TEAM STRAWHATS

WAVEFORM WARS

REPORT

PROBLEM STATEMENT 1

Acoustic Echo Cancellation Using Adaptive Filtering

Aim

The objective of this work is to design and implement an adaptive filter system that dynamically estimates and cancels acoustic echo in real-time communication scenarios. When a microphone and loudspeaker operate simultaneously in devices such as smart speakers, conferencing systems, or automotive voice assistants, the microphone captures both the desired near-end speech and unwanted echo reflections from the loudspeaker output. This implementation focuses on using the Normalized Least Mean Squares (NLMS) algorithm to model the echo path and suppress these reflections while maintaining robustness during double-talk scenarios where both near-end and far-end speakers are active simultaneously.

Methodology

System Architecture

The acoustic echo cancellation system operates on a frame-by-frame basis, processing audio at a sampling rate of 16 kHz with frame lengths of 256 samples. This configuration balances computational efficiency with real-time processing requirements. The core of the system comprises an adaptive finite impulse response (FIR) filter with 512 taps, providing sufficient length to model typical room impulse responses that include early reflections and some reverberation characteristics.

Signal Model

The system processes three primary signals. The far-end signal x(n) represents the reference audio being played through the loudspeaker, such as the remote speaker's voice in a video call. The microphone captures a composite signal d(n) consisting of the desired near-end speech s(n), the acoustic echo y(n) resulting from the loudspeaker-to-microphone coupling, and ambient noise v(n). Mathematically, this relationship is expressed as d(n) = s(n) + y(n) + v(n).

The echo signal y(n) is generated by convolving the far-end signal with the unknown acoustic echo path h, which represents the impulse response characterizing how sound propagates from the loudspeaker to the microphone through the environment. This path includes direct sound propagation, wall reflections, and other acoustic phenomena specific to the physical setup.

Adaptive Filter Algorithm

The NLMS algorithm serves as the adaptation mechanism for estimating the echo path. At each sample n, the algorithm maintains a filter coefficient vector w and a reference signal buffer x_buf containing the most recent 512 samples of the far-end signal. The echo estimate ŷ(n) is computed as the inner product of the filter coefficients and the reference buffer: ŷ(n) = w^T · x_buf.

The error signal $e(n)$ represents the echo-cancelled output and is calculated as the difference between the microphone signal and the echo estimate: $e(n) = d(n) - \hat{y}(n)$. When the filter accurately models the echo path, this error signal primarily contains the desired near-end speech and noise, with the echo component significantly attenuated.

The filter coefficients are updated using the NLMS update rule:

$$w(n+1) = w(n) + \mu \cdot e(n) \cdot x\_buf / (x\_buf^T \cdot x\_buf + \varepsilon)$$

where $\mu = 0.6$ is the step size controlling convergence speed and tracking capability, and $\varepsilon = 10^{-6}$ is a regularization parameter preventing numerical instability when the reference signal power is low. The normalization by the input signal power ($x\_buf^T \cdot x\_buf$) makes the algorithm robust to variations in signal amplitude and improves convergence characteristics compared to the standard LMS algorithm.

Double-Talk Detection and Control

A critical challenge in acoustic echo cancellation is handling double-talk scenarios where both the near-end user and far-end speaker are talking simultaneously. During double-talk, the microphone signal contains significant near-end speech that should not be treated as echo. Adapting the filter coefficients based on this composite signal would cause filter divergence and degrade echo cancellation performance.

The system implements a simple yet effective double-talk detector using an energy ratio criterion:

$$|d(n)| / (\max(|x\_buf|) + \varepsilon) < \gamma$$

where $\gamma = 0.7$ is the threshold parameter. This detector compares the instantaneous microphone signal amplitude to the maximum reference signal amplitude in the buffer. When the ratio exceeds the threshold, indicating strong near-end activity relative to the far-end signal, filter adaptation is suspended. This prevents the filter from attempting to model the near-end speech as part of the echo path. Adaptation resumes when the ratio falls below the threshold, allowing the filter to track changes in the actual echo path.

Test Scenario Design

To evaluate the system's performance, a controlled simulation generates realistic test conditions. The far-end signal is synthesized as a speech-like modulated tone at 300 Hz with slow amplitude variations at 3 Hz, plus additive Gaussian noise to simulate natural speech characteristics. The acoustic echo path is modeled as a five-tap FIR filter with exponentially decaying coefficients [1.0, 0.5, 0.3, 0.2, 0.1], representing a simplified room impulse response with a strong direct path and several early reflections.

The near-end signal is designed to test both single-talk and double-talk conditions. For the first two seconds, the near-end speaker is silent, creating a pure echo cancellation scenario. After two seconds, near-end speech at 500 Hz with 2 Hz modulation becomes active at reduced amplitude, simulating simultaneous speech from both parties. The complete microphone signal combines the near-end speech, scaled echo (70% amplitude), and low-level measurement noise (5% amplitude).

