

# A Bilinear Source Separation, Dereverberation, and Background Noise Suppression Algorithm for Augmented Reality Applications

Alon Nemirovsky      Gal Itzhak      Israel Cohen

Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering  
Technion – Israel Institute of Technology, Haifa 3200003, Israel

## ABSTRACT

This paper introduces a new bilinear algorithm that combines distortionless response beamforming with weighted prediction error using the Kronecker product operator to improve speech signal quality in different acoustic environments. The algorithm utilizes recursive least squares cost functions and handles real dynamic, noisy, and reverberant conditions effectively. Compared to a recently proposed approach, our approach outperforms separating the desired source from undesired sources while providing dereverberation and noise suppression. It is also preferable regarding the quality of the enhanced signals it produces. In contrast, it exhibits a more considerable desired signal distortion and reduced intelligibility. The algorithm's computational efficiency and robustness make it suitable for real-time applications, as validated using real recordings from the Speech Enhancement for Augmented Reality (SPEAR) challenge.

**Index Terms**—Source separation, beamforming, dereverberation, speech enhancement, SPEAR challenge.

## I. INTRODUCTION

Real-time source separation, dereverberation, and background noise suppression are essential for improving the quality and intelligibility of speech signals in various acoustic environments [1]–[5]. These tasks involve recovering individual sound sources from a mixture of signals, which is critical for applications such as augmented reality [6], conference calls, and hearing aids. In particular, the demand for effective solutions is heightened in static and dynamic environments, with the latter posing a more significant challenge due to rapidly changing acoustic conditions.

Deep neural networks (DNNs) have shown promise for source separation in recent years due to their ability to model complex acoustic data [7]. However, their high data and computational demands and limited generalization to unseen scenarios make them less suitable for real-time applications in dynamic environments. Beamforming algorithms, on the other hand, provide an efficient, low-latency solution that adapts well to changing conditions and requires minimal training. This makes beamforming a practical choice for dynamic and resource-constrained scenarios.

Beamforming algorithms have been central to isolating desired speech and enhancing speech quality in noisy environments. The minimum variance distortionless response (MVDR) beamformer effectively separates sources when the exact noise covariance matrix is provided. However, modeling noise covariance often requires additional data, such as a voice activity detector (VAD) [8], making it less practical. On the other hand, while applicable, the minimum power distortionless response (MPDR) beamformer is sensitive to inaccuracies in the steering vector [9]. This sensitivity highlights the need for more robust methods to address steering vector errors while maintaining performance.

In addition to the challenges of source separation, reverberation significantly affects speech quality and intelligibility, particularly in

enclosed environments. To address this, dereverberation techniques suppress late reflections while preserving both the direct signal and early reflections. Among these methods, multichannel linear prediction (MCLP) has gained prominence for its ability to predict and subtract late reverberation components from observed signals [2]. MCLP can be applied in both the time domain and the short-time Fourier transform (STFT) domain [10]. In particular, the weighted prediction error (WPE) algorithm has demonstrated strong performance in reducing late reverberation effects [11], [12]. However, computational complexity remains a significant challenge, especially for real-time applications on embedded devices. Additionally, additive noise in the environment can further degrade performance. As a result, in noisy environments, dereverberation must be combined with additive noise cancellation to achieve optimal performance.

To achieve denoising and dereverberation simultaneously, various beamforming techniques such as the generalized sidelobe canceller (GSC) [13], MVDR [14], and weighted MPDR (wMPDR) [15], [16] have been combined with MCLP-based dereverberation methods. While these methods enhance overall performance, they also increase computational complexity. A more efficient approach involves using the Kronecker product (KP) to merge spatial and temporal filters, allowing sub-optimal solutions to reduce computational demands [17]–[21]. Integrating spatial and temporal filtering through KP has shown promise in reducing noise and improving dereverberation [22]. However, there are still challenges in accurately separating individual sound sources.

We propose a bilinear framework based on recursive least squares (RLS) to address these challenges. This framework integrates the MPDR beamformer with the weighted prediction error (WPE) method using the KP. This approach allows for joint source separation, noise reduction, and suppression of late reverberation while introducing tradeoffs between source separation and noise suppression. We validate our method using real recordings from the SPEAR challenge [23], [24], showcasing its effectiveness in real-world scenarios that require robust and efficient audio processing.

The remainder of the paper is organized as follows: Section II presents the signal model. Section III introduces the bilinear framework model that integrates beamforming with dereverberation. Section IV details the application of the bilinear framework in developing a real-time RLS-based algorithm for dynamic environments. In Section V, we present the results of the proposed algorithm and discuss the tradeoffs involved in the source separation task.

