2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Low-Complexity Adaptive Beamformer for Joint Reverberation and Noise Suppression

Fan Zhang*, Chao Pan*, Jingdong Chen*, and Jacob Benesty†
*Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China
†INRS-EMT, University of Quebec, Canada

***Abstract*—Reverberation and background noise can severely affect the quality and intelligibility of recorded speech, potentially impairing speech communication and human-machine interaction systems. The minimum variance distortionless response (MVDR) beamformer is commonly used to jointly suppress reverberation and noise, but it is computationally intensive due to the need for matrix inversion at each time frame and subband. In this paper, we introduce a beamforming method that combines an MVDR beamformer optimized for noise reduction with a fixed maximum directivity-factor beamformer for reverberation suppression. This hybrid beamformer offers a more efficient implementation than the traditional MVDR beamformer and adapts to varying levels of reverberation and noise by adjusting a single weighting factor. Simulation results demonstrate that the proposed beamformer achieves significantly faster processing speeds compared to the optimal MVDR beamformer, with only minor performance degradation in noise reduction and reverberation suppression.**

I. Introduction

In hands-free applications like speech communication or recognition systems, the speech signal captured by a distant microphone can be affected by pervasive background noise and reverberation due to multi-path propagation in the acoustic environment. These factors can significantly degrade speech quality, intelligibility, and the performance of speech communication and human-machine interaction systems.

When a microphone array is available, beamforming techniques [1–4] are commonly used to recover the desired signal from noisy and reverberant multichannel observations. A prominent method is the Wiener beamformer [4, 5], which seeks to minimize the mean-squared error between the desired speech component and the beamformer's output. Reverberation is often modeled as a diffuse sound field [5–11], leading the Wiener beamformer to be expressed as a function of the (time-varying) noise covariance matrix and the (time-varying) reverberation variance. This approach can be decomposed into two main components: 1) a minimum variance distortionless response (MVDR) beamformer, which aims to minimize the sum of the noise and reverberation variances in the output, while keeping the desired signal undistorted; and 2) a single-channel Wiener post-filter [12, 13] that further reduces residual noise and reverberation, albeit with some distortion to the desired signal. One challenge with the MVDR beamformer is that its implementation requires inverting the sum of two full-rank matrices for every time frame and subband, which can impose a significant computational burden, particularly with large arrays containing many microphones.

To tackle the computational challenges of the MVDR beamformer, we introduce a distortionless beamformer that combines a fixed maximum directivity-factor beamformer with an MVDR beamformer specifically optimized for noise reduction. Although the concept of using a weighted sum of different sub-beamformers has been employed in various applications, such as adaptive differential microphone arrays [14–16], robust superdirective beamforming [17, 18], and speech enhancement in underdetermined scenarios [19], it has not been previously applied to joint noise and reverberation suppression with a focus on computational efficiency. Simulation results demonstrate that the proposed combined beamformer provides significantly faster processing speeds, particularly with large microphone arrays, while only marginally sacrificing performance compared to the traditional MVDR beamformer.

II. Signal Model and Problem Formulation

Consider a reverberant environment where a speech signal from a distant source is captured by a uniform linear array (ULA) of M omnidirectional microphones. In the short-time Fourier transform (STFT) domain, the observed signal vector of length M can be represented as

$$\mathbf{y}(\ell) = \mathbf{d}_{\theta_S} S(\ell) + \mathbf{r}(\ell) + \mathbf{v}(\ell), \tag{1}$$

where ℓ is the time-frame index, $S(\ell)$ is the direct-path speech signal at the reference microphone, $\mathbf{d}_{\theta_S}$ is the signal propagation vector of the speech source, analogous to a typical steering vector, $\mathbf{r}(\ell)$ is the reverberation vector, and $\mathbf{v}(\ell)$ is the additive noise vector. Note that the STFT subband index is omitted in this model for simplicity, as the subsequent processing occurs on a subband basis, with each subband processed independently using the same algorithm. In the field of speech enhancement, it is common to assume that $S(\ell)$, $\mathbf{r}(\ell)$, and $\mathbf{v}(\ell)$ have zero mean, and are mtually uncorrelated [5]. Under these assumptions, the covariance matrix of $\mathbf{y}(\ell)$ can be expressed as

