Pertanika J. Sci. & Technol. 34 (2): 951 - 973 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Improved Speech Enhancement using Parallel MVDR Beamforming and Coherent-to-Diffuse Power Ratio Post-filtering

Sureshkumar Natarajan[1], Mohd Khair Hassan[2], Raja Kamil[2], Faisul Arif Ahmad[1], Syaril Azrad[3], June Francis Macleans[4], Hussna Elnoor M. Abdalla[1], Nurbek Saparkhojayev[5], and Syed Abdul Rahman Al-Haddad[1]*

[1]Department of Computer and Communication Systems Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Department of Electrical and Electronic Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[3]Department of Aerospace Engineering, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[4]Independent Researcher, 50250 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia
[5]Chairman of the Board-Rector of the NCJSC, Rudny Industrial University, 111500 Rudny City, Kazakhstan

ABSTRACT

Speech communication involves the exchange of information between two or more individuals; however, background noise and reverberation often degrade speech clarity and intelligibility. The impact of these degradations depends on factors such as the number, intensity, and spatial characteristics of noise sources, as well as reflections from surrounding surfaces. These effects pose significant challenges for applications including teleconferencing, hearing-aid devices, and human-machine voice interfaces. This study aims to address the combined effects of background noise and reverberation by proposing a speech enhancement framework that integrates Minimum Variance Distortionless Response (MVDR) beamforming with a Coherent-to-Diffuse Power Ratio (CDR) based post-filter. The methodology relies on parallel processing, where MVDR beamforming performs spatial noise suppression, while CDR values are estimated from microphone-domain signals and used to compute post-filter gains for suppressing residual diffuse noise and reverberation in the beamformer output. The proposed algorithm was evaluated over an input signal-to-noise ratio range from 0 dB to 40 dB and compared against the Integrated Sidelobe Cancellation Linear Prediction (ISCLP) baseline using four objective metrics: Perceptual Evaluation of Speech Quality (PESQ), Extended Short-Time Objective Intelligibility (ESTOI),

ARTICLE INFO

Article history:
Received: 04 August 2025
Accepted: 04 February 2026
Published: 17 April 2026

DOI: https://doi.org/10.47836/pjst.34.2.15

E-mail addresses:
suresh45619@gmail.com (Sureshkumar Natarajan)
khair@upm.edu.my (Mohd Khair Hassan)
kamil@upm.edu.my (Raja Kamil)
faisul@upm.edu.my (Faisul Arif Ahmad)
syaril@upm.edu.my (Syaril Azrad)
macleans30june@gmail.com (June Francis Macleans)
hussnaa@hotmail.com (Hussna Elnoor M. Abdalla)
nursp81@gmail.com (Nurbek Saparkhojayev)
sar@upm.edu.my (Syed Abdul Rahman Al-Haddad)
* Corresponding author

e-ISSN: 2231-8526

Cepstral Distance (CD), and Weighted Spectral Slope Distance (WSS). The results demonstrate consistent improvements in PESQ and reductions in CD and WSS across all tested conditions, while gains in ESTOI remain modest. These findings indicate improved spectral fidelity and speech naturalness, highlighting the practical relevance of the proposed framework for real-time, low-complexity speech enhancement in noise and reverberation-prone communication systems.

Keywords: Acoustics, CDR, dereverberation, MVDR beamformer, noise suppression, speech enhancement

INTRODUCTION

Various speech-based application devices, such as voice-controlled robotics and day-to-day use of hands-free mobiles, process the signal captured by the microphone. Furthermore, the speech recognition and speaker recognition systems also depend on the quality of the microphone recorded signal, where the speech signal is often contaminated with both background noise and reverberant signal, which cannot be directly used for speech application devices. Hence, the requirement of a speech enhancement algorithm is crucial and essential, which filters the unwanted component and thus aims to enhance the quality and clarity of the speech. Over the years, there has been tremendous advancement in the development of various speech enhancement algorithms, right from the traditional speech enhancement, such as spectral subtraction and Wiener filter, to modern deep learning-based speech enhancement algorithms, which are driven by data (O'Shaughnessy, 2024). Despite these advancements, the same clarity of signal uttered by the source speaker will not reach the receiving end of the speaker due to the surrounding background noise and reverberation, which distorts the source signal (Beutelmann & Brand, 2006). To tackle these issues and to strengthen the desired speaker utterance, multi-microphone techniques using spatial filtering effects were extensively investigated (Delcroix et al., 2015; Habets & Benesty, 2013). Another technique based on a recursive Kalman filtering framework was proposed by Cohen et al. (2021), specifically designed to deal with both reverberation effects as well as interfering sources. The two most fundamental ideas used in modern multi-microphone speech enhancement systems are the spatial filtering concept and deconvolution.

The spatial filtering was usually realised using multiple microphones, which is called the beamforming technique, which is basically used to enhance the desired speech arriving from a particular direction and suppress the interfering sources arriving from other unknown directions. Beamforming, in general, is more successful in removing directional sources arriving from outside the region of interest as compared to reverberated signals. Among various types of beamforming techniques, the MVDR beamformer is known to be a widely used technique as it offers better improvement in speech intelligibility and clarity in a complex noise environment (Cauchi et al., 2015). Some other techniques which integrate a linear prediction framework with a beamforming method have also shown improvement in obtaining signal clarity within reverberant conditions (Yang et al., 2022).

There have been some efforts taken to integrate noise reduction with dereverberation, which has led to the advancement of hybrid techniques. For example, combining pre-whitening and Generalised Sidelobe Canceller (GSC) beamformer helps to enhance the desired speech by removing reverberation (Dietzen et al., 2015). Similarly, fusion based technique was employed to deal with both noise suppression and reverberation, which uses Multichannel Linear Prediction (MCLP) and GSC within a Kalman filtering framework (Dietzen et al., 2018). Another technique that uses a cascade approach that employs a Multiple-Input Multiple-Output (MIMO) based Weighted Prediction Error (WPE) algorithm is used as a preprocessing stage, followed by Minimum Power Distortion Less Response (MPDR) beamforming to simultaneously tackle both reverberation and background noise (Delcroix et al., 2015).

The recent studies have indicated that further enhancement of the signal can be obtained as compared to the cascaded approach. A study has shown Weighted Power Minimisation Distortion Less Response (WPD) beamformer, which jointly optimises the convolutional WPE filter and MPDR beamformer, yielded in superior performance (Nakatani & Kinoshita, 2019a). Extension to this model was carried out by including the application of sparse priors to further refine the WPD beamformer's performance (Gode et al., 2021). To obtain effective enhancement of the signal, an adaptive version has been introduced which uses RLS-WPD and the recursive WPD approach to simultaneously suppress both background noise and reverberation (Nakatani & Kinoshita, 2019b). One more interesting approach was introduced in recent times called the Weighted Binaural Linearly Constrained Minimum Power (wBLCMP) beamformer, which has shown the ability to jointly handle background noise, dereverberation and other interfering sources while maintaining binaural cues. This enhancement was achieved by extending the WPD formulation by combining the optimisation of the convolutional WPE filter with the Linearly Constrained Minimum Power (LCMP) beamformer (Gode & Doclo, 2023).