## II. SIGNAL MODEL

We consider using SPEAR real data to represent a dynamic, reverberant, and noisy environment. In this setup,  $M$  microphones are deployed to capture multiple speech sources. The signals are analyzed in the STFT domain, where the signal received by the  $m$ -th microphone is denoted as  $Y_m(\ell, k)$ , with  $\ell$  representing the time

This research was supported by the Israel Science Foundation (grant no. 1449/23) and the Pazy Research Foundation.

frame and  $k$  representing the frequency bin. The observed signals can be modeled as follows:

$$\mathbf{y}(\ell, k) = \mathbf{d}(\ell, k)X(\ell, k) + \mathbf{r}(\ell, k) + \gamma(\ell, k) + \sum_i \mathbf{m}^{(i)}(\ell, k). \quad (1)$$

The vector of the microphone signals,  $\mathbf{y}(\ell, k)$ , is defined as:

$$\mathbf{y}(\ell, k) = [Y_1(\ell, k) \quad \cdots \quad Y_M(\ell, k)]^T. \quad (2)$$

The vector  $\mathbf{d}(\ell, k)$  represents the time-dependent relative transfer function (RTF) for each microphone concerning a reference microphone and is defined as:

$$\mathbf{d}(\ell, k) = \left[ \frac{\mathbf{H}_1(\ell, k)}{\mathbf{H}_{ref}(\ell, k)} \quad \cdots \quad \frac{\mathbf{H}_M(\ell, k)}{\mathbf{H}_{ref}(\ell, k)} \right]^T. \quad (3)$$

This formulation normalizes each microphone's transfer function by the reference microphone's transfer function, reflecting the relative acoustic transfer paths.  $X(\ell, k)$  denotes the STFT coefficient of the desired speech source at the reference microphone. The vectors  $\mathbf{r}(\ell, k)$  and  $\gamma(\ell, k)$  denote the late reverberant components and background noise, respectively. The term  $\sum_i \mathbf{m}^{(i)}(\ell, k)$  models other speech sources that are considered as directional speech interferences.

The task of blind joint dereverberation, noise reduction, and interference suppression is to estimate  $X(\ell, k)$ , the desired source signal at the reference microphone, from the noisy and reverberant observations  $\mathbf{y}(\ell, k)$ . The challenge lies in suppressing late reverberation, background noise, and directional interferences while preserving the desired source signal. For simplicity, the frequency index  $k$  is omitted in the following sections.

### III. BEAMFORMING AND DEREVERBERATION BILINEAR FRAMEWORK

Inspired by the bilinear framework introduced in [22], we propose combining spatial and temporal adaptive filters using the KP. This framework facilitates joint directional noise reduction and dereverberation by estimating  $X(\ell, k)$  from  $\mathbf{y}(\ell, k)$ . First, we apply a spatial filter,  $\mathbf{h}$ , to the observation signal vector:

$$\begin{aligned} Z(\ell) &= \mathbf{h}^H \mathbf{y}(\ell) \\ &= \mathbf{h}^H \left( X(\ell) \mathbf{d}(\ell) + \mathbf{r}(\ell) + \gamma(\ell) + \sum_i \mathbf{m}^{(i)}(\ell) \right) \\ &= X(\ell) + \mathbf{h}^H \mathbf{v}(\ell), \end{aligned} \quad (4)$$

where  $\mathbf{v}(\ell)$  represents the linear combination of the undesired signals, and the superscript  $H$  denotes the conjugate-transpose operator. We assume uncorrelated zero-mean signals and enforce the distortionless constrain  $\mathbf{h}^H \mathbf{d}(\ell) = 1$  for each frequency bin.

Next, dereverberation is achieved by subtracting the estimated reverberant signal components using a linear filter of length  $L$  on the beamformer output signal:

$$\hat{X}(\ell) = Z(\ell) - \mathbf{g}^H \mathbf{z}(\ell - \Delta). \quad (5)$$

Here,  $\mathbf{g}$  is the prediction filter, and  $\Delta$  is the prediction delay, which aims to preserve the correlation between the samples of the clean speech signal. The vector  $\mathbf{z}(\ell - \Delta)$  includes the beamforming outputs from the previous consecutive frames and is defined as:

$$\mathbf{z}(\ell - \Delta) = [Z(\ell - \Delta) \quad \cdots \quad Z(\ell - \Delta - L + 1)]^T. \quad (6)$$

Assuming fixed  $\mathbf{h}^*$ , the second term in (5) is linear in  $\mathbf{g}^*$ , or vice versa. This term can be further optimized using the KP representation to estimate the desired signal:

$$\hat{X}(\ell) = Z(\ell) - (\mathbf{g} \otimes \mathbf{h})^H \bar{\mathbf{y}}(\ell - \Delta), \quad (7)$$

where the superscript  $*$  is the complex conjugate operator,  $\otimes$  is the KP operator, and  $\bar{\mathbf{y}}(\ell - \Delta)$  is the vectorized delayed observation signal matrix with a length of  $ML$ :

$$\bar{\mathbf{y}}(\ell - \Delta) = \text{vec}([\mathbf{y}(\ell - \Delta) \quad \cdots \quad \mathbf{y}(\ell - \Delta - L + 1)]). \quad (8)$$