$$\mathbf{\Phi}_{\mathbf{y}}(\ell) = \phi_S(\ell)\mathbf{d}_{\theta_S}\mathbf{d}_{\theta_S}^H + \mathbf{\Phi}_{\mathbf{r}}(\ell) + \mathbf{\Phi}_{\mathbf{v}}(\ell), \tag{2}$$

where $\phi_S(\ell)$ is the variance of $S(\ell)$, the superscript ^{H} denotes the conjugate transpose, and $\mathbf{\Phi}_{\mathbf{r}}(\ell)$ and $\mathbf{\Phi}_{\mathbf{v}}(\ell)$ are the covariance matrices of $\mathbf{r}(\ell)$ and $\mathbf{v}(\ell)$, respectively. The reverberation is usually modeled as a spherically isotropic (diffuse) noise [20]. Hence,

$$\mathbf{\Phi}_{\mathbf{r}}(\ell) = \phi_R(\ell)\mathbf{\Gamma}_{\mathrm{d}}, \tag{3}$$

where $\phi_R(\ell)$ is the common variance of the elements in $\mathbf{r}(\ell)$, and $\mathbf{\Gamma}_{\mathrm{d}}$ is the coherence matrix for the diffuse field, which can be determined based on the array geometry information.

With the above signal model, beamforming with a filter $\mathbf{h}(\ell)$ is typically employed to estimate the desired signal $S(\ell)$ from the multichannel observations $\mathbf{y}(\ell)$, which can be written as

$$\begin{aligned} Z(\ell) &= \mathbf{h}^H(\ell)\mathbf{y}(\ell) \\ &= S_{\mathrm{o}}(\ell) + R_{\mathrm{o}}(\ell) + V_{\mathrm{o}}(\ell), \end{aligned} \tag{4}$$

where $Z(\ell)$ is the beamformer output, and

$$S_{\mathrm{o}}(\ell) \triangleq \mathbf{h}^H(t)\mathbf{d}_{\theta_S}S(\ell), \tag{5}$$

$$R_{\mathrm{o}}(\ell) \triangleq \mathbf{h}^H(\ell)\mathbf{r}(\ell), \tag{6}$$

$$V_{\mathrm{o}}(\ell) \triangleq \mathbf{h}^H(\ell)\mathbf{v}(\ell) \tag{7}$$

denote the components in the beamformer's output that correspond to the direct-path signal, reverberation, and noise, respectively.

From (4), we can define the narrowband output signal-to-reverberation ratio (SRR), output signal-to-noise ratio (SNR), and output signal-to-reverberation-plus-noise ratio (SRNR) as follows:

$$\mathrm{SRR}_{\mathrm{o}}\left[\mathbf{h}(\ell)\right] \triangleq \frac{\phi_{S_{\mathrm{o}}}(\ell)}{\phi_{R_{\mathrm{o}}}(\ell)}, \tag{8}$$

$$\mathrm{SNR}_{\mathrm{o}}\left[\mathbf{h}(\ell)\right] \triangleq \frac{\phi_{S_{\mathrm{o}}}(\ell)}{\phi_{V_{\mathrm{o}}}(\ell)}, \tag{9}$$

$$\mathrm{SRNR}_{\mathrm{o}}\left[\mathbf{h}(\ell)\right] \triangleq \frac{\phi_{S_{\mathrm{o}}}(\ell)}{\phi_{R_{\mathrm{o}}}(\ell) + \phi_{V_{\mathrm{o}}}(\ell)}, \tag{10}$$

where

$$\phi_{S_{\mathrm{o}}}(\ell) = \phi_S(\ell)\left|\mathbf{h}^H(\ell)\mathbf{d}_{\theta_S}\right|^2, \tag{11}$$

$$\phi_{R_{\mathrm{o}}}(\ell) = \phi_R(\ell)\mathbf{h}^H(\ell)\mathbf{\Gamma}_{\mathrm{d}}\mathbf{h}(\ell), \tag{12}$$

$$\phi_{V_{\mathrm{o}}}(\ell) = \mathbf{h}^H(\ell)\mathbf{\Phi}_{\mathbf{v}}(\ell)\mathbf{h}(\ell) \tag{13}$$

are the variances of $S_{\mathrm{o}}(\ell)$, $R_{\mathrm{o}}(\ell)$, and $V_{\mathrm{o}}(\ell)$, respectively.

