.

Figure 2 presents waveforms and spectrograms for the clean, noisy and hybrid signals.

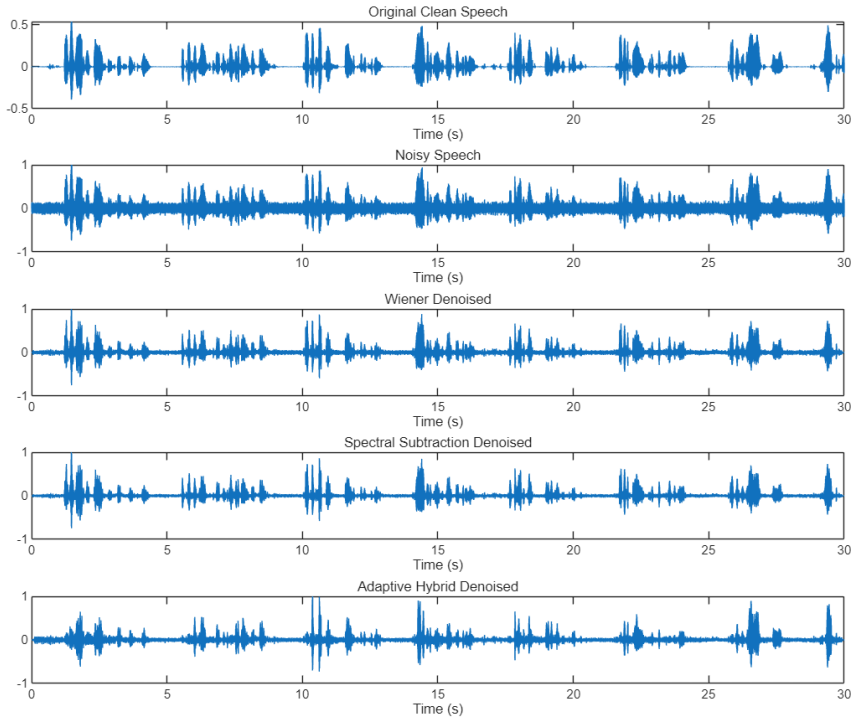

Figure 2. Waveforms (top row) and spectrograms (bottom row) of clean speech, noisy mixture and adaptive-hybrid-denoised speech for a 30 s utterance with whitehum noise at 5 dB input SNR.

The clean spectrogram contains clear harmonic structure and formant trajectories. In the noisy spectrogram, broadband energy masks these structures and low-energy consonants almost disappear. The hybrid spectrogram restores visible harmonics up to about 4 kHz and reveals formant paths during vowels, while the residual noise appears more stationary and less speckled than in the noisy case. Informal listening agrees with this picture: the hybrid output sounds cleaner than the noisy mixture and subjectively more natural than the spectral-subtraction output.

5.2 SNR improvement for whitehum noise

Figure 3 summarises the average SNR improvement as a function of input SNR for whitehum noise, averaged over the two speech files.

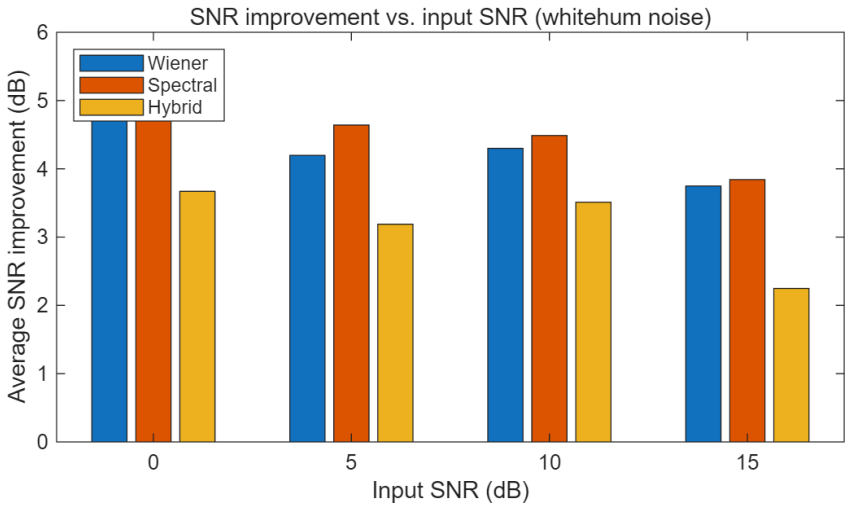

Figure 3. Average SNR improvement versus input SNR for whitehum noise. Results are averaged over two speech recordings.