By leveraging the relationship [25]:

$$\mathbf{g} \otimes \mathbf{h} = (\mathbf{I}_L \otimes \mathbf{h}) \mathbf{g} = (\mathbf{g} \otimes \mathbf{I}_M) \mathbf{h}, \quad (9)$$

we can decouple the spatial and temporal filters, allowing for efficient adaptive processing in real-time applications where  $\mathbf{I}$  is the identity matrix with the appropriate dimension:

$$\begin{aligned} \hat{X}_\ell &= Z_{\mathbf{h}}(\ell) - \mathbf{g}^H (\mathbf{I}_L \otimes \mathbf{h})^H \bar{\mathbf{y}}(\ell - \Delta) \\ &= Z_{\mathbf{h}}(\ell) - \mathbf{g}^H \bar{\mathbf{y}}_{\mathbf{h}}(\ell - \Delta), \end{aligned} \quad (10)$$

where  $\bar{\mathbf{y}}_{\mathbf{h}} = (\mathbf{I}_L \otimes \mathbf{h})^H \bar{\mathbf{y}}(\ell - \Delta)$  is the observation signal vector filtered by  $\mathbf{h}$  and  $Z_{\mathbf{h}}(\ell)$  is the beamformer output of  $\mathbf{h}$ . Similarly, we can also write (7) as

$$\begin{aligned} \hat{X}_\ell &= \mathbf{h}^H \mathbf{y}(\ell) - \mathbf{h}^H (\mathbf{g} \otimes \mathbf{I}_M)^H \bar{\mathbf{y}}(\ell - \Delta) \\ &= \mathbf{h}^H \bar{\mathbf{y}}_{\mathbf{g}}(\ell - \Delta), \end{aligned} \quad (11)$$

where  $\bar{\mathbf{y}}_{\mathbf{g}}(\ell - \Delta) = \mathbf{y}(\ell) - (\mathbf{g} \otimes \mathbf{I}_M)^H \bar{\mathbf{y}}(\ell - \Delta)$  is the observation signal vector filtered by  $\mathbf{g}$ .

### IV. REAL-TIME RLS BASED BILINEAR FRAMEWORK

The algorithm can be derived for real-time source separation and dereverberation using the RLS algorithm [26] with the time-varying spatial filter  $\mathbf{h}(\ell)$  and the temporal filter  $\mathbf{g}(\ell)$ . These filters can be iteratively optimized by defining the following cost functions:

$$J[\mathbf{h}(\ell)|\mathbf{g}(\ell - 1)] = \sum_{i=1}^{\ell} \alpha^{\ell-i} |\mathbf{h}^H(\ell) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(i - \Delta)|^2, \quad (12)$$

$$J[\mathbf{g}(\ell)|\mathbf{h}(\ell - 1)] = \sum_{i=1}^{\ell} \alpha^{\ell-i} \frac{|Z_{\mathbf{h}(\ell-1)}(i) - \mathbf{g}^H(\ell) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(i - \Delta)|^2}{\lambda(i)}, \quad (13)$$

where

$$\bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(i - \Delta) = \mathbf{y}(\ell) - (\mathbf{g}(\ell - 1) \otimes \mathbf{I}_M)^H \bar{\mathbf{y}}(\ell - \Delta) \quad (14)$$

$$\bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(i - \Delta) = (\mathbf{I}_L \otimes \mathbf{h}(\ell - 1))^H \bar{\mathbf{y}}(i - \Delta) \quad (15)$$

$\lambda(\ell) = |\hat{X}(\ell)|^2$  is the variance of the a priori estimate of the desired signal, and  $\alpha$  is the forgetting factor. The solution for the temporal filter  $\mathbf{g}(\ell)$  can be obtained by minimizing the cost function  $J[\mathbf{g}(\ell)|\mathbf{h}(\ell - 1)]$ :

$$\mathbf{g}(\ell) = \mathbf{R}_{\mathbf{h}}^{-1}(\ell) \mathbf{p}_{\mathbf{h}}(\ell), \quad (16)$$

where  $\mathbf{R}_{\mathbf{h}}(\ell)$  is the weighted covariance matrix of  $\bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell - \Delta)$  and  $\mathbf{p}_{\mathbf{h}}(\ell)$  is the weighted correlation vector between  $\bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell - \Delta)$  and  $Z_{\mathbf{h}(\ell-1)}(\ell)$

$$\mathbf{R}_{\mathbf{h}}(\ell) = \alpha \mathbf{R}_{\mathbf{h}}(\ell - 1) + \frac{\bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell - \Delta) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}^H(\ell - \Delta)}{\lambda(\ell)} \quad (17)$$

$$\mathbf{p}_{\mathbf{h}}(\ell) = \sum_{i=1}^{\ell} \alpha^{\ell-i} \frac{\bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(i - \Delta) Z_{\mathbf{h}(\ell-1)}^*(i)}{\lambda(i)} \quad (18)$$