The MVDR beamformer for joint speech dereverberation and noise reduction is given by [5]

$$\mathbf{h}_{\mathrm{MVDR}}(\ell) = \frac{\left[\phi_R(\ell)\mathbf{\Gamma}_{\mathrm{d}} + \mathbf{\Phi}_{\mathbf{v}}(\ell)\right]^{-1}\mathbf{d}_{\theta_S}}{\mathbf{d}_{\theta_S}^H\left[\phi_R(\ell)\mathbf{\Gamma}_{\mathrm{d}} + \mathbf{\Phi}_{\mathbf{v}}(\ell)\right]^{-1}\mathbf{d}_{\theta_S}}. \tag{14}$$

This beamformer is optimal in terms of maximizing the narrowband output SRNR defined in (10). However, its practical implementation involves inverting the matrix within the brackets for each time frame and subband, which has a computational complexity of $\mathcal{O}(M^3)$. This can impose a significant computational burden for real-time applications, particularly because no fast algorithms are currently available.

In the following section, we will present a combined beamformer that reduces computational complexity while effectively suppressing both reverberation and noise.

III. Combined Beamforming

In this section, we analyze the behavior of the MVDR beamformer in (14) under two extreme scenarios to identify two beamformers that individually maximize the narrowband output SNR and output SRR. By linearly combining these two beamformers, we create a new beamformer that depends on a real-valued weighting factor. We then discuss how to determine the optimal value for this weighting factor by minimizing the variance of the beamformer output.

A. Two Sub-Beamformers

We define the narrowband input reverberation-to-noise ratio (RNR) at frame ℓ as

$$\mathrm{RNR}(\ell) \triangleq \frac{\phi_R(\ell)}{\phi_V(\ell)}, \tag{15}$$

where $\phi_V(\ell)$ is the noise variance at frame ℓ. We now consider two extreme scenarios.

- When $\mathrm{RNR}(\ell) = 0$, meaning there is no reverberation, the MVDR beamformer in (14) degenerates to [21, 22]

$$\mathbf{h}_{\mathrm{MVDR,NR}}(\ell) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(\ell)\mathbf{d}_{\theta_S}}{\mathbf{d}_{\theta_S}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}(\ell)\mathbf{d}_{\theta_S}}, \tag{16}$$

  which is the MVDR beamformer designed for noise reduction only. This beamformer maximizes the output SNR defined in (9).
- When $\mathrm{RNR}(\ell)$ approaches infinity, meaning there is no additive noise, the MVDR beamformer in (14) reduces to [21, 23]

$$\mathbf{h}_{\mathrm{MDF}} = \frac{\mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{d}_{\theta_S}}{\mathbf{d}_{\theta_S}^H\mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{d}_{\theta_S}}, \tag{17}$$

  which is the maximum-directivity-factor (MDF) beamformer. This beamformer maximizes the output SRR defined in (8).

The two beamformers in (16) and (17) will serve as sub-beamformers in constructing the combined beamformer.

B. Combined Beamformer

We construct a new beamformer by linearly combining the two beamformers in (16) and (17):

$$\mathbf{h}_{\alpha}(\ell) = \alpha(\ell)\mathbf{h}_{\mathrm{MVDR,NR}}(\ell) + \left[1 - \alpha(\ell)\right]\mathbf{h}_{\mathrm{MDF}}, \tag{18}$$

where $0 \leq \alpha(\ell) \leq 1$ is a real weighting factor used to balance between noise reduction and reverberation suppression. It is straightforward to verify that when $\alpha(\ell) = 0$, we obtain the MDF beamformer, while setting $\alpha(\ell) = 1$ yields the noise reduction MVDR beamformer in (16).