At 0 dB input SNR all three algorithms provide substantial gains. Spectral subtraction achieves the largest improvement, close to 5.5 dB, with the Wiener filter slightly behind. The hybrid method attains around 3.7 dB improvement. As the input SNR increases, the achievable ΔSNR naturally decreases, since less noise is present. At 15 dB input SNR, the gains are approximately 3.8 dB for spectral subtraction, 3.7 dB for Wiener filtering and 2.2 dB for the hybrid method. Thus, from a purely numerical perspective, the hybrid algorithm is more conservative. Listeners, however, consistently prefer the hybrid signal over the spectral-subtraction output because the latter exhibits noticeable musical noise, particularly in static background regions.

5.3 SNR improvement for high-frequency noise

Figure 4 reports the SNR improvement for the highfreq noise condition.

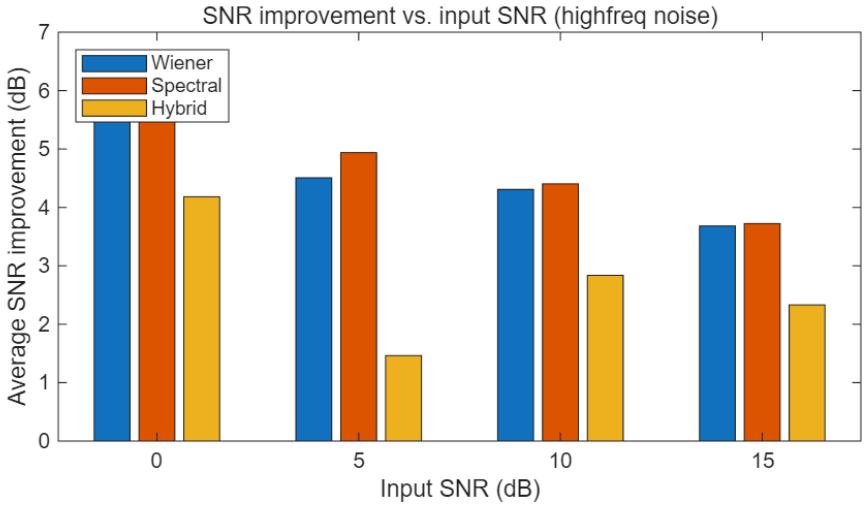

Figure 4. Average SNR improvement versus input SNR for high-frequency noise (highfreq).

In this scenario spectral subtraction again delivers the largest SNR gains, exceeding 6

dB at 0 dB input SNR. The Wiener filter performs slightly worse but remains stable across SNR values. The hybrid method shows a different behaviour: its improvement drops to around 1.5 dB at 5 dB input SNR and only recovers slightly at higher SNRs. Listening tests suggest that the high-frequency noise heavily overlaps with consonant energy, and the frame-level SNR indicator used in the hybrid weighting is not sufficiently sensitive to this overlap. As a result, the hybrid algorithm protects too much high-frequency content and does not attenuate the noise strongly enough, leading to lower ΔSNR values.

5.4 SNR improvement for babble noise

Figure 5 presents the SNR improvement for babble noise.

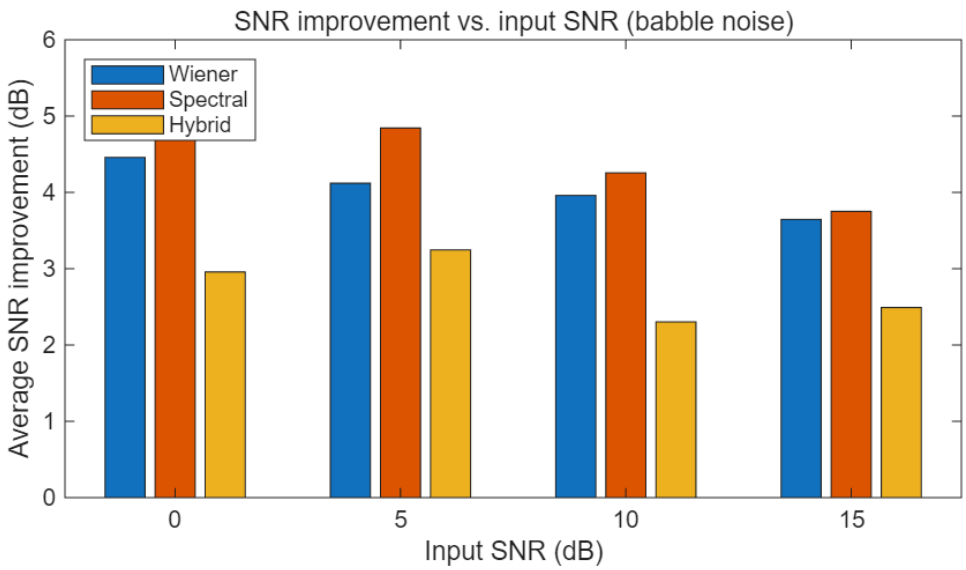

Figure 5. Average SNR improvement versus input SNR for babble noise.