The spatial beamformer can be optimized by minimizing  $J[\mathbf{h}(\ell)|\mathbf{g}(\ell - 1)]$  under the dynamic distortionless constraint:

$$\min_{\mathbf{h}(\ell)} J[\mathbf{h}(\ell)|\mathbf{g}(\ell - 1)] \quad \text{s.t.} \quad \mathbf{h}^H(\ell) \mathbf{d}(\ell) = 1, \quad (19)$$

whose solution is the MPDR beamformer:

$$\mathbf{h}(\ell) = \frac{\mathbf{R}_{\mathbf{g}}^{-1}(\ell) \mathbf{d}(\ell)}{\mathbf{d}(\ell)^H \mathbf{R}_{\mathbf{g}}^{-1}(\ell) \mathbf{d}(\ell)}, \quad (20)$$

---

**Algorithm 1** MPDR-WPE Bilinear Framework
 

---

**Initialization:**  $\mathbf{g}(0)$ ,  $\mathbf{h}(0)$ ,  $\mathbf{R}_{\mathbf{g}}^{-1}(0)$ ,  $\mathbf{R}_{\mathbf{h}}^{-1}(0)$ ,  $\epsilon^{-1}$

- 1: **for**  $\ell = 1, 2, \dots$  **do**
- 2:    $\bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell - \Delta) = \mathbf{y}(\ell) - (\mathbf{g}(\ell - 1) \otimes \mathbf{I}_M)^H \bar{\mathbf{y}}(\ell - \Delta)$
- 3:    $\bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell - \Delta) = (\mathbf{I}_L \otimes \mathbf{h}(\ell - 1))^H \bar{\mathbf{y}}(\ell - \Delta)$
- 4:    $\lambda(\ell) = \|\mathbf{h}^H(\ell - 1) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell - \Delta)\|^2$
- 5:    $\mathbf{k}_{\mathbf{g}}(\ell) = \frac{\mathbf{R}_{\mathbf{g}}^{-1}(\ell-1) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell-\Delta)}{\alpha + \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}^H(\ell-\Delta) \mathbf{R}_{\mathbf{g}}^{-1}(\ell-1) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell-\Delta)}$
- 6:    $\mathbf{k}_{\mathbf{h}}(\ell) = \frac{\mathbf{R}_{\mathbf{h}}^{-1}(\ell-1) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell-\Delta)}{\alpha \lambda(\ell) + \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}^H(\ell-\Delta) \mathbf{R}_{\mathbf{h}}^{-1}(\ell-1) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell-\Delta)}$
- 7:    $\mathbf{R}_{\mathbf{g}}^{-1}(\ell) = \frac{\mathbf{R}_{\mathbf{g}}^{-1}(\ell-1) - \mathbf{k}_{\mathbf{g}}(\ell) (\mathbf{R}_{\mathbf{g}}^{-1}(\ell-1) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell-\Delta))^H}{\alpha}$
- 8:    $\mathbf{R}_{\mathbf{h}}^{-1}(\ell) = \frac{\mathbf{R}_{\mathbf{h}}^{-1}(\ell-1) - \mathbf{k}_{\mathbf{h}}(\ell) (\mathbf{R}_{\mathbf{h}}^{-1}(\ell-1) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell-\Delta))^H}{\alpha}$
- 9:    $\mathbf{g}(\ell) = \mathbf{g}(\ell - 1) + \mathbf{k}_{\mathbf{g}}(\ell) \hat{X}^*(\ell)$
- 10:    $\mathbf{R}_{\sigma}^{-1}(\ell) = \mathbf{R}_{\mathbf{g}}^{-1}(\ell) - \mathbf{R}_{\mathbf{g}}^{-1}(\ell) (\epsilon^{-1} \mathbf{I}^{-1} + \mathbf{R}_{\mathbf{g}}^{-1}(\ell))^{-1} \mathbf{R}_{\mathbf{g}}^{-1}(\ell)$
- 11:    $\mathbf{h}(\ell) = \frac{\mathbf{R}_{\sigma}^{-1}(\ell) \mathbf{d}(\ell)}{\mathbf{d}^H(\ell) \mathbf{R}_{\sigma}^{-1}(\ell) \mathbf{d}(\ell)}$
- 12: **end for**

---

with

$$\mathbf{R}_{\mathbf{g}}(\ell) = \alpha \mathbf{R}_{\mathbf{g}}(\ell - 1) + \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell - \Delta) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}^H(\ell - \Delta) \quad (21)$$

being the covariance matrix of  $\bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell - \Delta)$ .