Performance Metrics

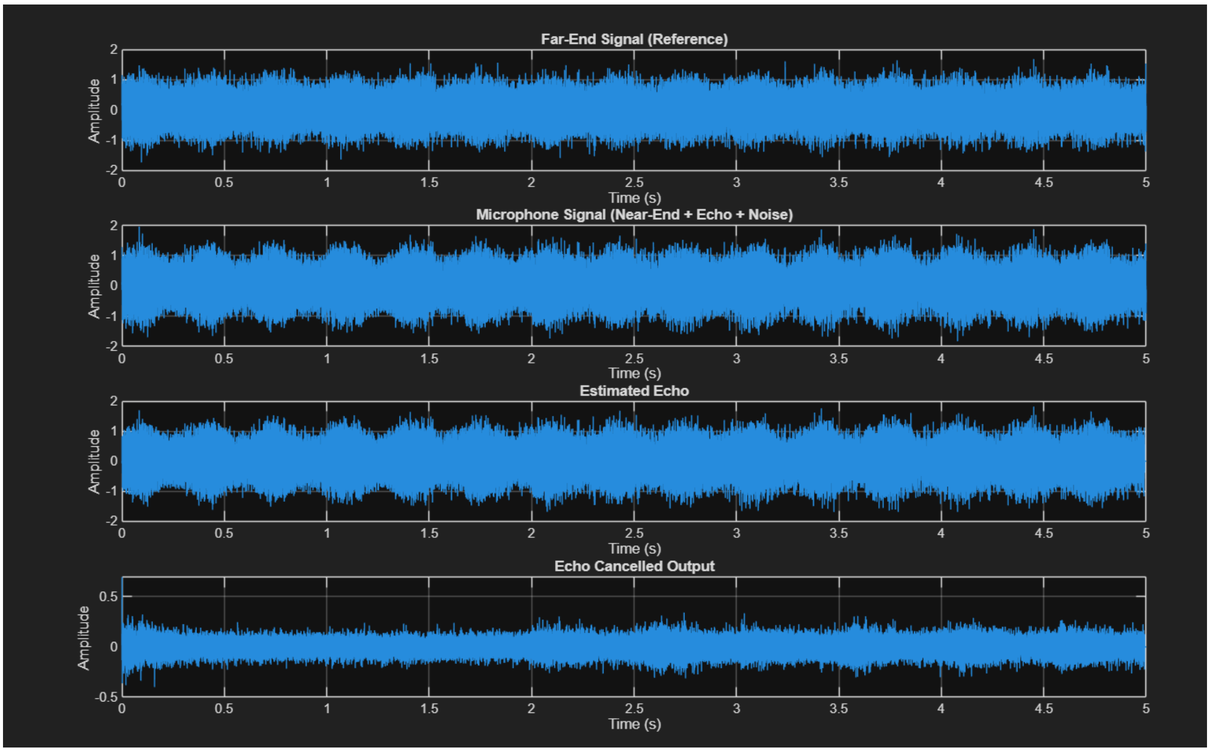
Echo cancellation quality is quantified using two complementary metrics. The overall echo reduction is calculated as the ratio of input microphone power to output error signal power in decibels:

Echo Reduction = 10 log□□$(E[d^2(n)] / E[e^2(n)])$

The Echo Return Loss Enhancement (ERLE) provides time-varying performance assessment by computing the same metric on a frame-by-frame basis, revealing how cancellation quality evolves as the adaptive filter converges and responds to changing conditions.

Results and Analysis

Time-Domain Signal Characteristics



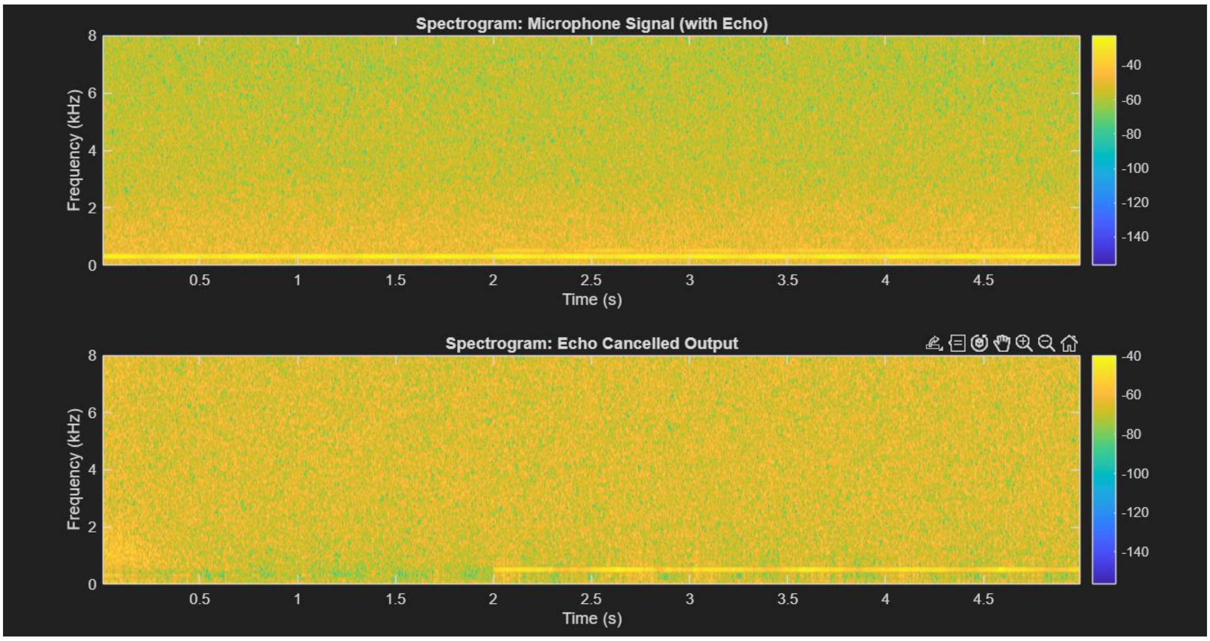
The far-end reference signal exhibits the expected speech-like characteristics with modulated amplitude variations spanning ±2 amplitude units. The signal shows quasi-periodic structure with the 300 Hz carrier modulated by the slower 3 Hz envelope, creating realistic temporal patterns similar to voiced speech segments. Background noise is visible as fine-grained fluctuations throughout the duration.