The MDF beamformer is a fixed beamformer with time-invariant filter coefficients. In contrast, the noise reduction MVDR beamformer involves computing the inverse of the noise covariance matrix. This inversion can be efficiently performed using the matrix inversion lemma [2], which has a complexity of $\mathcal{O}(M^2)$. Consequently, from a computational complexity standpoint, the combined beamformer $\mathbf{h}_{\alpha}(\ell)$ presents a valuable alternative to the MVDR beamformer in (14).

C. Determination of the Optimal Value of $\alpha(\ell)$

The goal of this section is to determine the optimal value of $\alpha(\ell)$ for the combined beamformer to maximize the output SRNR, as defined in (10). Given that the combined beamformer is distortionless, this optimization problem can be equivalently framed as minimizing the variance of its output.

The output of the combined beamformer $\mathbf{h}_\alpha(\ell)$ can be expressed as

$$\begin{aligned} Z(\ell) &= \alpha(\ell) Z_{\mathrm{NR}}(\ell) + [1-\alpha(\ell)]\, Z_{\mathrm{DR}}(\ell) \\ &= Z_{\mathrm{DR}}(\ell) - \alpha(\ell) U(\ell), \end{aligned} \tag{19}$$

where

$$Z_{\mathrm{NR}}(\ell) = \mathbf{h}_{\mathrm{MVDR,NR}}^H(\ell)\mathbf{y}(\ell), \tag{20}$$

$$Z_{\mathrm{DR}}(\ell) = \mathbf{h}_{\mathrm{MDF}}^H\mathbf{y}(\ell) \tag{21}$$

are the outputs of the noise reduction MVDR beamformer in (16) and the MDF beamformer, respectively, and

$$\begin{aligned} U(\ell) &= Z_{\mathrm{DR}}(\ell) - Z_{\mathrm{NR}}(\ell) \\ &= \left[\mathbf{h}_{\mathrm{MDF}} - \mathbf{h}_{\mathrm{MVDR,NR}}(\ell)\right]^H \left[\mathbf{r}(\ell) + \mathbf{v}(\ell)\right], \end{aligned} \tag{22}$$

The term $U(\ell)$ does not include any component of the desired signal, and can thus be considered a reference signal used for noise cancellation in $Z_{\mathrm{DR}}(\ell)$ [24]. Leveraging the fact that $\alpha(\ell)$ is real-valued, the variance of $Z(\ell)$ in (19) can be expressed as

$$\phi_Z(\ell) = \alpha^2(\ell)\phi_U(\ell) - 2\alpha(\ell)\Re\left[\phi_{U,Z_{\mathrm{DR}}}(\ell)\right] + \phi_{Z_{\mathrm{DR}}}(\ell), \tag{23}$$

where $\phi_U(\ell)$ is the variance of $U(\ell)$, $\phi_{U,Z_{\mathrm{DR}}}(\ell)$ is the covariance between $U(\ell)$ and $Z_{\mathrm{DR}}(\ell)$, $\Re[\cdot]$ is the real part operator, and $\phi_{Z_{\mathrm{DR}}}(\ell)$ is the variance of $Z_{\mathrm{DR}}(\ell)$.

The variance $\phi_Z(\ell)$ in (23) is a quadratic function of $\alpha(\ell)$. Given the constraint that $\alpha(\ell)$ must lie between 0 and 1, the optimal value of $\alpha(\ell)$ can be determined by

$$\alpha_{\mathrm{c}}(\ell) = \begin{cases} 0 & \alpha_{\mathrm{uc}}(\ell) < 0 \\ 1 & \alpha_{\mathrm{uc}}(\ell) > 1 \\ \alpha_{\mathrm{uc}}(\ell) & \text{otherwise} \end{cases}, \tag{24}$$

where

$$\alpha_{\mathrm{uc}}(\ell) = \frac{\Re\left[\phi_{U,Z_{\mathrm{DR}}}(\ell)\right]}{\phi_U(\ell)}, \tag{25}$$

and the subscripts $_{\mathrm{c}}$ and $_{\mathrm{uc}}$ denote "constrained" and "unconstrained," respectively. In practice, the quantities $\phi_U(\ell)$ and $\phi_{U,Z_{\mathrm{DR}}}(\ell)$ are typically unknown and can be estimated recursively as follows:

$$\widehat{\phi}_U(\ell) = \lambda_r\widehat{\phi}_U(\ell-1) + (1-\lambda_r)\left|U(\ell)\right|^2, \tag{26}$$

$$\widehat{\phi}_{U,Z_{\mathrm{DR}}}(\ell) = \lambda_r\widehat{\phi}_{U,Z_{\mathrm{DR}}}(\ell-1) + (1-\lambda_r)U(\ell)Z_{\mathrm{DR}}^*(\ell), \tag{27}$$

where $0 < \lambda_r < 1$ is a smoothing factor, and the superscript * denotes complex conjugate.