In practice, due to inevitable steering errors, regularization of  $\mathbf{R}_{\mathbf{g}}(\ell)$  is necessary. This regularization increases the white noise gain (WNG), making the MPDR more robust and broadening the beampattern near the desired direction. The regularized covariance matrix is:

$$\mathbf{R}_{\sigma}(\ell) = \mathbf{R}_{\mathbf{g}}(\ell) + \epsilon \mathbf{I}_M, \quad (22)$$

where  $\epsilon \mathbf{I}_M$  is a diagonal matrix with small values corresponding to the dimension of the sensors. To improve computational complexity, we can apply Woodbury's identity [26]. The updates of  $\mathbf{R}_{\mathbf{h}}^{-1}(\ell)$ ,  $\mathbf{R}_{\mathbf{g}}^{-1}(\ell)$ , and  $\mathbf{R}_{\sigma}^{-1}(\ell)$  are given by:

$$\mathbf{R}_{\mathbf{h}}^{-1}(\ell) = \frac{\mathbf{I}_L - \mathbf{k}_{\mathbf{h}}(\ell) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}^H(\ell - \Delta)}{\alpha} \mathbf{R}_{\mathbf{h}}^{-1}(\ell - 1) \quad (23)$$

$$\mathbf{R}_{\mathbf{g}}^{-1}(\ell) = \frac{\mathbf{I}_M - \mathbf{k}_{\mathbf{g}}(\ell) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}^H(\ell - \Delta)}{\alpha} \mathbf{R}_{\mathbf{g}}^{-1}(\ell - 1) \quad (24)$$

$$\mathbf{R}_{\sigma}^{-1}(\ell) = \mathbf{R}_{\mathbf{g}}^{-1}(\ell) - \mathbf{R}_{\mathbf{g}}^{-1}(\ell) (\epsilon^{-1} \mathbf{I}_M + \mathbf{R}_{\mathbf{g}}^{-1}(\ell))^{-1} \mathbf{R}_{\mathbf{g}}^{-1}(\ell) \quad (25)$$

where

$$\mathbf{k}_{\mathbf{h}}(\ell) = \frac{\mathbf{R}_{\mathbf{h}}^{-1}(\ell - 1) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell - \Delta)}{\alpha \lambda(\ell) + \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}^H(\ell - \Delta) \mathbf{R}_{\mathbf{h}}^{-1}(\ell - 1) \bar{\mathbf{y}}_{\mathbf{h}(\ell-1)}(\ell - \Delta)} \quad (26)$$

$$\mathbf{k}_{\mathbf{g}}(\ell) = \frac{\mathbf{R}_{\mathbf{g}}^{-1}(\ell - 1) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell - \Delta)}{\alpha + \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}^H(\ell - \Delta) \mathbf{R}_{\mathbf{g}}^{-1}(\ell - 1) \bar{\mathbf{y}}_{\mathbf{g}(\ell-1)}(\ell - \Delta)} \quad (27)$$

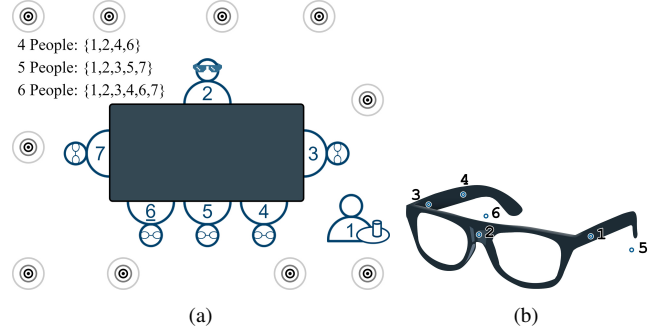
are the Kalman gains. Therefore, the temporal filter can be updated using the derived Kalman gain:

$$\mathbf{g}(\ell) = \mathbf{g}(\ell - 1) + \mathbf{k}_{\mathbf{g}}(\ell) \hat{X}^*(\ell). \quad (28)$$

As shown in [22], the computational complexity is significantly reduced due to combining two sub-optimal solutions using the KP.