Although the proposed work is based on traditional non-deep learning (DL) based methods, in recent years, there has been a lot of work-based DL-based algorithms that have shown tremendous advancement in enhancing the desired speaker speech. For instance, an attention-based MVDR beamforming architecture (ABIC-MVDR) was introduced by Bai et al. (2025) that integrates spatial covariance matrix estimation with In-Place Gated Convolutional Recurrent Networks and frequency-independent LSTM modules. They employed an end-to-end design that yields a significant improvement in challenging conditions, including the movement of the speaker and varying acoustic environments. In another study, the MVDR beamforming technique was employed within a U-Net architecture to obtain multichannel speech in severe noise and a reverberant environment (Lee et al., 2024). A deep beamforming framework proposed to jointly enhance speech signals and localise target speakers by learning spatial filtering directly from multichannel data, which uses a convolutional recurrent network for training (CRN) (Chang et al., 2024).

A lightweight two-stage CRN was proposed for the hearing aid application that uses a microphone array for enhancing the speech signal (Xi et al., 2024). A WPD beamforming approach was proposed to jointly perform dereverberation and denoising, where the unified filter is estimated using a deep neural network (Fujita et al., 2024). A detailed systematic review that uses 187 studies highlighted the impact of various deep learning-based techniques, such as CNN, LSTM, and many more, for developing a speech enhancement algorithm to improve the quality of the speech signal, and the study also discussed speech enhancements based on beamforming techniques (Natarajan et al., 2025). In another study, which specifically discussed the impact of the beamforming technique related to enhancement in the study, this review mainly discussed the comparison of various techniques with the beamforming approach, which has shown an effective method in enhancing the clarity of speech in a far-field microphone captured signal (Natarajan et al., 2023).

Furthermore, Huang et al. (2025) have conducted a comprehensive review based on microphone array processing. The author has shown tracing evolution from the traditional beamforming approach to the latest advancement of the spatial filter using the DL approach. Although a notable performance gain through the use neural-based beamforming technique can be found to be better as compared to a traditional statistical-based approach. However, DL methods often rely on a huge amount of data, which eventually results in increasing the computational complexity, and they are less interpretable. Due to this fact, it is more often preferred to use a non-DL method as it is easier to carry out interpretation and analysis, which helps to address the issues. Due to these practical considerations, a traditional-based approach was employed in this work.

The CDR is one of the well-known techniques capable of suppressing the reverberant components (Schwarz & Kellermann, 2015). Using the idea of the CDR technique, an algorithm was developed recently that uses CDR as a preprocessing and the CDR output is fed to a recursive least squares (RLS) adaptive filter to remove noise without depending on an explicit noise reference signal (Natarajan et al., 2024). However, it is a known fact that CDR alone cannot be effective when recording is carried out in an environment where multiple speakers talk simultaneously. To handle multiple speakers talking scenarios, the MVDR beamforming method serves as one of the effective techniques in enhancing the desired speaker's speech by isolating the interfering speaker signal and helps to mitigate some reverberant components.

The idea is to take advantage of using both techniques by proposing a hybrid approach that will help to suppress the remaining residual reverberant component that exists at the output of the MVDR beamformer. In this work, a noise-corrupted signal will be processed by both the MVDR beamformer as well as CDR estimation block in parallel, and later the CDR output will be used as SNR value which in turn calculates the gain values of the

CDR is later used as CDR gain as a post filter to obtain the final output where the MVDR beamformer output is multiplied with the post-filtered gain value. In this proposed work, the aim is to suppress the babble noise and reverberation effectively as compared to the baseline ISCLP algorithm, using the same experimental conditions (Dietzen et al., 2020).

RELATED WORK

Speech enhancement algorithms play a crucial role in enhancing the quality and clarity of speech by reducing background noise in a challenging acoustic setting (Shankar et al., 2021). Single-channel techniques are the simplest and easiest to use in the system, as they depend solely on the spectral and temporal characteristics of a corrupted signal to estimate clean speech. However, their performance was limited because they do not employ spatial information, and their performance will be affected severely, specifically in a complex acoustic setting where the impact of residual noise and speech distortion is more significant (Shankar et al., 2021). To tackle these issues, the Modified Noise Reduction Method (MNRM) was proposed by Kumar and Chitra (2022), which was built upon the traditional Wiener filtering framework. This method succeeds in enhancing the speech by suppressing the noise where Wiener gain was calculated before and after SNR-based filtering, and it incorporates a decision-directed strategy that uses a twin-stage suppression mechanism approach. This approach has shown more resilience in practical applications. However, the performance of the MNRM method still depends on estimating the value of power spectral densities (PSDs) of clean speech and noise, which is one of the biggest challenges to accurately determine the PSD values under real-time and non-stationary scenarios. In a related work, the Wiener gain was determined by considering both noise and reverberation, and the gain was obtained using estimates of both SNR and CDR (Xiang et al., 2025).

In contrast, the use of microphone array processing was preferred over single-channel as it uses multiple microphones to capture spatial diversity and it offers better noise removal capability to yield higher performance over single-channel-based techniques, specifically when sources are arriving from different directions, and also it helps in handling diffuse noise sources or non-stationary signals (Song et al., 2021). Multi-channel processing allows for the spatial filtering, which ensures the isolation of the desired speaker's speech from non-desired speech. However, it is difficult to deal with when the sources are moving simultaneously because the impact of overlapping spectral content and frequent directional changes in real-time will be greater, which will reduce the performance of spatial filtering. Traditional algorithms such as multichannel Wiener and Kalman filters have been adopted for use in the multichannel domain, but they cannot produce distorted output speech signals when the estimation of noise power is inaccurate (Dietzen et al., 2020). Recently, Thimmraja et al. (2022) extended the spectral subtraction (SS) algorithm approach by combining phase information between the noisy speech and the noise model. This modified approach has

shown some improvement in cross-term estimation in the gain function, which in turn helps to align these modified to perform more closely to that of minimum mean square error (MMSE)-based filtering, and the results of this method have shown better musical noise suppression and better performance despite adverse noise conditions, specifically at low SNRs.

Research using an adaptive beamforming approach has shown significant advancement due to the requirement of robust performance in reverberant or uncertain acoustic settings. Zhao et al. (2020) focused on adaptive strategies to address the challenges, such as steering vector mismatches, statistical estimation errors, and late reverberation. To tackle the steering vector model uncertainties, the norm-constrained optimisation technique was introduced first, which results in improving the robustness against directional mismatches and multipath effects. Spatial smoothing was employed as the second step to stabilise the estimation of the noise covariance matrix, which results in reducing the diffuse interference. This method has shown notable performance improvement in terms of PESQ and STOI metrics as compared to the standard beamforming approach (Zhao et al., 2020).

In the last few years, there have been significant advancements in DL-based multichannel speech enhancement approaches as they learn the complex pattern of spatial details and help to determine the spectral relationship from the data. Some of the neural network-based approaches, such as the Gated Convolutional Recurrent Network (GCRN) proposed by Tan and Wang (2019), and the Deep Path Convolutional Recurrent Network (DPCRN) introduced by Le et al. (2021), which focused mainly on spectral mapping by considering each microphone channel as an independent input signal. Although these approaches have proved to be effective, especially for spectral modelling, they often neglect the important spatial dependencies between the microphone channels. To tackle these issues, Liu and Zhang (2021) developed the In-Place Gated Convolutional Recurrent Neural Network (IGCRN) that helps to retain the spatial information at every frequency bin without performing either down-sampling or up-sampling. Later, Tan et al. (2022) introduced a neural spectro-spatial filtering approach using a multi-microphone system that adopts a learning approach to learn both spectral and spatial filters within a convolutional network, which has shown improvement in the performance gain of the speech enhancement task.