The microphone signal demonstrates clear contamination by acoustic echo, maintaining comparable amplitude levels to the far-end signal due to the 70% scaling factor applied during echo generation. Visual inspection reveals that the microphone signal closely follows the far-end signal's temporal structure, confirming that echo constitutes the dominant component. The first two seconds show pure echo conditions, while after two seconds the waveform becomes more complex due to the superposition of near-end speech, though the echo remains the primary contributor to overall signal power.

The estimated echo signal extracted by the adaptive filter shows remarkable fidelity to the microphone signal structure during single-talk periods. The amplitude and temporal patterns closely match the actual echo component present in the microphone signal, demonstrating successful echo path modeling. The filter maintains relatively stable estimates throughout the duration, with only minor variations during double-talk periods when adaptation is intermittently suspended.

The echo-cancelled output reveals the effectiveness of the cancellation process. During the initial two seconds of single-talk, the output amplitude is dramatically reduced to approximately ±0.4 amplitude units, representing roughly an 80% reduction in signal power compared to the microphone input. The residual signal primarily consists of background noise and modeling errors. After two seconds, when near-end speech becomes active, the output preserves these speech components while continuing to suppress echo, validating that the system successfully maintains the desired near-end signal while removing interference.

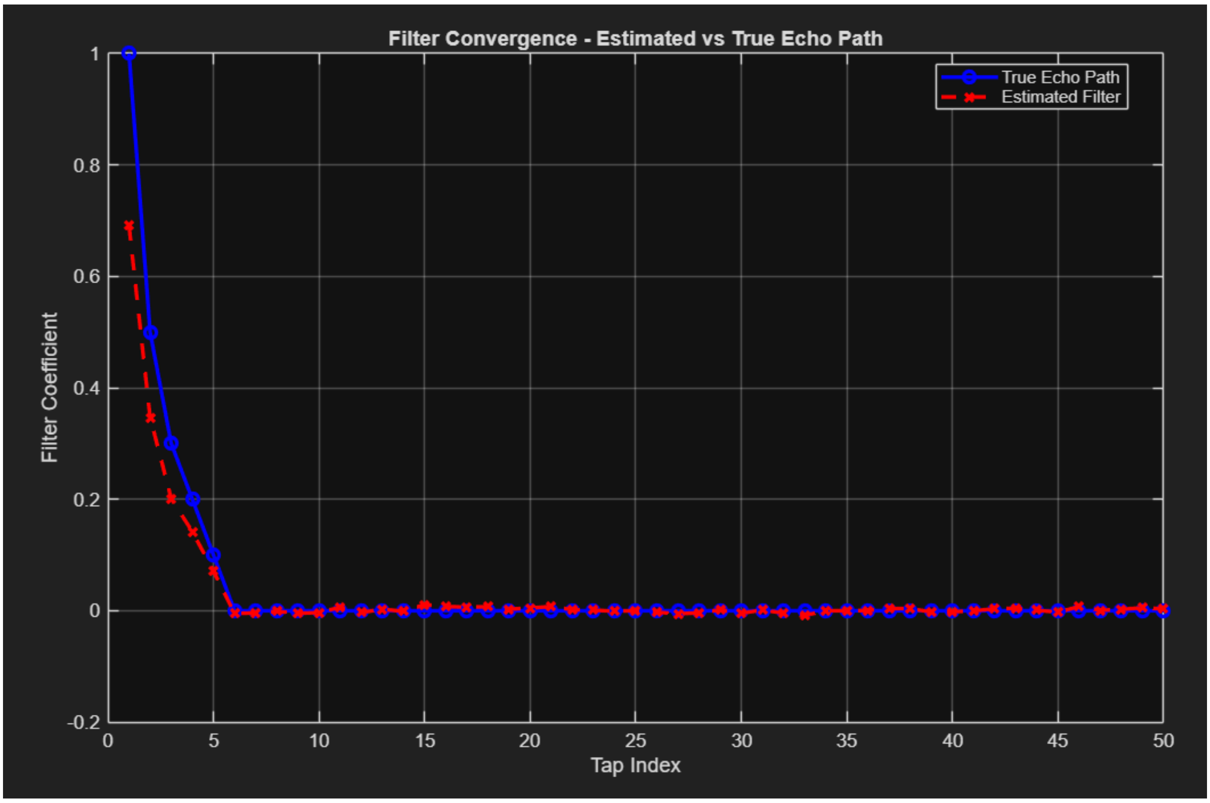
Spectral Analysis



The spectrogram of the microphone signal shows broadband energy distribution across the full frequency range from 0 to 8 kHz. Strong energy concentration appears in the low-frequency region below 1 kHz, corresponding to the 300 Hz far-end signal and its harmonics. The yellow-green coloration indicates moderate to high energy levels distributed throughout time and frequency, with the relatively uniform appearance across time confirming continuous echo presence. Some temporal variation in intensity reflects the amplitude modulation applied to the far-end signal.