## V. EXPERIMENTAL RESULTS

We conducted experiments using the SPEAR challenge first dataset to present the tradeoffs involved in real-time source separation, dereverberation, and noise suppression. This dataset contains real six-channel recordings captured in dynamic, noisy, and reverberant environments using a head-worn microphone array as shown in Fig. 1(b). The reverberant room has dimensions of  $6.11 \times 7.74 \times 3.44$  meters and a reverberation time of 645 ms. The environment



**Fig. 1.** (a) The SPEAR Challenge scene setup and (b) the microphone array worn by person number 2 [23], [24].

presented a challenging scenario, including speech from both the microphone wearer and distant speakers. Moreover, the microphone array wearer's head is in motion during the conversation, further increasing the complexity of the scenario. In addition, 10 loudspeakers were placed at various heights throughout the room, playing uncorrelated restaurant-like background sounds. To capture the spatial characteristics of the microphone array, the dataset included 1020 sphere-sampled points with impulse responses (IRs) for each direction, measured using a mannequin in an anechoic room. Using the Haversine formula [27], we determined the nearest neighbor impulse response and then calculated the reference sensor's relative transfer function (RTF). However, steering errors were unavoidable due to the finite number of samples, the dynamic nature of the environment, and the room's reverberation.

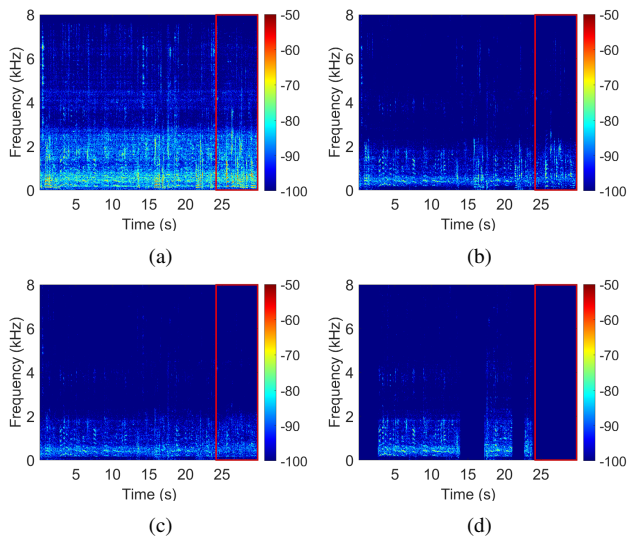
Each audio signal in the dataset is 1-minute long and initially recorded at a sampling rate of 48 kHz, which we downsampled to 16 kHz. The observation signals were then divided into overlapping frames of 1024 samples, with a 75% overlap, and transformed into the STFT domain with a Hamming window. The prediction delay was set to  $\Delta = 2$ , and the prediction filter lengths were set to 14, 12, and 10 for the frequency ranges of 0 – 1 kHz, 1 – 3 kHz, and 3 – 8 kHz, respectively. The prediction filter was initialized to zeros ( $\mathbf{g}(0) = [0 \dots 0]$ ), while the spatial filter length was set to  $M = 6$  and was initialized as a delay-and-sum beamformer ( $\mathbf{h}(0) = \frac{\mathbf{d}(0)}{M}$ ). The covariance matrices were initialized as  $\mathbf{R}_{\mathbf{g}}^{-1}(0) = 10^{-2} \mathbf{I}_M$ ,  $\mathbf{R}_{\mathbf{h}}^{-1}(0) = 10^{-4} \mathbf{I}_L$ . The forgetting factor  $\alpha = 0.994$  was applied to the RLS cost functions, and the regularization factor was set to  $\epsilon^{-1} = 10^2$ . Those parameters were chosen to stabilize the signal and maintain an appropriate tradeoff between desired signal preservation and noise reduction. The algorithm converges within a few hundred frames, reducing noise and interference.

In our experimental setup, the desired speaker is represented as person number 6. In contrast, person number 4 represents the distant undesired speaker, and person number 2 represents the undesired microphone wearer (the scene illustrated in Fig. 1(a)). We compared the performance of the algorithms using the following metrics: perceptual evaluation of speech quality (PESQ) [28], short-time objective intelligibility (STOI) [29], and average maximum correlation (AMC). The algorithm's output was compared against the MVDR-VAD beamformer, which adapts its time-varying noise covariance matrix during frames where the relevant speaker is silent, similar to the approach in [8], where the signal is also nulled during periods of silence from the appropriate speaker. The final results were averaged across seven audio files for evaluation.

The AMC is calculated as the average of the maximum absolute cross-correlation values between the 1024-sample frames of the algorithm's output, denoted as  $\hat{x}_j$ , for active frames of source  $i$ , and the corresponding frames from the MVDR-VAD output, denoted as

**Table I.** Comparison of PESQ, STOI, and AMC metrics between the reference microphone, wMPDR-WPE [22], and the proposed MPDR-WPE. The upward arrow ( $\uparrow$ ) indicates a higher value is desired, while the downward arrow ( $\downarrow$ ) indicates a lower value.