IV. Simulations

In this section, we assess the performance of the combined beamformer in simulated reverberant and noisy acoustic environments.

A. Setup

We consider a room with dimensions of $5 \times 6 \times 4$ meters, where the reflection coefficients are uniformly set at 0.85. This configuration results in a reverberation time (T_{60}) of approximately 475 ms. A uniform linear array (ULA) with an inter-element spacing of 2 cm is placed in the room at a height of 1.4 m. A speech source is located directly in front of the ULA, at a distance of 1.5 m from the array's center. Additionally, an interference source is positioned at an angle of 180° relative to the ULA. Both sources are at the same height as the array. The impulse responses between each source and the sensors are simulated using the image-model method [25, 26].

The clean speech signals are obtained from the NOIZEUS database [27, 28], which are downsampled from their original higher sampling rate to 16 kHz. Given that the database contains only short samples, we compile 30 clean speech recordings from the same speaker and concatenate them into six longer segments by randomly combining five recordings each. The reverberant speech signals are created by convolving these source signals with the room impulse responses. Similarly, the interference signal is generated by convolving an air-conditioner noise recording, captured in an office environment, with the room impulse responses from the interference source position to the microphone sensors' positions. These reverberant speech and interference signals are then mixed, and white Gaussian noise is added to produce the final observation signals. The power ratio between the interference signal and the white Gaussian noise is set to 10 dB.

For evaluation, the impulse response is split into two components: the direct path and reflections. We derive the direct-path signal and reverberation by convolving the clean speech signal with these respective impulse response segments. This allows us to calculate broadband input SNR, SRR, and SRNR. The improvements in SNR, SRR, and SRNR, denoted as $\Delta\mathrm{SNR}$, $\Delta\mathrm{SRR}$, and $\Delta\mathrm{SRNR}$, are used to measure the extent of noise and reverberation attenuation.

To apply beamforming in the STFT domain, we first divide the observation signals into overlapping frames, each consisting of 256 points with a 75% overlap. A Hamming window of length 256 is then applied to each frame, and the resulting windowed frames are transformed into the STFT domain using the fast Fourier transform (FFT). Our proposed beamformer combines both the MDF beamformer and the noise reduction MVDR beamformer. To enhance robustness, we use a regularized MDF beamformer with a regularization parameter set to 10^{-4} [21]. For the noise reduction MVDR beamformer, we need to estimate the noise covariance matrix. We utilize an oracle estimator that assumes noise samples are available, providing an upper bound on performance. The inverse of the estimated noise covariance matrix is calculated using the matrix inversion lemma [2]:

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(\ell) = \frac{1}{\lambda_v}\left[\widehat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(\ell-1) - \frac{\widetilde{\mathbf{v}}(\ell)\widetilde{\mathbf{v}}^H(\ell)}{\mathbf{v}^H(\ell)\widetilde{\mathbf{v}}(\ell) + \widetilde{\lambda}_v}\right], \tag{28}$$

where $\widetilde{\mathbf{v}}(\ell) \triangleq \widehat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(\ell-1)\mathbf{v}(\ell)$ and $\widetilde{\lambda}_v \triangleq \lambda_v/(1-\lambda_v)$. In