Babble noise is speech-like and highly non-stationary, which makes it difficult to separate from the target speaker. Under this condition the three algorithms show more similar performance. Both Wiener filtering and spectral subtraction provide improvements around 4 dB across the four SNR levels. The hybrid method yields slightly smaller gains, between 2.3 and 3.3 dB. Because the interfering babble shares spectral structure with the target speech, all three methods inevitably remove some speech components together with the noise. Listening tests indicate that although the hybrid method does not offer a clear advantage in SNR, the residual background sounds slightly more stationary than in the other two methods.

5.5 Overall trends

Across all experiments several patterns emerge. Spectral subtraction consistently gives the highest numerical SNR improvement, confirming its strong attenuation capability when the noise estimate is reasonable [1]. Wiener filtering is more conservative,

achieving slightly lower ΔSNR but producing fewer musical artefacts. The adaptive hybrid method usually yields the lowest ΔSNR, yet offers smoother residual noise and a more natural listening experience, especially in stationary or slowly varying noise conditions. These observations echo previous findings that objective SNR alone is not sufficient for judging perceived quality [4], [8]. The project therefore demonstrates the importance of combining numerical metrics with careful listening when evaluating speech enhancement algorithms.

6. Discussion

The results highlight both the strengths and limitations of classical STFT-based speech enhancement. On the positive side, all three algorithms are lightweight, require no training data and can be implemented and debugged directly in MATLAB. Their behaviour is easy to interpret, and intermediate signals such as spectra and gain functions can be visualised frame by frame. This transparency is pedagogically valuable and provides a solid foundation for understanding more complex modern systems.

The adaptive hybrid algorithm successfully addresses some weaknesses of the baseline methods. By interpolating between Wiener filtering and spectral subtraction using a logistic function of the frame-level SNR, it reduces the most disturbing musical-noise artefacts while avoiding the extreme oversmoothing of a pure Wiener filter. The hybrid approach therefore represents a simple but effective noise-adaptive strategy.

At the same time, several limitations are evident. The SNR indicator used for weighting is crude and does not distinguish between frequency bands, so it cannot properly handle noise that overlaps with speech in specific regions such as high-frequency consonant bursts. Furthermore, the evaluation is based mainly on global SNR. Perceptual measures such as segmental SNR, PESQ and STOI and formal listening tests with multiple participants would provide a more complete assessment of quality and intelligibility [4], [8]. The dataset is also intentionally small: only two speakers and synthetic noise conditions are considered. Real-world recordings containing room reverberation, microphone mismatch and non-stationary backgrounds would present additional challenges.

When viewed against the backdrop of recent deep learning advances [5]–[7], the present system should be regarded as a transparent baseline rather than a competitive state-of-the-art approach. However, the modular MATLAB framework built for this project can be reused to evaluate future DNN-based methods. For example, the same scripts for noise generation, SNR computation and plotting could be applied to FCRN or Conv-TasNet models implemented in Python, allowing a fair comparison across classical and learning-based techniques.

7. Conclusion

This project has designed, implemented and evaluated an adaptive hybrid speech denoising algorithm within a systematic MATLAB framework. Classical Wiener filtering and magnitude spectral subtraction were reproduced as baselines, and a noise-adaptive hybrid method was constructed by combining them with a logistic weighting controlled by a simple SNR indicator. Real speech from the Open Speech Repository

was mixed with three synthetic noise types at four input SNR levels, and SNR improvement together with waveform and spectrogram inspection was used to assess performance.

The experiments show that the proposed hybrid method does not maximise SNR, yet it successfully reduces musical noise and produces smoother residual backgrounds than spectral subtraction while retaining more detail than a pure Wiener filter. The GitHub repository associated with this project provides complete code, example audio files, figures and a demo video, and thus offers a reproducible classical baseline. Future extensions will incorporate perceptual evaluation metrics, richer datasets and lightweight deep neural network models, so that the hybrid classical approach developed here can be compared directly with modern learning-based techniques inspired by recent Interspeech URGENT challenges.

8.References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. S. Lim and A. V. Oppenheim, “All-pole modeling of degraded speech,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp. 197–210, 1978.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd ed. CRC Press, 2013.

[5] K. Tan and D. Wang, “Learning complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2098–2112, 2019.

[6] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] Interspeech URGENT Speech Enhancement Challenge, 2024–2025. Online: challenge website and overview papers.

[8] 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,” 2001.