Algorithm	PESQ $\uparrow$	STOI $\uparrow$	AMC <sub>6</sub> $\uparrow$	AMC <sub>4</sub> $\downarrow$	AMC <sub>2</sub> $\downarrow$
reference microphone	1.68 $\pm$ 0.50	0.54 $\pm$ 0.04	0.51 $\pm$ 0.02	0.47 $\pm$ 0.06	0.71 $\pm$ 0.04
wMPDR-WPE [22]	2.47 $\pm$ 0.36	<b>0.75 <math>\pm</math> 0.02</b>	<b>0.79 <math>\pm</math> 0.02</b>	0.42 $\pm$ 0.03	0.45 $\pm$ 0.04
MPDR-WPE	<b>2.63 <math>\pm</math> 0.60</b>	0.67 $\pm$ 0.04	0.72 $\pm$ 0.05	<b>0.34 <math>\pm</math> 0.05</b>	<b>0.38 <math>\pm</math> 0.06</b>



**Fig. 2.** Spectrograms comparison across various algorithms. (a) Reference microphone, (b) wMPDR-WPE [22], (c) proposed MPDR-WPE, and (d) MVDR-VAD. The red boxes indicate the final five seconds, highlighting the attenuation of the undesired speech signal associated with the microphone-array wearer.

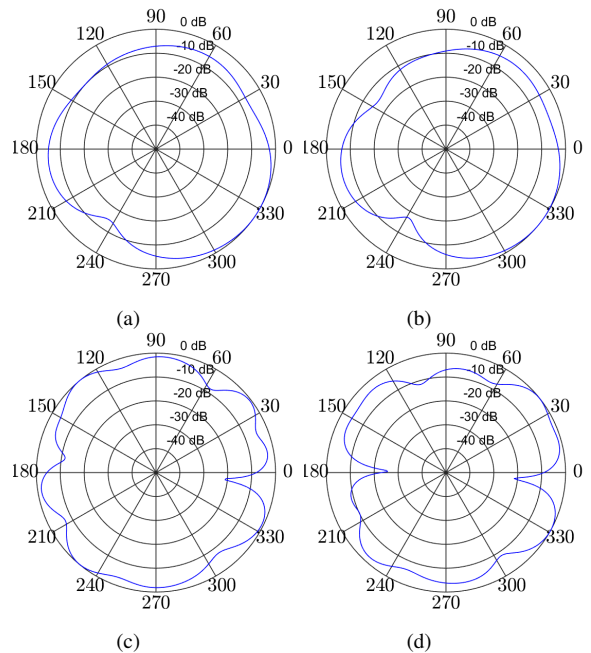
$x_{i,j}$ , when the VAD of person  $i$  is active. This metric quantitatively evaluates the algorithm's source separation performance: lower AMC values indicate more significant source attenuation, while higher AMC values suggest better source preservation. Mathematically, the AMC concerning speaker  $i$  is defined as:

$$\text{AMC}_i = \frac{1}{P} \sum_{j=1}^P \max_{\tau} \left| \frac{\sum_{n=0}^{N-1} \hat{x}_j(n) x_{i,j}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1} \hat{x}_j^2(n) \sum_{n=0}^{N-1} x_{i,j}^2(n)}} \right|, \quad (29)$$

where  $P$  represents the number of non-silent frames of source  $i$ ,  $N$  is the frame length (set to 1024), and  $\tau$  denotes the cross-correlation lag.

As shown in Table I, our proposed algorithm, MPDR-WPE, achieves the lowest AMC values for the undesired speakers (persons 2 and 4), indicating superior source separation performance. At the same time, it maintains high signal quality and low distortion for the desired speaker (person 6). This superior performance is further reflected in Fig. 2, where the MPDR-WPE algorithm demonstrates greater source separation and significant attenuation of the microphone wearer's signal (around 3 dB) compared to wMPDR-WPE. The MPDR-WPE algorithm's integration of WPE with the spatial filter notably enhances dereverberation performance compared to MVDR-VAD, which focuses mainly on source separation.

The azimuth beampatterns shown in Fig. 3 demonstrate the directional sensitivity of the wMPDR-WPE and MPDR-WPE algorithms at 1 kHz and 4 kHz frequencies. The desired speaker is positioned at an azimuth of 327 degrees, the distant undesired speaker at 12



**Fig. 3.** Azimuth beampatterns for (a) wMPDR-WPE at 1 kHz, (b) MPDR-WPE at 1 kHz, (c) wMPDR-WPE at 4 kHz, and (d) MPDR-WPE at 4 kHz, all at fixed elevation.

degrees, and the microphone wearer, also an undesired source, at 0 degrees. These beampatterns confirm that MPDR-WPE is more effective at suppressing directional interferences, providing more focused attenuation of unwanted signals. In contrast, wMPDR-WPE shows a relatively uniform spatial response across different directions, indicating less effective suppression of directional interferences. Notably, both algorithms concentrate their spatial filtering at 4 kHz on nulling the microphone wearer (azimuth 0), with MPDR-WPE demonstrating superior performance.