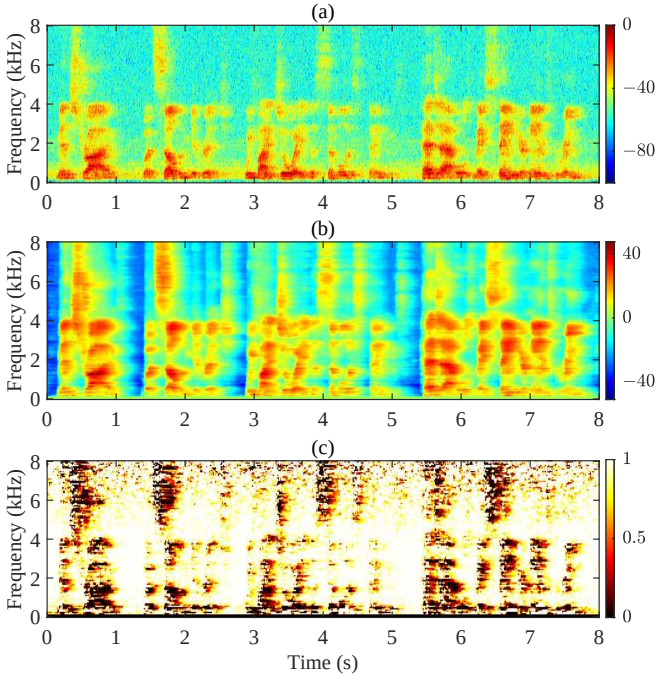

Fig. 1. Example of: (a) the spectrogram of the observed signal at sensor 1, (b) the narrowband input RNR, and (c) the estimate of α_c. The input SNR is 15 dB and $M = 6$.

practice, the noise covariance matrix is typically updated when speech is absent, which requires a voice activity detector or a more advanced speech presence probability estimator [29–33]. In our simulations, we empirically set the smoothing factors: λ_r and λ_v to 0.85 and 0.98, respectively.

For comparison purposes, we select the MVDR beamformer described in (14) as our baseline algorithm. Implementing this beamformer requires an estimate of the reverberation variance, denoted as $\phi_R(\ell)$. Several methods for estimating $\phi_R(\ell)$ have been proposed in the literature, such as those detailed in [5] and its references. To achieve optimal performance, we use an oracle estimator, which assumes that reverberation samples are available.

B. Simulation Results

The initial set of simulations is designed to evaluate the performance of the proposed beamformer. Figure 1(c) presents a plot of the estimated parameter α_c. Comparing this with the narrowband input RNR shown in Figure 1(b), we observe that α_c tends to be larger in regions with low RNR and smaller in regions with high RNR. Consequently, the proposed beamformer behaves similarly to the noise reduction MVDR beamformer in situations where noise predominates over reverberation, and approximates the MDF beamformer when reverberation is more significant compared to noise.

Figure 2 shows the performance of various beamformers as a function of input SNR with $M = 6$, based on averaging results from six speech source signals. The noise reduction MVDR beamformer provides the highest SNR gain, while the MDF beamformer achieves the most significant SRR improvement. Althoughthe hybrid beamformer performs slightly worse in

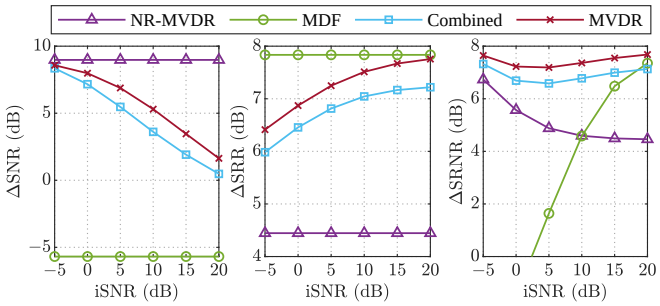

Fig. 2. Speech enhancement performance of various beamforming methods as a function of the input SNR. The number of microphones is $M = 6$.

TABLE I
REAL-TIME FACTORS OF TWO BEAMFORMERS FOR DIFFERENT VALUES OF M. THE INPUT SNR IS 15 DB.

M	2	4	8	16
Proposed	**0.22**	**0.24**	**0.30**	**0.75**
MVDR in (14)	0.34	0.40	0.53	1.39

SRNR improvement compared to the MVDR beamformer in (14), it generally outperforms both of its constituent sub-beamformers, which focus exclusively on either speech dereverberation or noise reduction.