To attain better speech enhancement performance, more advanced DL architectures such as FasNet (Luo et al., 2019), EabNet (Li et al., 2022), and MIMO-Unet (Ren et al., 2021) are explicitly integrated with traditional beamforming methods. This integration ensures the efficient use of spatial diversity in noisy acoustic settings. Despite such progress, the existing DL-based systems still process multichannel input by concatenating Short-Time Fourier Transform (STFT) features from each microphone channel. Furthermore, these approaches often fail to consider two important spatial clues, namely inter-channel phase information and spatial coherence, which are essential for obtaining accurate

source localisation, and they help to isolate the target speech from the background noise in a reverberant environment. Moreover, extracting the spatial features using a DL-based model is computationally expensive, and it may introduce latency, resulting in difficulty in employing it in a real-time system. Another issue with the DL model is related to data, because the performance of the DL model mostly depends on a huge number of labelled datasets required for effective training of the model. More importantly, the dependency of labelled data used for training may limit the performance of the model when a model is fed with new unseen data.

Non-neural network or analytical approaches have received more attention due to the fact they do not need large datasets for operation, and it is less computational cost and more importantly it is possible to perform the interpretation and signal analysis at any stage using an analytical approach whereas in case of DL model, interpretation of the signal is difficult as the model itself acts like a black-box. One such notable speech enhancement filtering technique is the ISCLP algorithm (Dietzen et al., 2020) that integrates MCLP with GSC in a Kalman filter framework. This design allows for performing spatial and temporal filtering operations simultaneously, and this approach has resulted in significant performance as compared to traditional sequential processing pipelines, specifically in suppressing the diffuse reverberation and interfering sources in distant talking acoustic settings. However, its practical implementation may face some constraints and difficulties because the accuracy of this method heavily depends on the precise estimation of Relative Early Transfer Functions (RETFs). If the RETFs are not estimated accurately, then in such scenarios, self-cancellation artefact issues can occur, which result in disrupting the filter updates, which can be seen during pauses in speech. Furthermore, the MCLP filters depend on active speech segments to learn active filter coefficients, making it difficult to jointly optimise with other models in real-time. Another drawback of the ISCLP algorithm is computationally complex, which puts a limitation on real-time implementation.

To overcome these issues, the proposed method uses the MVDR beamformer and CDR to implement a hybrid approach. The role of the MVDR beamformer in this work is to preserve the desired speech by suppressing the directional interferences. However, the performance of the MVDR beamformer tends to degrade in a scenario if the level of reverberation is high, where the spatial selectivity reduces due to early reflection effects. To compensate for the issue of reverberation, the CDR estimation technique is incorporated in this work, which is well-known for dereverberation. However, the CDR technique cannot handle any background noise or interfering sources. So, the proposed method considers the advantage of both methods, and in this research work, the MVDR beamformer and the CDR technique were used to implement a hybrid approach where both the MVDR beamformer and the CDR process the noise-corrupted signal. Since the output of the MVDR beamformer does contain some residual noise, which will be further reduced by

employing CDR as a post-filter. The integration of the CDR technique with the MVDR beamformer was to improve the desired speech clarity and suppress the reverberation and background noise in acoustic settings where a single dominant speech source is present.

To evaluate the effectiveness of the proposed method and to compare its performance with the baseline ISCLP algorithm, we used the same ISCLP experimental settings. The experimental setup mirrors the ISCLP evaluation environment, featuring a single target speaker within 2 meters of a microphone array, moderate reverberation (RT60 = 0.6 seconds), and multi-talker babble interference. The objective is to assess whether the MVDR-CDR-SS framework can achieve comparable or improved performance relative to ISCLP, thereby establishing its viability as a lightweight and competitive alternative for real-time, far-field speech enhancement.

SIGNAL MODELLING

In this section, the speech enhancement system is discussed. Consider the microphone recorded signals consist of mixed components of desired speech and noise/interfering sources in a reverberant environment. The desired speech gets convolved with the room impulse response (RIR), and these signals can be mathematically expressed in the STFT domain at frame m and frequency f, and the i^{th} microphone signal can be written using Equation 1 and Equation 2:

$$X_i(m,f) = o_i(m,f)S_i\,(m,f) + N_i(m,f) \quad [1]$$

$$X(m,f) = [X_0(m,f), X_1(m,f), \ldots\ldots X_{N-1}(m,f)]^T \quad [2]$$

where, $X_i(m,f)$ represents the direct signal component $S_i(m,f)$, where $o_i(m,f)$ represents the steering vector of the sound coming from the target speaker and $N_i(m,f)$ representing noise and late reverberation signals.

Minimum Variance Distortionless Response Beamformer

To process a mixed signal containing the target speaker's speech and background noise, the MVDR beamformer is used. The mixed-noise signal is first transformed into shorter segments using STFT for the Fourier transformation. The MVDR filter weights are then computed to minimise the power of the noise component while maintaining the signal power of the target speaker. This Equation 3 can be expressed as:

$$W_{MVDR}(m,f) = argmin\, w^H(m,f)\, N_i(m,f) \quad [3]$$

Based on the received multichannel covariance matrix of the noise component $N_i(m, f)$ From the microphone array, we calculate the steering vector for both the target speaker and the noise signal using the eigenvector method suggested in (Sarradj, 2010). To estimate the covariance matrix of the target signal, we need to compute the covariance matrix of the combined signal and the noise signal. Once we have the target signal's covariance matrix, we can compute the steering vector using the formulas and concepts derived. The eigenvalue decomposition of the microphone array's covariance matrix for the target speaker can be expressed using Equation 4:

$$G = \mathbb{V}\Lambda\mathbb{V}^H \quad [4]$$

where G, is the covariance matrix for the target speaker in the microphone array, and it can be written in Equation 5. Λ is a diagonal matrix that consists of the positive eigenvalues of G can be mathematically written using Equation 6, while $\mathbb{V}$ is the eigenvector corresponding to these eigenvalues. We assume that the smallest eigenvalue associated with the noise component of the signal is n^2, following the method proposed in (Su & Morf, 1983). Therefore, the eigenvalue can be split into two parts in Equation 5, one for the signal and the other for the noise. The steering vector is estimated using eigenvalue decomposition of the spatial covariance matrix, following the approach proposed by Su and Morf (1983). This method assumes sufficient separation between the signal and noise subspaces. In low-SNR or highly diffuse reverberant environments, partial overlap between these subspaces may occur, potentially affecting estimation accuracy. While this limitation is inherent to subspace-based methods, the impact is mitigated in the proposed framework through subsequent diffuseness-based post-filtering.

$$G = \mathbb{V}\begin{bmatrix}\Lambda_s & 0\\ 0 & 0\end{bmatrix}\mathbb{V}^H + n^2\mathbb{V}\mathbb{V}^H = \mathbb{V}_s\Lambda_s\mathbb{V}^H{}_s + n^2\mathrm{I} \quad [5]$$

$$\Lambda = \begin{bmatrix}\Lambda_s + n^2 I & 0\\ 0 & n^2 I\end{bmatrix} \quad [6]$$

Each part corresponds to either the signal or the noise, where the eigenvectors of $\mathbb{V}_s$ belong to the signal subspace $\mathbb{V}$, while the remaining ones belong to the noise subspace. In real experiments, the eigenvalues in the noise subspace have minimal impact on the eigenvalues in the signal subspace. We calculate the smallest eigenvalue for the noise signal in the matrix and identify the highest eigenvalues as the target source values in the matrix. Next, we assign high values to frequency bin values associated with the target speaker and low values to those associated with noise. The weights are first derived from the mixed input signal, and each frequency bin is then assigned its own corresponding weight. These

weights are applied back to the mixed signal so that every frequency bin is scaled according to its computed value. In this way, the enhancement is distributed more precisely across the frequency range, ensuring balanced improvement throughout the spectrum. Finally, we obtain a time-domain signal by converting the beamforming signal to an inverse short-time Fourier transform, following the weight calculation with the signal.