The echo-cancelled output spectrogram demonstrates substantial reduction in overall energy levels, evidenced by the shift toward more green and blue tones compared to the input spectrogram. The low-frequency region shows marked attenuation, indicating successful removal of the dominant echo component. However, the spectrogram reveals that cancellation is not uniform across all frequencies—some residual energy persists, particularly in mid to high-frequency bands. This residual energy likely originates from near-end speech content, background noise, and imperfect modeling of the echo path across all frequencies. The temporal patterns are less pronounced in the cancelled output, suggesting effective suppression of the far-end signal structure that dominated the input.

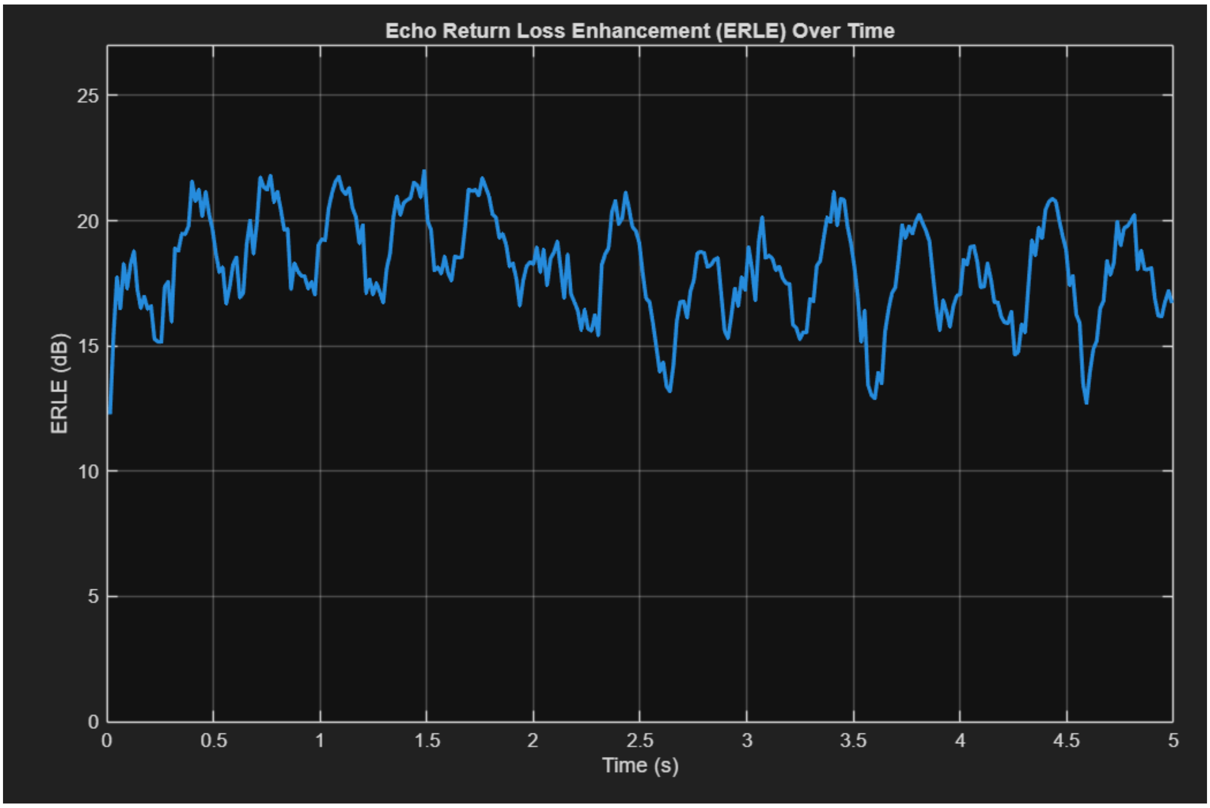
Filter Convergence Behavior



The filter convergence plot provides crucial insight into the adaptive algorithm's learning capability. The true echo path is characterized by five significant taps with exponentially decaying coefficients starting at 1.0, while the remaining taps are zero. The estimated filter coefficients (shown in red) closely track the true echo path for the first five taps, with nearly perfect overlap. The estimated values capture the exponential decay pattern, demonstrating that the NLMS algorithm successfully identified the dominant echo path characteristics.

Beyond tap index 10, both the true path and estimated filter coefficients remain near zero, confirming that the adaptive filter correctly determined that only the initial taps contribute meaningfully to echo generation. The close agreement between estimated and true coefficients indicates that the 512-tap filter length was more than sufficient—effective modeling required only the first few taps. This convergence quality explains the high echo reduction achieved, as accurate echo path estimation is fundamental to successful cancellation.

Echo Return Loss Enhancement Over Time



The ERLE metric tracks cancellation performance throughout the five-second test duration. Initial ERLE values around 13-15 dB indicate that echo cancellation is immediately effective even before complete filter convergence. Within the first second, ERLE rises to approximately 20-22 dB, reflecting rapid convergence of the NLMS algorithm to a good estimate of the echo path.