The final set of simulations is conducted to compare the computational complexity of the proposed beamformer with that of the MVDR beamformer in (14). The real-time factor (RTF) [30, 34] is defined as the ratio between the processing time of an algorithm/system and the duration of the input signal, with a smaller RTF indicating a faster algorithm. For real-time applications, it is essential to maintain an RTF value below one. Table I shows the average RTF values for both beamformers across different numbers of microphones. The simulations were performed on a laptop PC (AMD Ryzen 5 4600H CPU at 3.00 GHz, 16.0 GB RAM). The results demonstrate that the proposed beamformer consistently achieves a lower RTF compared to the MVDR beamformer, with the difference becoming more significant as the number of microphones increases.

V. Conclusions

This paper presented a beamformer designed to reduce computational cost while simultaneously suppressing both reverberation and noise. The proposed beamformer is created by linearly combining a noise-reduction MVDR beamformer and a fixed MDF beamformer using an optimal real-valued weighting factor. We derived this optimal weighting factor by minimizing the variance of the beamformer output. By applying the matrix inversion lemma, the proposed beamformer achieves enhanced implementation efficiency. Speech enhancement simulations show that the proposed beamformer outperforms both the noise-reduction MVDR and MDF beamformers, and performs only slightly below the jointly optimized MVDR beamformer.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, 2002.

[3] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 945–958, May 2013.

[4] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 240–251, Feb. 2015.

[5] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1056–1071, June 2018.

[6] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2374–2384, Nov. 2011.

[7] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. EUSIPCO*, pp. 1–5, 2013.

[8] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. EUSIPCO*, pp. 61–65, 2014.

[9] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1006–1018, June 2015.

[10] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. IEEE ICASSP*, pp. 151–155, 2016.

[11] O. Schwartz, S. Gannot, and E. A. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm," in *Proc. EUSIPCO*, pp. 1123–1127, 2016.

[12] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), ch. 3, pp. 39–60, Springer, 2001.

[13] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," in *Proc. IEEE ICASSP*, pp. 380–384, 2016.

[14] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in *Proc. IEEE WASPAA*, pp. 169–172, 1995.

[15] G. W. Elko and J. Meyer, "Second-order differential adaptive microphone array," in *Proc. IEEE ICASSP*, pp. 73–76, 2009.

[16] J. Jin, X. Luo, G. Huang, J. Chen, and J. Benesty, "Beamforming through online convex combination of differential beamformers," in *Proc. IEEE ICASSP*, pp. 8561–8565, 2024.

[17] R. Berkun, I. Cohen, and J. Benesty, "Combined beamformers for robust broadband regularized superdirective beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 877–886, May 2015.

[18] W. Meng, C. Zheng, and X. Li, "Fully automatic balance between directivity factor and white noise gain for large-scale microphone arrays in diffuse noise fields," in *Interspeech*, pp. 5433–5437, 2022.

[19] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3461–3475, 2021.

[20] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 108, no. 1, pp. 204–210, 2000.

[21] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. John Wiley & Sons, 2018.

[22] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 67–79, Jan. 2014.

[23] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 393–398, June 1986.

[24] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propag.*, vol. 30, pp. 27–34, Jan. 1982.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[26] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin, Germany: Springer-Verlag, 2006.

[27] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, 2007.

[28] P. C. Loizou, *Speech Enhancement: Theory and Practice (2nd ed.)*. CRC Press, 2013.

[29] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2159–2169, Sep. 2011.

[30] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE ICASSP*, pp. 5210–5214, 2016.

[31] M. Taseska and E. A. P. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 2223–2236, Nov. 2017.

[32] J. M. Mart n-Doñas, J. Jensen, Z.-H. Tan, A. M. Gomez, and A. M. Peinado, "Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 3080–3094, 2020.

[33] F. Zhang, C. Pan, J. Benesty, and J. Chen, "Directional gain based noise covariance matrix estimation for MVDR beamforming," in *Proc. IEEE ICASSP*, pp. 511–515, 2024.

[34] O. Cheng, W. Abdulla, and Z. Salcic, "Hardware-software codesign of automatic speech recognition system for embedded real-time applications," *IEEE Trans. Ind. Electron.*, vol. 58, pp. 850–859, Mar. 2011.