Coherent-to-Diffuse Power Ratio

To enhance the results of the MVDR beamformer, we employ a postfilter based on the spectral subtraction technique to remove diffuse noise and reverberation. We start by determining a short-time SNR value, which is used in the spectral subtraction method to calculate the filter gain. Furthermore, the estimation of the CDR only necessitates the computation of the SNR, which in turn determines the expression of the gain $G(m,f)$ which can be written using Equation 7:

$$G(m,f) = max\left\{1 - \mu \frac{1}{1+SNR(m,f)}, G_{min}\right\} \quad [7]$$

The gain floor has a predetermined value of 0.1 and an overestimation factor μ is set to 1.3 (Schwarz & Kellermann, 2015). The SNR value is derived from the CDR, and the gain floor G_{min} has a predetermined value of 0.1 while μ is set to 1.3. A mathematical definition for the CDR between two omnidirectional microphones is suggested in (Schwarz & Kellermann, 2015) that can be written using Equation 8:

$$CDR(m,f) = \frac{\Gamma_n(m,f) - \Gamma_x(m,f)}{\Gamma_x(m,f) - \Gamma_s(m,f)} \quad [8]$$

To estimate the correlation between two omnidirectional microphones, we compute the coherence function of the signal recorded by the microphone $\Gamma_x\ (m,f)$ and the diffuse noise $\Gamma_n(m,f)$. We also use short-time spatial coherence to estimate the coherence function of the diffuse field. The spatial coherence function of the diffuse field can be represented using Equation 9:-

$$\Gamma_{diffuse}\ (f) = \Gamma_n(m,f) = \frac{sin\left(2\pi f \frac{d_{1,2}}{c}\right)}{\left(2\pi f \frac{d_{1,2}}{c}\right)} \quad [9]$$

The coherence function of the desired signal component is defined in Equation 10:-

$$\Gamma_s(f) = e^{jkdsin\ (\theta)} = e^{j\,2\pi f\,\Delta t} \quad [10]$$

where, Δt represents the time difference of arrival (TDOA). Using an automatic cross-power spectrum estimator that can be estimated from the microphone signal through recursive averaging to compute the post-filter gain function, and it can be determined using Equation 11:

$$\widehat{\Phi}_{x_p x_q}(m,f) = \lambda \widehat{\Phi}_{x_1 x_2}(m-1,f) + (1-\lambda) X_1(m,f) X^*{}_2(m,f) \quad [11]$$

The microphone signals are processed with a smoothing factor λ, which is a constant value between 0 and 1. We used λ = 0.68 in the experiment. The temporal smoothing factor λ was fixed to 0.68 for all experiments, consistent with prior work on CDR-based dereverberation (Schwarz & Kellermann, 2015). Larger values of λ improve estimation stability but reduce responsiveness to rapid acoustic changes, whereas smaller values enhance temporal tracking at the risk of increased estimation variance and artefacts. In this study, λ was kept constant to ensure stable behaviour and fair comparison with the baseline. The coherence function can be determined by using Equation 12:

$$\hat{\Gamma}_x(m,f) = \frac{\widehat{\Phi}_{x_1 x_2}(m,f)}{\sqrt{\widehat{\Phi}_{x_1 x_1}(m,f)\widehat{\Phi}_{x_2 x_2}(m,f)}} \quad [12]$$

A direction-of-arrival (DOA) independent CDR estimator was employed to reduce computational complexity and improve robustness under diffuse reverberation. While this approach is effective in late-reverberant conditions, it may limit dereverberation accuracy when early reflections dominate. Incorporating DOA-aware diffuseness estimation represents a promising direction for future improvement. In this work, we use the CDR estimator represented in Equation 13 due to its independence from the DOA (Schwarz & Kellermann, 2015).

$$\widehat{CDR}_{DOAindep} = max\left(0, \frac{\Gamma_n Re\{\widehat{\Gamma_x}\} - |\widehat{\Gamma_x}|^2 - \sqrt{\Gamma_n{}^2 Re\{\widehat{\Gamma_x}\}^2 - \Gamma_n{}^2 |\widehat{\Gamma_x}|^2 + \Gamma_n{}^2 - 2\Gamma_n Re\{\widehat{\Gamma_x}\} + |\widehat{\Gamma_x}|^2}}{|\widehat{\Gamma_x}|^2 - 1}\right) \quad [13]$$

We use the real part (Re{.}), magnitude (|.|), and phase (arg{.}) operations along with the maximum operation to avoid negative values. The CDR estimator was initially designed only for microphones (Schwarz & Kellermann, 2015). However, to ensure optimal performance, we need to estimate the CDR for more than just microphones. To achieve this, we estimate the CDR for each pair of microphones, determining the diffuse value from each estimate. Then, we take the arithmetic mean of all the diffuse values of the microphone pairs to obtain the average diffuse estimate.

Instead of the average CDR for a specific pair of microphones, we will directly calculate the average diffusion estimate. This is because CDR can take any value from 0 to infinity, while diffusion values lie within a specific interval. To obtain the average diffuseness value (Schwarz & Kellermann, 2015), we apply the following Equation 14:

$$D(m,f) = (1 + \widehat{CDR}\ (m,f))^{-1} \quad [14]$$

where $\widehat{CDR}\ (m,f)$ represents the estimated CDR value. After obtaining the diffuseness value, the CDR value for each microphone can be computed using Equation 15:

$$\widehat{CDR}_{In}(m,f) = \frac{1-D(m,f)}{D(m,f)} \quad [15]$$

where $\widehat{CDR}_{In}(m,f)$ is the CDR value obtained from the microphone input, not the output of the beamformer. To account for this, a correction factor $A_\Gamma(m,f)$ is applied to $\widehat{CDR}_{In}(m,f)$. The CDR estimate from the beamformer output, $\widehat{CDR}_B(m,f)$ can be defined using Equation 16:

$$\widehat{CDR}_B(m,f) = \frac{\widehat{CDR}_{In}(m,f)}{A_\Gamma(m,f)} \quad [16]$$

The correction factor $A_\Gamma(m,f)$ can be calculated using Equation 17,

$$A_\Gamma(m,f) = w^H(m,f)\, J_{diff}\ (f)\, w(m,f) \quad [17]$$

where $w^H(m,f)$ represents the beamformer weight and $J_{diff}\ (f)$ is the $N \times N$ spatial coherence matrix of a diffuse noise field with the $(1,2)th$ element given by Equation 9. Figure 1 shows the flow process starting from the data acquired by microphones, which consists of a mixed version of target speaker speech, interfering speaker speech and diffuse noise. The corrupted signal is converted into the STFT domain using an STFT block and then applied to coherence estimation to calculate the CDR estimation value without using DOA. Also, the STFT signal is applied to the MVDR beamformer and the post filter gain, which is calculated using the spectral subtraction technique, which uses the SNR value that is obtained from the CDR estimator. Finally, the time domain output is determined by applying the inverse STFT to the spectral enhancement output.