These experiments emphasize two main tradeoffs that occur in real-time source separation for dynamic and real-world recordings. The first tradeoff is between source separation and background noise reduction. Concentrating on reducing directional interference may have a negative effect on reducing diffuse noise. The second tradeoff is between source separation and the preservation of the desired source, which can be adjusted using regularization to compensate for MPDR steering error sensitivity.

## VI. CONCLUSIONS

We have developed an adaptive algorithm for separating and reducing echoes in real-time recordings from multiple channels. The algorithm uses a combination of an RLS-based MPDR beamformer and a temporal weighted prediction filter to effectively balance between reducing interfering background noises, minimizing distortion in the desired signal, and suppressing directional interferences. Using the KP operator, our approach reduces computational complexity, making it suitable for real-time applications. Experimental results show that our approach can effectively separate sources in challenging real-life acoustic environments with multiple noise sources and significant reverberation. This suggests that the algorithm can be used in augmented reality, conference calls, and hearing aids. In the future, we aim to improve the algorithm's adaptive time for better responsiveness in dynamic environments and explore ways to enhance its robustness against inaccuracies in the steering vector.

## VII. REFERENCES

- [1] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1320–1335, 2014.
- [2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [3] H. W. Löllmann, A. Brendel, and W. Kellermann, "Generalized coherence-based signal enhancement," in *Proc. 45th IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 4–8, 2020, pp. 201–205.
- [4] R. Ikeshita, K. Kinoshita, N. Kamo, and T. Nakatani, "Online speech dereverberation using mixture of multichannel linear prediction models," *IEEE Signal Process. Lett.*, vol. 28, pp. 1580–1584, 2021.
- [5] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. 46th IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, June 6–11, 2021, pp. 6129–6133.
- [6] S. Greengard, *Virtual Reality*. Mit Press, 2019.
- [7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2013.
- [8] A. H. Moore, S. Hafezi, R. R. Vos, P. A. Naylor, and M. Brookes, "A compact noise covariance matrix model for MVDR beamforming," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 30, pp. 2049–2061, 2022.
- [9] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. 26th IEEE Convention of Electrical and Electronics Engineers in Israel*, November 17–20, 2010, pp. 000416–000420.
- [10] S. Braun and E. A. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating kalman filters," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1119–1129, 2018.
- [11] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, 2009.
- [12] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. 33rd IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, March 30 – April 4, 2008, pp. 85–88.
- [13] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 544–558, 2018.
- [14] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined weighted prediction error and minimum variance distortionless response for dereverberation," in *Proc. 42nd IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, March 5–9, 2017, pp. 446–450.
- [15] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. 45th IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 4–8, 2020, pp. 216–220.
- [16] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, 2019.
- [17] C. F. Van Loan, "The ubiquitous Kronecker product," *J. Comput. Appl. Math.*, vol. 123, no. 1–2, pp. 85–100, 2000.
- [18] J. Benesty, I. Cohen, and J. Chen, *Array Processing — Kronecker Product Beamforming*. Berlin, Germany: Springer-Verlag, 2019.
- [19] I. Cohen, J. Benesty, and J. Chen, "Differential Kronecker product beamforming," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 892–902, 2019.
- [20] X. Wang, G. Huang, I. Cohen, J. Benesty, and J. Chen, "Kronecker product adaptive beamforming for microphone arrays," in *Proc. 13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, December 14–17, 2021, pp. 49–54.
- [21] G. Huang, J. Benesty, I. Cohen, and J. Chen, "Kronecker product multichannel linear filtering for adaptive weighted prediction error-based speech dereverberation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 30, pp. 1277–1289, 2022.
- [22] W. Yang, G. Huang, A. Brendel, J. Chen, J. Benesty, W. Kellermann, and I. Cohen, "A bilinear framework for adaptive speech dereverberation combining beamforming and linear prediction," in *Proc. 17th International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 5–8, 2022, pp. 1–5.
- [23] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An introduction to the speech enhancement for augmented reality (SPEAR) challenge," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 5–8, 2022, pp. 1–5.
- [24] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv:2107.04174*, 2021.
- [25] D. A. Harville, "Matrix algebra from a statistician's perspective," 1998.
- [26] S. Haykin, "Adaptive filter theory," 4th ed. Upper Saddle River, New Jersey: Prentice Hall, vol. 2, pp. 453–455, 2002.
- [27] C. C. Robusto, "The cosine-haversine formula," *Amer. Math. Monthly*, vol. 64, no. 1, pp. 38–40, 1957.
- [28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. 26th IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2, May 7–11, 2001, pp. 749–752.
- [29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.