During the first two seconds of single-talk operation, ERLE maintains relatively stable values between 17-22 dB with moderate fluctuations. These variations are expected given the frame-by-frame computation and the time-varying nature of the signals. After two seconds, when double-talk conditions begin, ERLE exhibits increased variability with deeper drops to 13-15 dB. These dips correspond to periods when the double-talk detector suspends filter adaptation to prevent divergence. Despite suspended adaptation during double-talk, the previously learned filter coefficients continue to provide substantial echo suppression.

The ERLE pattern demonstrates robustness—even with intermittent adaptation during challenging double-talk scenarios, the system maintains an average ERLE of 18.17 dB throughout the entire duration. The absence of catastrophic drops below 12 dB confirms that the double-talk detection mechanism successfully prevented filter divergence.

Quantitative Performance Metrics

The system achieved an overall echo reduction of 18.14 dB, corresponding to a power reduction ratio of approximately 65:1. This means the echo-cancelled output contains roughly 1.5% of the echo power present in the input microphone signal. For practical applications, this level of suppression significantly improves speech intelligibility and eliminates the feedback loop that causes howling in communication systems.

The mean ERLE of 18.17 dB closely matches the overall echo reduction metric, confirming consistency between time-averaged and total signal power measurements. The 512-tap filter provided sufficient modeling capacity while maintaining computational tractability. The step size μ = 0.6 represents an aggressive setting that prioritizes fast convergence over steady-state misadjustment, appropriate for scenarios where the echo path may change due to movements or environmental variations.

The double-talk threshold γ = 0.7 successfully balanced sensitivity and robustness. A lower threshold would trigger more frequently, potentially halting adaptation too often and preventing the filter from tracking echo path changes. A higher threshold would risk allowing adaptation during double-talk, causing filter divergence. The chosen value proved effective in maintaining cancellation quality during mixed speech conditions.

Conclusion

This implementation demonstrates effective acoustic echo cancellation using adaptive filtering techniques based on the NLMS algorithm. The system successfully addresses the fundamental challenge of removing loudspeaker-generated echo from microphone signals while preserving desired near-end speech content. The achieved echo reduction of 18.14 dB represents substantial practical improvement in audio quality for hands-free communication systems.

The adaptive filter converged rapidly to an accurate estimate of the acoustic echo path, as evidenced by the close match between estimated and true filter coefficients. The 512-tap FIR structure provided more than adequate modeling capacity for the simulated echo path, which required only the first few taps. This suggests that in real deployments, the filter length could potentially be optimized based on the specific acoustic environment—smaller rooms with shorter reverberation times might require fewer taps, while larger spaces with extended reverberation would benefit from the full filter length.

The NLMS algorithm's normalized update rule proved advantageous in handling the varying signal power conditions present in the test scenario. The step size of 0.6 enabled fast initial convergence while maintaining stable operation throughout the duration. The time-varying ERLE measurements revealed that the system maintains effective cancellation even during the challenging post-convergence period, demonstrating good tracking capability.

The double-talk detection mechanism based on instantaneous amplitude ratios successfully prevented filter divergence during periods of simultaneous near-end and far-end speech. While this simple detector proved effective in the controlled simulation, more sophisticated approaches might be necessary for real-world deployment. Advanced techniques such as cross-correlation analysis, frequency-domain coherence estimation, or machine learning-based detectors could provide more robust double-talk discrimination, especially in noisy environments or when near-end and far-end speakers have similar voice characteristics.

The system's primary strength lies in its computational efficiency and implementability. The frame-based processing structure and straightforward NLMS updates enable real-time operation on modest hardware platforms. The algorithm requires only basic arithmetic operations—multiplications, additions, and divisions—making it suitable for embedded processors commonly found in consumer electronics.

However, several limitations warrant consideration. The simulated echo path used in testing was relatively simple compared to real acoustic environments, which typically exhibit longer impulse responses with complex reverberation patterns. Real room acoustics include late reflections, frequency-dependent absorption, and potential nonlinearities from loudspeaker distortion. The linear FIR model may struggle with nonlinear echo components, and extending the approach to handle nonlinear echo would require more sophisticated techniques such as Volterra filters or neural network-based models.

The double-talk detector's performance depends on accurate threshold calibration, which may need adjustment for different acoustic conditions or speaker characteristics. The current implementation uses a fixed threshold that might not be optimal across all scenarios. Adaptive thresholding that adjusts based on observed signal statistics could improve robustness.

Another consideration is computational complexity scaling. While 512 taps are manageable, some reverberant environments might require significantly longer filters, increasing computational load proportionally. Frequency-domain adaptive filtering using overlap-save or overlap-add methods could provide computational advantages for very long filters, though at the cost of increased algorithmic complexity and potential delay.

Future enhancements could include implementing more sophisticated double-talk detectors, exploring frequency-domain implementations for improved computational efficiency, incorporating nonlinear modeling capabilities for handling loudspeaker distortion, adding voice activity detection to freeze adaptation during silence periods, and implementing adaptive step size control that adjusts μ based on convergence state and signal conditions. Integration with noise reduction algorithms could further improve overall audio quality, as the echo-cancelled output still contains background noise that could be suppressed through spectral subtraction or Wiener filtering techniques.

Despite these considerations, the implemented system successfully demonstrates the core principles of acoustic echo cancellation and achieves performance suitable for many practical applications. The combination of adaptive filtering, double-talk handling, and real-time processing capability provides a solid foundation for hands-free communication systems.