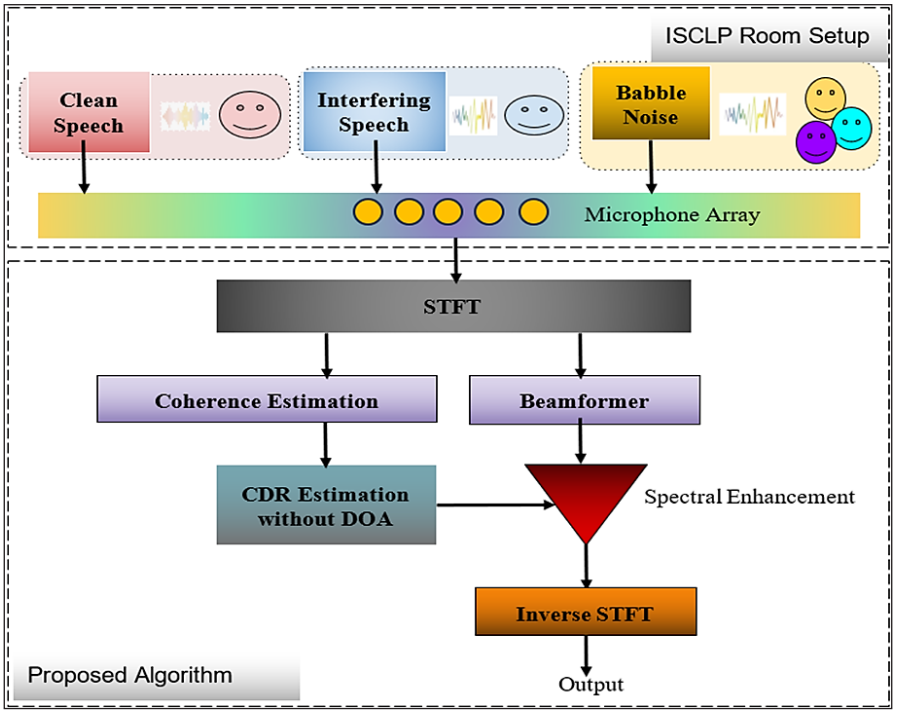

Figure 1. Diagram of proposed algorithm integrates CDR method with MVDR beamformer

Combination of MVDR Beamformer and CDR as Post-filter

The proposed framework adopts a parallel processing structure in which MVDR beamforming and CDR-based post-filtering operate independently. This design decouples spatial noise suppression from diffuseness estimation, as the CDR is computed directly from microphone-domain signals rather than from the beamformer output. Such an approach avoids bias and spatial colouration introduced by beamforming, limits error propagation between processing stages, and preserves the spatial coherence characteristics of the acoustic scene, leading to improved robustness in reverberant and diffuse noise conditions.

To obtain the final output $Z(m,f)$ of the speech enhancement approach, we multiply the weights of the MVDR beamformer $Y_B(m,f)$ signal by the post-filter gain $G(m,f)$, which is calculated using the CDR output of the microphone and output $Z(m,f)$ can be defined using Equation 18:

$$Z(m,f) = G(m,f)\, Y_B(m,f) \quad [18]$$

Finally, the signal is converted back to a time-domain waveform through the utilisation of the inverse short-time Fourier transform.

Experiment Setup and Data Collection

To allow a fair comparison of the results, we decided to use the same data generated in (Dietzen et al., 2020), where two methods were proposed, namely ISCLP SNR and ISCLP smooth SNR. Therefore, we considered the results from (Dietzen et al., 2020) as benchmark results to compare the results with our proposed algorithm. The experiment considers a target speaker, an interfering speaker, and noise in a reverberant environment. We denote the target reverberant speech, the interfering reverberant speech, and the babble noise component as $x1(t), x2(t)$, and v(t), respectively. To generate $x1(t)$ and $x2(t)$, using RIRs with a linear microphone array of M = 5 microphones, each 8 cm apart. We measured RIRs in a room with a reverberation time of 0.61s, with the source 2m away from the array. The target source and the interfering speaker were positioned at 2m from the microphone, with the target speaker placed at an angle of 0 degrees. Recorded babble noise at random locations 2m from the microphone. Both female and male speech source signals are considered in the experiment, and the babble noise is generated by a specific process. The simulations had a sampling frequency of fs = 16kHz, and square Hann windows with STFT = 512 samples and 50% overlap were used in STFT analysis and synthesis. All experiments were conducted offline on a MacBook Pro (14-inch, 2021) equipped with an Apple M1 Pro processor, 16 GB RAM, and a 10-core CPU (8 performance and 2 efficiency cores). The implementation was carried out in Python (version 3.9.12) using the *Pyroomacoustics* toolbox for acoustic simulation and signal processing.

Evaluation Metrics

To determine the effectiveness and to compare the results of the proposed method with the baseline work results, we consider four evaluation metrics, namely: PESQ (Rix et al., 2001), ESTOI, which calculates speech intelligibility without assuming mutual independence between frequency bands, which sets it apart from STOI's correlation analysis (Kobayashi & Kondo, 2019). Beyond these, other important evaluation measures include WSS and CD (Loizou, 2007). The PESQ metric is used to assess the perceived quality of the enhanced speech, with scores ranging from 1, representing very poor quality, to 4.5, representing excellent clarity. ESTOI produces values between 0 and 1, where scores closer to 1 reflect higher speech intelligibility and clarity. The CD was selected as an additional quality measure because it is computationally efficient and effectively captures spectral distortions in the enhanced signal. A higher CD value indicates greater deviation from the clean reference, while a lower value signifies minimal distortion and better signal fidelity.

RESULTS AND DISCUSSION

The proposed method results were compared with the baseline method results of the ISCLP SNR method, the ISCLP Smooth SNR method, which is illustrated using two evaluation

metrics, such as PESQ and ESTOI scores, as shown in Table 1. The performance of the proposed approach is evaluated against both baseline methods across various input SNR levels, ranging from 0 dB to 40 dB. As shown in Table 1, the proposed algorithm demonstrates slightly better performance than the baselines in terms of PESQ, indicating improved overall speech quality. Additionally, the ESTOI results of the proposed algorithm are slightly better than the baseline results at input SNR = 0dB, 5dB, and 40dB, with slightly lower results for other SNR values compared to the baseline result.

Table 2 presents the results of the ISCLP SNR method, the ISCLP Smooth SNR method, and the proposed method using CD and WSS scores. The performance of the proposed algorithm is compared to the baseline results of the ISCLP SNR method and the ISCLP smooth SNR method. The proposed algorithm was tested for different input SNR levels ranging from -20dB to 40dB. It can be observed from Table 2 that the performance of the proposed algorithm achieves the lowest value of CD for all input SNR levels ranging from 0dB to 40dB. The CD results indicate less distortion in the output compared to the two baseline methods. In addition, the WSS also attains the lowest value compared to the baseline results, indicating that the proposed method outperforms the baseline methods. From Table 1 and Table 2, it is clear that the proposed method performs better than the two baseline methods, especially for the metrics PESQ, CD, and WSS, across all input SNR levtheb), (c), and (d) display the time domain waveforms of the target speech, interfering speech, babble noise, and mixed speech, respectively, affected by an input SNR of 0dB and in a reverberant environment with rt60 = 0.61. The time domain waveforms of the output of the ISCLP SNR method, output of the ISCLP smooth SNR method, and output of the proposed method are shown in Figures 2e, 2f, and 2g, respectively. It is evident from Figure 2g compared to Figure 2e and 2f that the amplified speech signal is clearly seen at the output of the proposed method when compared to the results of the two baseline methods.

Table 1
Presents a comparison of the results obtained from the ISCLP SNR method, ISCLP smooth SNR method, and the proposed method using PESQ and ESTOI scores

SNR Level (dB)	ISCLP SNR PESQ	ISCLP Smooth PESQ	Proposed Method PESQ	ISCLP SNR ESTOI	ISCLP Smooth SNR ESTOI	Proposed Method ESTOI
0	1.0918	1.088	1.1192	0.4356	0.4389	0.5008
5	1.1574	1.1514	1.1785	0.5539	0.5535	0.5628
10	1.2166	1.2072	1.2199	0.6159	0.6199	0.6075
15	1.2444	1.2422	1.2601	0.648	0.6519	0.6334
20	1.2561	1.2562	1.2875	0.6614	0.6647	0.6484
25	1.2653	1.2632	1.2993	0.6666	0.6705	0.6574
30	1.2722	1.2726	1.3204	0.6686	0.6724	0.6643
35	1.2765	1.2754	1.3299	0.6674	0.6719	0.6689
40	1.2757	1.2762	1.3364	0.6663	0.6705	0.6719

Table 2
Presents a comparison of the results obtained from the ISCLP SNR method, ISCLP smooth SNR method, and the proposed method using CD and WSS scores

SNR Level (dB)	ISCLP SNR CD	ISCLP smooth SNR CD	Proposed method CD	ISCLP SNR WSS	ISCLP smooth SNR WSS	Proposed method WSS
0	6.3878	6.2441	5.7246	91.5309	85.5473	59.8161
5	5.6725	5.5536	5.2302	72.4086	67.8177	56.4334
10	5.2501	5.1336	4.9275	64.9385	60.8119	54.0634
15	5.0212	4.9056	4.7879	61.0293	56.8611	52.6873
20	4.8965	4.7925	4.7254	59.0557	55.2304	51.82
25	4.8181	4.7317	4.7008	57.6582	54.1183	51.0073
30	4.7847	4.6989	4.6887	57.0284	53.7325	50.6551
35	4.7674	4.6836	4.6742	56.7206	53.5247	50.2594
40	4.7571	4.6696	4.6549	56.6754	53.4617	50.1757

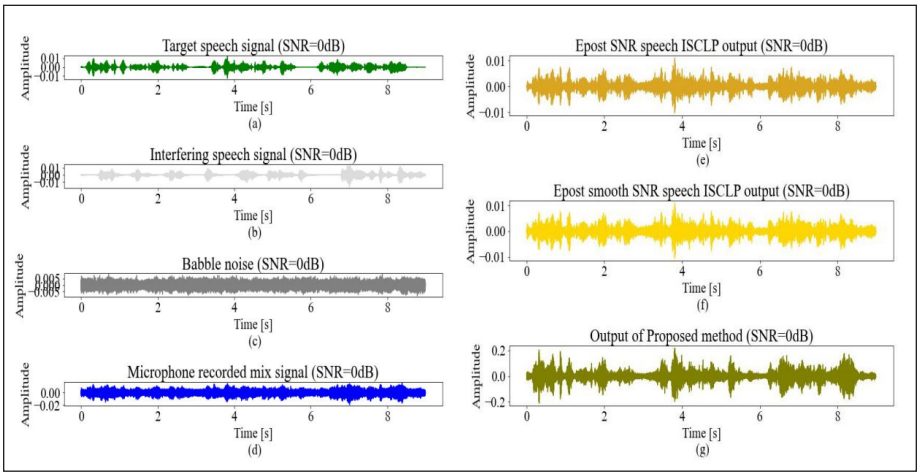

Figure 2. Different time-domain input and output waveforms processed at input SNR = 0 dB. (a) Target speech signal; (b) Interfering speech signal; (c) Babble noise; (d) Microphone-recorded mixed signal; (e) Epost SNR speech ISCLP output; (f) Epost smoothed SNR speech ISCLP output; (g) Proposed method output

Figure 3a, 3b, 3c, and 3d display the time domain waveforms of the target speech, interfering speech, babble noise, and mixed speech, respectively, affected by an input SNR of 40dB and in a reverberant environment with rt60 = 0.61. The time domain waveforms of the output of the ISCLP SNR method, the output of the ISCLP smooth SNR method, and the output of the proposed method are shown in Figures 3e, 3f, and 3g, respectively. Similarly, it is evident from Figure 3g compared to Figure 3e and 3f that the amplified speech signal is clearly observed at the output of the proposed method when compared to the results of the two baseline methods.

The proposed method consistently achieves higher PESQ scores than the ISCLP baseline across all tested SNR conditions, as shown in Table 1, while improvements in ESTOI remain modest. In contrast, Table 2 demonstrates a clear and consistent reduction in CD and WSS values. These reductions indicate improved spectral fidelity and reduced speech distortion, suggesting that the proposed framework enhances speech naturalness even when intelligibility gains are limited. In practical communication scenarios such as teleconferencing and hearing assistance, such improvements are associated with reduced listening effort and improved long-term listening comfort.

The proposed framework combines spatial filtering and diffuseness-based suppression to enhance speech in reverberant environments. Although the CDR estimator operates directly on microphone-domain signals while enhancement is applied to the MVDR beamformer output, the system exhibits stable behaviour under the evaluated conditions. The MVDR stage attenuates spatially off-target and incoherent components of the babble noise field, which reduces the sensitivity of the overall system to representation mismatch between microphone-domain CDR estimation and beamformer-domain enhancement. Performance under strong directional interference was not explicitly evaluated and represents an important direction for future investigation.

While MVDR beamforming contributes to spatial noise reduction, residual reverberation and diffuse noise components remain, particularly in environments with moderate to high reverberation. The consistent improvements observed in PESQ, CD and WSS across all tested SNR levels can therefore be primarily attributed to the CDR-based post-filter, which explicitly models and suppresses diffuse energy. By complementing MVDR beamforming, the post-filter improves spectral fidelity beyond what can be achieved through spatial selectivity alone.

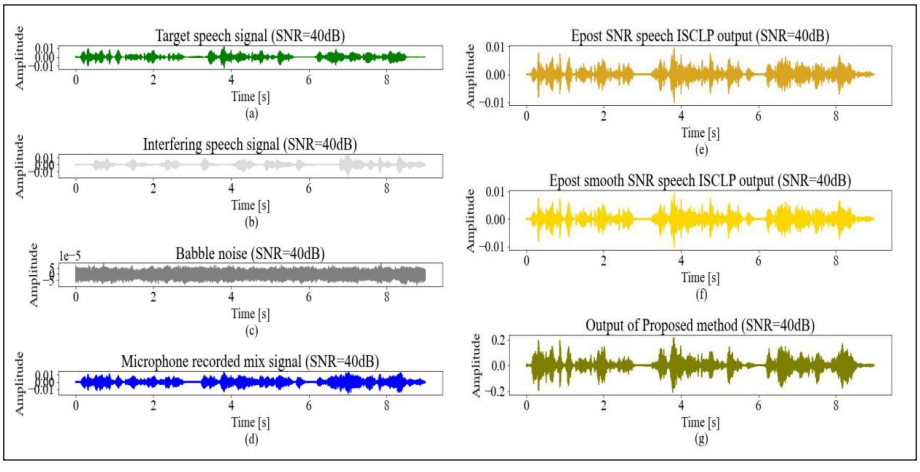

Figure 3. Different time-domain input and output waveforms processed at input SNR = 40 dB. (a) Target speech signal; (b) Interfering speech signal; (c) Babble noise; (d) Microphone-recorded mixed signal; (e) Epost SNR speech ISCLP output; (f) Epost smoothed SNR speech ISCLP output; (g) Proposed method output

Beyond objective metric comparison, the proposed framework differs qualitatively from ISCLP in its architectural behaviour during speech pauses and low-energy segments. The CDR estimation is derived from temporally smoothed cross-power spectral statistics of the microphone signals and does not rely on explicit speech activity detection. Consequently, the diffuseness-based post-filter remains active during silent or low-energy intervals and primarily reflects the characteristics of background noise and reverberation. Similarly, the MVDR beamformer continues to minimise spatially correlated noise power irrespective of speech presence. While pause-specific behaviour was not explicitly analysed, the system design supports continued suppression of diffuse and reverberant components during such segments.