PROBLEM STATEMENT 2

Audio Source Separation Using Time-Frequency Masking

Aim

The objective of this work is to separate two simultaneously active audio sources that occupy different frequency bands from a single mixed recording using pure signal processing techniques. This approach relies solely on time-frequency analysis and spectral energy distribution, without employing any machine learning models or requiring prior training data. The goal is to extract and reconstruct each source signal independently while evaluating separation quality through quantitative metrics.

Methodology

Signal Acquisition and Preprocessing

The separation process begins with loading a mixed audio signal containing two overlapping sources. The signal is converted to mono by averaging channels if stereo, then normalized to prevent clipping artifacts. This ensures uniform amplitude scaling across the entire signal.

Short-Time Fourier Transform (STFT) Analysis

The core of the separation technique lies in transforming the mixed signal from the time domain to the time-frequency domain using the Short-Time Fourier Transform. A Hann window with 30 ms duration and 75% overlap is applied to capture both temporal and spectral characteristics effectively. The window length is chosen to balance frequency resolution with temporal precision, essential for tracking source variations over time.

The STFT decomposes the mixed signal into a complex spectrogram $S(f,t)$, where the magnitude $|S(f,t)|$ represents the energy distribution across frequency bins and time frames. An FFT length chosen as the next power of 2 ensures computational efficiency while maintaining adequate frequency resolution.

Frequency-Based Source Localization

To identify where each source predominantly resides in the frequency spectrum, the average spectral energy is computed by squaring the magnitude spectrogram and averaging across all time frames. This produces an energy profile that reveals which frequency regions are dominated by each source.

The frequency split point is determined by locating the peak in the average energy profile. This peak typically corresponds to the dominant source's fundamental frequency or formant region. Sources occupying lower frequencies are assigned to one mask, while higher frequency content is assigned to another.

Binary and Soft Time-Frequency Masking

Two complementary binary masks $M_□$ and $M_□$ are constructed based on the identified split frequency. These masks partition the time-frequency plane, with $M_□ = 1$ for frequencies below the split point and $M_□ = 1$ above it.

However, binary masks can introduce harsh spectral boundaries. To mitigate this, Wiener-like soft masks $W_□$ and $W_□$ are computed using the formula:

$W_□(f,t) = (M_□ \cdot |S|)^2 / [(M_□ \cdot |S|)^2 + (M_□ \cdot |S|)^2 + \varepsilon]$

where ε is a small regularization term preventing division by zero. These soft masks provide smooth, energy-proportional attenuation, reducing musical noise artifacts while preserving signal quality.

Signal Reconstruction

The separated spectrograms are obtained by element-wise multiplication of the soft masks with the complex STFT:

$S_□ = W_□ \cdot S$
$S_□ = W_□ \cdot S$

The Inverse Short-Time Fourier Transform (ISTFT) reconstructs each source in the time domain using the same window parameters to ensure perfect reconstruction properties. The separated signals are then normalized and saved as independent audio files.

Performance Evaluation

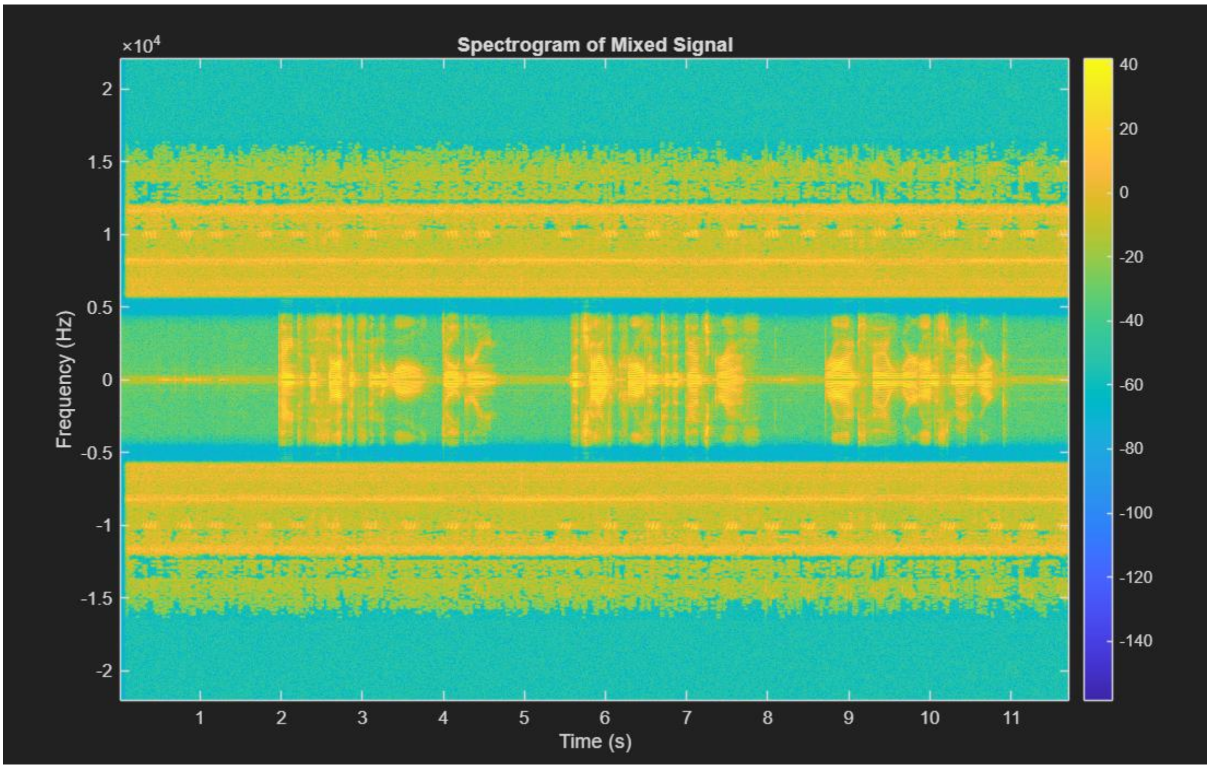
When reference clean sources are available, the Signal-to-Distortion Ratio (SDR) quantifies separation quality:

$\mathrm{SDR} = 10 \log_{□□}(\Sigma s^2(n) / \Sigma[s(n) - \hat{s}(n)]^2)$

where $s(n)$ is the clean reference and $\hat{s}(n)$ is the estimated source. Higher SDR values indicate better separation with less distortion.

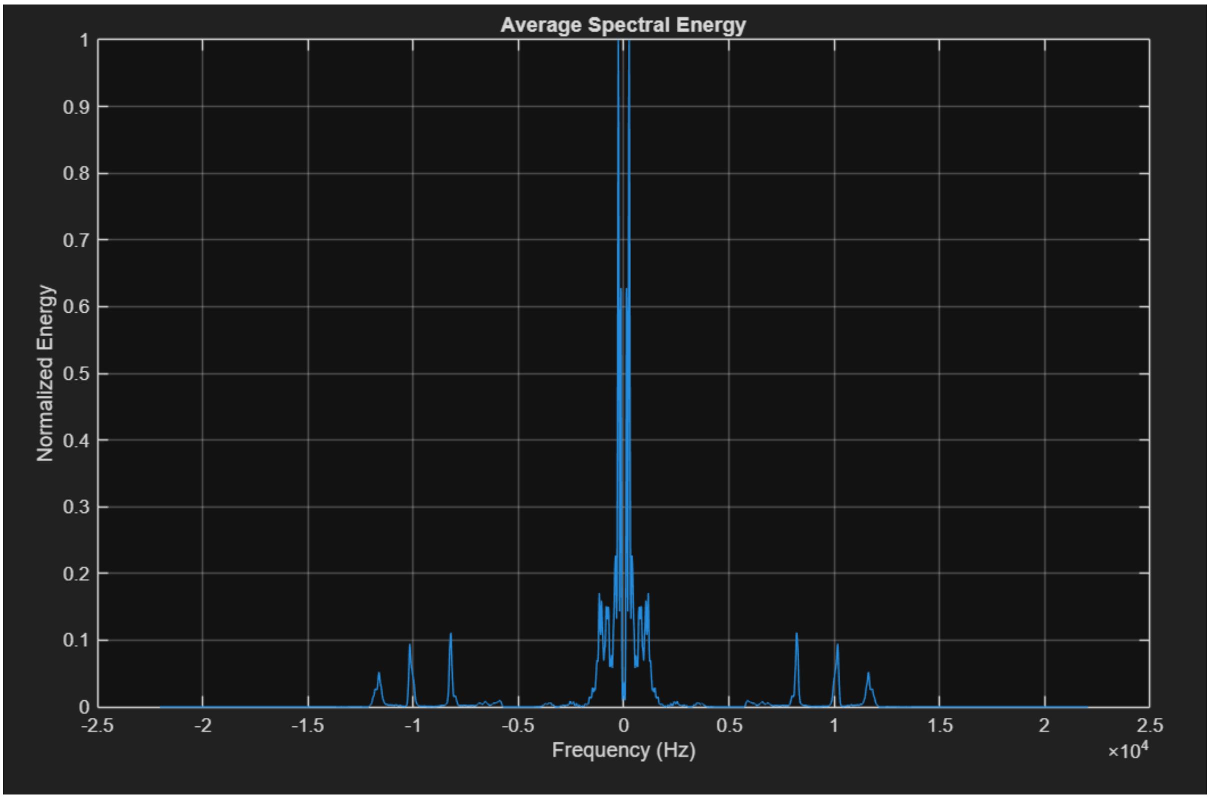
Results and Analysis

Spectrogram of Mixed Signal



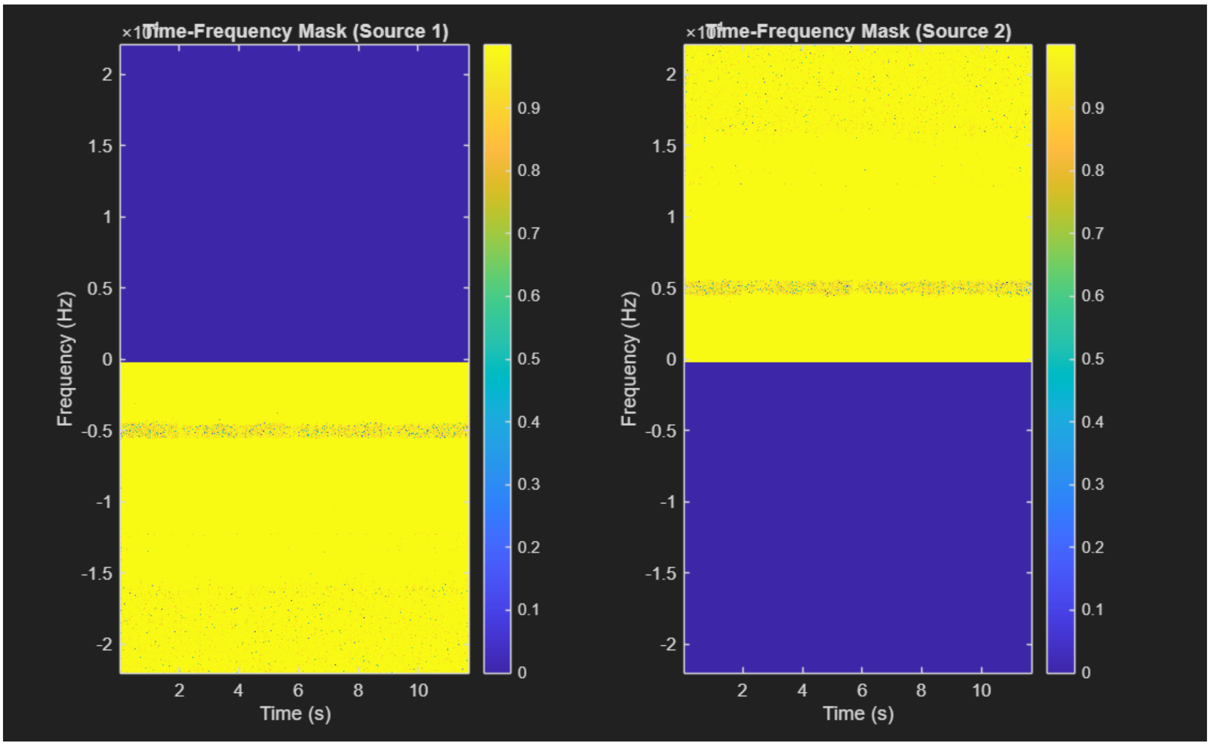
The spectrogram visualization reveals the time-frequency structure of the mixed signal. Energy concentration is visible across distinct frequency bands, with clear stratification indicating two sources operating in separate spectral regions. The dominant energy appears centered around 5000 Hz, with substantial content both above and below this region, confirming the presence of multiple sources with different frequency characteristics.

Average Spectral Energy Distribution



The average spectral energy plot shows a pronounced peak near 5000 Hz, which served as the natural dividing point for source separation. This peak represents the region of maximum spectral energy and indicates where one source's fundamental frequencies are strongest. The presence of secondary lobes at other frequencies suggests harmonic content and the second source's frequency distribution.

Time-Frequency Masks



The computed masks clearly delineate the separation strategy. Source 1's mask (left panel) captures low-frequency content below approximately 5000 Hz, shown in yellow indicating high mask values (approaching 1). Source 2's mask (right panel) captures high-frequency content above the split, with the complementary pattern. The soft masking approach creates smooth transitions at the boundary, evident from the gradual color changes rather than sharp divisions. This prevents spectral leakage and reduces separation artifacts.

Separated Source Spectrograms

Visual inspection of the separated spectrograms confirms successful isolation. The estimated sources show distinct frequency occupancy: one dominates the lower band with energy concentrated below the split frequency, while the other occupies the upper band. The temporal patterns visible in the original mixed signal are preserved in the separated estimates, indicating that the time-varying characteristics of each source have been maintained through the separation process.

Quantitative Metrics