Formal listening tests were not conducted as part of this study. However, informal inspection of the enhanced signals did not reveal severe musical noise artefacts. Nevertheless, subtle speech colouration effects may still be present and are not fully captured by objective metrics such as PESQ and ESTOI. A comprehensive perceptual evaluation is therefore identified as an important direction for future work.

From a computational perspective, the proposed framework relies on closed-form MVDR beamforming and coherence-based CDR estimation and does not involve iterative state-space tracking or adaptive filtering stages such as those employed in the ISCLP baseline, which incorporates Kalman filtering for interference power estimation. Consequently, the proposed method primarily consists of matrix operations and recursive spectral averaging, resulting in reduced computational complexity. Although all experiments were conducted offline and explicit runtime measurements were not reported, the algorithmic structure suggests suitability for real-time implementation on moderate embedded hardware.

Overall, the proposed framework effectively integrates spatial filtering and diffuseness-based suppression to improve speech quality under reverberant conditions while maintaining stable behaviour and low computational complexity.

CONCLUSION

This study presents a new framework that integrates the CDR estimator with the MVDR beamformer to address both reverberation and suppression of babble noise in complex acoustic settings. The CDR estimator in the proposed system supports spectral enhancement through SNR estimation and acts as a post-filter to suppress residual reverberation. The proposed approach improves robustness against babble noise and late reflections. The experimental results have confirmed that the proposed method has shown slightly better evaluation metric scores of PESQ, CD, and WSS as compared to two ISCLP-based baseline systems, especially the perceptual quality, and spectral accuracy improvement was observed to be better in the proposed method over the baseline method. The experiment has been tested considering moderate reverberant conditions with RT60 = 0.61 s, with one source

and babble noise. The experiment was tested by recording the voice of the speaker at 2m from the microphone. The proposed method assumes the presence of a single dominant target speaker. In scenarios involving multiple speakers with comparable power and partial spatial overlap, steering vector estimation and coherence-based separation may become unreliable, potentially leading to target suppression or interference leakage. Addressing multi-speaker conditions remains an important extension of this work. In addition, future studies will investigate the robustness of the proposed framework under increased source-to-microphone distances beyond 2 m, diverse noise conditions, and higher reverberation levels. The correction factor used to adapt the microphone-domain CDR estimate to the beamformer output was evaluated under fixed array geometry and stationary acoustic conditions. While stable performance was observed in this setting, rapidly changing acoustic scenes or microphone array perturbations were not investigated. Future work will focus on adaptive correction mechanisms to improve robustness under dynamic conditions. The experimental evaluation was limited to babble noise, which predominantly exhibits diffuse characteristics. Performance under transient, impulsive, or strongly directional interference sources was not evaluated and therefore remains an open research question to be explored in future studies. The proposed framework assumes a fixed linear array geometry consisting of five microphones with an inter-element spacing of 8 cm. Changes in array spacing, the number of microphones, or physical array deformation can affect spatial covariance estimation and steering vector computation, which in turn may influence the performance of MVDR beamforming and the subsequent CDR-based post-filter. The current study does not explicitly evaluate sensitivity to such array variations, and investigating robustness under different array configurations remains an important direction for future work.

The proposed framework offers several practical advantages over learning-based approaches due to its fully analytical formulation. The method provides interpretability and predictable behaviour, does not require training data, and exhibits low computational complexity. These properties make it particularly suitable for real-time and resource-constrained applications such as hearing aids, embedded communication devices, and safety-critical speech enhancement systems, where robustness, transparency, and reliable behaviour across acoustic conditions are essential.

ACKNOWLEDGEMENT

This research was funded by the Research Management Centre, Universiti Putra Malaysia, Grant Putra 9695500.

REFERENCES

Bai, J., Li, H., Zhang, X., & Chen, F. (2025). Attention-based beamformer for multi-channel speech enhancement. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5. https://doi.org/10.1109/ICASSP49660.2025.10890720

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 120(1), 331-342. https://doi.org/10.1121/1.2202888

Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukic, A., Gerkmann, T., Doclo, S., & Goetze, S. (2015). Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP Journal on Advances in Signal Processing*, 2015, Article 61, 1-12. https://doi.org/10.1186/s13634-015-0242-x

Chang, H., Hsu, Y., & Bai, M. R. (2024). Deep beamforming for speech enhancement and speaker localisation with an array response-aware loss function. *Frontiers in Signal Processing*, 4, 1-8. https://doi.org/10.3389/frsip.2024.1413983

Cohen, N., Hazan, G., Schwartz, B., & Gannot, S. (2021). An online algorithm for echo cancellation, dereverberation and noise reduction based on a Kalman-EM method. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021, Article 33, 1-17. https://doi.org/10.1186/s13636-021-00219-2

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., & Nakatani, T. (2015). Strategies for distant speech recognition in reverberant environments. *EURASIP Journal on Advances in Signal Processing*, 2015, Article 1, 1-15. https://doi.org/10.1186/s13634-015-0245-7

Dietzen, T., Huleihel, N., Doclo, S., Moonen, M., Waterschoot, T. V., Spriet, A., & Tirry, W. (2015). Speech dereverberation by data-dependent beamforming with signal pre-whitening. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, Nice, France. https://doi.org/10.1109/EUSIPCO.2015.7362827

Dietzen, T., Doclo, S., Moonen, M., & Waterschoot, T. V. (2018). Joint multimicrophone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction. In *Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC)* (pp. 221-225). https://doi.org/10.1109/IWAENC.2018.8521250

Dietzen, T., Doclo, S., Moonen, M., & Waterschoot, T. V. (2020). Integrated sidelobe cancellation and linear prediction Kalman filter for joint multimicrophone speech dereverberation, interfering speech cancellation, and noise reduction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 740-754. https://doi.org/10.1109/TASLP.2020.2966869

Fujita, Y., Nugraha, A. A., Di Carlo, D., Bando, Y., Fontaine, M., & Yoshii, K. (2024). Run-time adaptation of neural beamforming for robust speech dereverberation and denoising. In *Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. https://doi.org/10.1109/APSIPAASC63619.2025.10849318

Gode, H., Tammen, M., & Doclo, S. (2021). Joint multi-channel dereverberation and noise reduction using a unified convolutional beamformer with sparse priors. In *Proceedings of the ITG Conference on Speech Communication* (pp. 144-148). https://doi.org/10.48550/arXiv.2106.01902

Gode, H., & Doclo, S. (2023). Adaptive dereverberation, noise and interferer reduction using sparse weighted linearly constrained minimum power beamforming. *arXiv*. https://doi.org/10.23919/EUSIPCO55093.2022.9909809

Habets, E. A. P., & Benesty, J. (2013). A two-stage beamforming approach for noise reduction and dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5), 945-958. https://doi.org/10.1109/TASL.2013.2239292

Huang, G., Jensen, J. R., Chen, J., Benesty, J., Christensen, M. G., Sugiyama, A., Elko, G., & Gaensler, T. (2025). Advances in microphone array processing and multichannel speech enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 1-5). https://doi.org/10.1109/ICASSP49660.2025.10888510

Kobayashi, Y., & Kondo, K. (2019). Japanese speech intelligibility estimation and prediction using objective intelligibility indices under noisy and reverberant conditions. *Applied Acoustics*, 156, 327-335. https://doi.org/10.1016/j.apacoust.2019.07.034

Kumar, C. R., & Chitra, M. P. (2022). Implementation of modified Wiener filtering in frequency domain in speech enhancement. *International Journal of Advanced Computer Science and Applications*, 13(2), 434-439.https://doi.org/10.14569/IJACSA.2022.0130251

Le, X., Chen, H., Chen, K., & Lu, J. (2021). DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. In *Proceedings of Interspeech* (pp. 2811-2815). https://doi.org/10.21437/Interspeech.2021-296

Lee, C. H., Patel, K., Yang, C., Shen, Y., & Jin, H. (2024). An MVDR-embedded U-Net beamformer for effective and robust multichannel speech enhancement. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8541-8545). https://doi.org/10.1109/ICASSP48485.2024.10448366

Li, A., Liu, W., Zheng, C., & Li, X. (2022). Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 6487-6491). https://doi.org/10.1109/ICASSP43922.2022.9746432

Liu, J., & Zhang, X. (2021). Inplace gated convolutional recurrent neural network for dual-channel speech enhancement. *arXiv*. https://doi.org/10.21437/Interspeech.2021-899

Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. CRC Press.

Luo, Y., Han, C., Mesgarani, N., Ceolini, S. C., & Liu, E. (2019). FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 260-267). https://doi.org/10.1109/ASRU46091.2019.9003849

Nakatani, T., & Kinoshita, K. (2019a). A unified convolutional beamformer for simultaneous denoising and dereverberation. *IEEE Signal Processing Letters*, 26(6), 903-907. https://doi.org/10.21437/Interspeech.2019-1286

Nakatani, T., & Kinoshita, K. (2019b). Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer. In *Proceedings of Interspeech*, Graz, Austria (pp. 111-115). https://doi.org/10.21437/Interspeech.2019-1286

Natarajan, S., Al-Haddad, S. A. R., Hassan, M. K., Kamil, R., Azrad, S., Ahmad, F. A., & Macleans, J. F. (2023). A comprehensive review of beamforming-based speech enhancement techniques, IoT, and smart

city applications. In *Proceedings of the 2023 IEEE 2nd Industrial Electronics Society Annual On-Line Conference (ONCON)* (pp. 1-6). https://doi.org/10.1109/ONCON60463.2023.10431158

Natarajan, S., Al-Haddad, S. A. R., Hassan, M. K., Kamil, R., Azrad, S., Ahmad, F. A., & Macleans, J. F. (2024). Revolutionising speech clarity: Unveiling a novel approach with hybrid coherent-to-diffuse power ratio and recursive least square algorithm for reverberation removal. In *Proceedings of the 2024 8th International Conference on Digital Signal Processing (ICDSP '24)* (pp. 140-145). https://doi.org/10.1145/3653876.3653887

Natarajan, S., Al-Haddad, S. A. R., Ahmad, F. A., Kamil, R., Hassan, M. K., Azrad, S., Macleans, J. F., Abdulhussain, S. H., Mahmmod, B. M., Saparkhojayev, N., & Dauitbayeva, A. (2025). Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Engineering Journal*, 16(7), Article 103405, 1-35. https://doi.org/10.1016/j.asej.2025.103405

O'Shaughnessy, D. (2024). Speech enhancement-A review of modern methods. *IEEE Transactions on Human-Machine Systems*, 54(1), 110-120. https://doi.org/10.1109/THMS.2023.3339663

Ren, X., Zhang, X., Chen, L., Zheng, X., Zhang, C., Guo, L., & Yu, B. (2021). A causal U-Net based neural beamforming network for real-time multi-channel speech enhancement. In *Proceedings of Interspeech* (pp. 1832-1836). https://doi.org/10.21437/Interspeech.2021-1457

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vol. 7, pp. 749-752). https://doi.org/10.1109/ICASSP.2001.941023

Sarradj, E. (2010). A fast signal subspace approach for the determination of absolute levels from phased microphone array measurements. *Journal of Sound and Vibration*, 329(9), 1553-1569. https://doi.org/10.1016/j.jsv.2009.11.009

Schwarz, A., & Kellermann, W. (2015). Coherent-to-diffuse power ratio estimation for dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6), 1006-1018. https://doi.org/10.1109/TASLP.2015.2418571

Shankar, N., Bhat, G. S., Panahi, I. M. S., Tittle, S., & Thibodeau, L. M. (2021). Smartphone-based single-channel speech enhancement application for hearing aids. *The Journal of the Acoustical Society of America*, 150(3), 1663-1673.https://doi.org/10.1121/10.0006045

Song, S., Cheng, L., Luan, S., Yao, D., Li, J., & Yan, Y. (2021). An integrated multi-channel approach for joint noise reduction and dereverberation. *Applied Acoustics*, 171, Article 107526. https://doi.org/10.1016/j.apacoust.2020.107526

Su, G., & Morf, M. (1983). The signal subspace approaches multiple wide-band emitter locations. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(6), 1502-1522. https://doi.org/10.1109/TASSP.1983.1164233

Tan, K., & Wang, D. L. (2019). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 380-390. https://doi.org/10.1109/TASLP.2019.2955276

Tan, K., Wang, Z. Q., & Wang, D. L. (2022). Neural spectrospatial filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 605-621. https://doi.org/10.1109/TASLP.2022.3145319

Thimmraja, Y. G., Nagaraja, B. G., & Jayanna, H. S. (2022). A spatial procedure to spectral subtraction for speech enhancement. *Multimedia Tools and Applications*, 81, 23633-23647. https://doi.org/10.1007/s11042-022-12152-3

Xi, J., Xu, Z., Zhang, W., Zhao, L., & Xie, Y. (2024). Speech enhancement algorithm based on microphone array and lightweight CRN for hearing aid. *Electronics*, 13(22), Article 4394. https://doi.org/10.3390/electronics13224394

Xiang, Q., Chen, J., Benesty, J., Lei, T., & Pan, C. (2025). Design of the Wiener gain in noisy and reverberant environments. *Applied Acoustics*, 231, Article 110491. https://doi.org/10.1016/j.apacoust.2024.110491

Yang, W., Huang, G., Brendel, A., Chen, J., Benesty, J., Kellermann, W., & Cohen, I. (2022). A bilinear framework for adaptive speech dereverberation combining beamforming and linear prediction. In *Proceedings of the 2022 International Workshop on Acoustic Signal Enhancement (IWAENC)* (pp. 1-5). https://doi.org/10.1109/IWAENC53105.2022.9914728

Zhao, Y., Jensen, J. R., Jensen, T. L., Chen, J., & Christensen, M. G. (2020). Experimental study of robust acoustic beamforming for speech acquisition in reverberant and noisy environments. *Applied Acoustics*, 170, Article 107531, 1-13. https://doi.org/10.1016/j.apacoust.2020.107531