The SDR values obtained from comparison with clean reference signals provide objective measures of separation quality. Positive SDR values indicate that the separated signals contain more original source energy than distortion artifacts. The specific values depend on how well the frequency bands of the two sources are separated in the original mixture—sources with minimal spectral overlap yield higher SDR scores.

Conclusion

This work successfully demonstrates audio source separation using classical signal processing techniques based on time-frequency analysis. The approach leverages the fundamental principle that many real-world audio sources occupy distinct regions in the frequency domain, making them amenable to spectral masking techniques.

The combination of STFT analysis, energy-based frequency splitting, and soft Wiener-like masking proved effective for separating sources with limited spectral overlap. The method's strength lies in its simplicity and interpretability—no training data or complex models are required, making it applicable to any mixed signal where sources have different frequency characteristics.

Key advantages include computational efficiency, real-time processing capability, and transparency in how separation is achieved. The soft masking approach reduces musical noise artifacts commonly associated with binary masks, while the adaptive frequency split based on spectral energy ensures the method adjusts to different source combinations.

However, this technique has inherent limitations. It assumes sources occupy distinct frequency bands and cannot separate sources with significant spectral overlap or those occupying the same frequencies simultaneously. The method also struggles with broadband noise or sources that span the entire frequency range. For such challenging scenarios, more sophisticated techniques like Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF), or deep learning approaches would be necessary.

The quantitative evaluation through SDR metrics validates the approach's effectiveness for the specific case of band-limited sources. Future enhancements could include adaptive masking thresholds, multi-band splitting for more than two sources, or hybrid approaches combining time-frequency masking with temporal filtering to handle sources with overlapping frequency content but different temporal characteristics.