GLOBALLY OPTIMIZED LEAST-SQUARES POST-FILTERING FOR MICROPHONE ARRAY SPEECH ENHANCEMENT

Yiteng (Arden) Huang, Alejandro Luebs, Jan Skoglund, W. Bastiaan Kleijn

Google Inc., USA
{ardenhuang, aluebs, jks, kleijnb}@google.com

ABSTRACT

Existing post-filtering techniques for microphone array speech enhancement have two common deficiencies. First, they assume that the noise is either white or diffuse and cannot deal with point interferers. Second, they estimate the post-filter coefficients using only two microphones at a time and then perform averaging over all microphone pairs, yielding a suboptimal solution at best. In this paper, we present a novel post-filtering algorithm that alleviates the first limitation by using a more generalized signal model including not only white and diffuse but also point interferers, and overcomes the second deficiency by offering a globally optimized least-squares solution over all microphones. It is shown by simulations that the proposed method outperforms the existing algorithms in many different acoustic scenarios.

Index Terms— Microphone array, post-filter, beamforming, least-squares

1. INTRODUCTION

Microphone arrays are increasingly being recognized as an effective tool to combat noise, interference, and reverberation for speech acquisition in adverse acoustic environments. They have been proven useful in a variety of applications including robust speech recognition, hands-free voice communication and teleconferencing, hearing aids, to name just a few. Beamforming is a traditional microphone array processing technique that provides a form of spatial filtering: receiving signals coming from a specific direction while attenuating signals from other directions. While plausible, spatial filtering is not optimal in the minimum mean square error (MMSE) sense from a signal reconstruction perspective. The optimal solution is the so-called multichannel Wiener filter (MCWF), which can be decomposed into a minimum variance distortionless response (MVDR) beamformer and a single-channel post-filter [1]. The state-of-the-art post-filtering algorithms in [2, 3, 4] have demonstrated that they could significantly improve speech quality after beamforming.

While these methods are in general successful, they have two common limitations or deficiencies. First, they assume the noise is either spatially white (incoherent) or diffuse and cannot deal with point interferers. Second, they estimate post-filter coefficients using two microphones at a time and then perform averaging over all microphone pairs, leading to a suboptimal solution due to this heuristic approach. In this paper, we present a novel post-filtering algorithm that uses a more generalized signal model consisting of not only diffuse and white noise but also point interfering sources. Moreover the new algorithm offers a globally optimized least-squares (LS) solution over all microphones. The performance of the proposed approach is evaluated on real recorded impulse responses for the desired and interfering sources but synthesized diffuse and white noise.

2. SIGNAL MODELS

Suppose that we intend to use a microphone array of M elements to capture the signal $s(t)$ from a desired point sound source in a noisy acoustic environment. The output of the mth microphone in the time domain is written as

$$x_m(t) = g_{s,m} * s(t) + \psi_m(t), \quad m = 1, 2, \cdots, M, \tag{1}$$

where $g_{s,m}$ denotes the impulse response from the desired source to the mth microphone, $*$ denotes linear convolution, and $\psi_m(t)$ is the unwanted additive noise. Without loss of generality, the additive noise commonly consists of three different types of sound components: namely, coherent noise from a point interfering source $v(t)$[1], diffuse noise $u_m(t)$, and white noise $w_m(t)$. Then we have

$$\psi_m(t) \triangleq g_{v,m} * v(t) + u_m(t) + w_m(t), \tag{2}$$

where $g_{v,m}$ is the impulse response from the point noise source to the mth microphone. Presumably the desired signal and these noise components are short-time stationary and mutually uncorrelated.

In the frequency domain, this generalized microphone array signal model (1) is transformed into

$$\begin{aligned} X_m(j\omega) &= G_{s,m}(j\omega)S(j\omega) + \Psi(j\omega) \\ &= G_{s,m}(j\omega)S(j\omega) + G_{v,m}(j\omega)V(j\omega) + \\ &\quad\; U_m(j\omega) + W_m(j\omega), \end{aligned} \tag{3}$$

where $j \triangleq \sqrt{-1}$, ω is the angular frequency, and $X_m(j\omega)$, $G_{s,m}(j\omega)$, $S(j\omega)$, $G_{v,m}(j\omega)$, $V(j\omega)$, $U(j\omega)$, $W(j\omega)$ are the discrete-time Fourier transforms (DTFTs) of $x_m(t)$, $g_{s,m}$, $s(t)$, $g_{v,m}$, $v(t)$, $u(t)$, and $w(t)$, respectively. Let us put (3) in a vector/matrix form as follows

$$\mathbf{x}(j\omega) = S(j\omega)\mathbf{g}_s(j\omega) + V(j\omega)\mathbf{g}_v(j\omega) + \mathbf{u}(j\omega) + \mathbf{w}(j\omega), \tag{4}$$

where

$$\begin{aligned} \mathbf{z}(j\omega) &\triangleq \left[\, Z_1(j\omega)\ Z_2(j\omega)\ \cdots\ Z_M(j\omega) \,\right]^T, \ z \in \{x, u, w\}, \\ \mathbf{g}_z(j\omega) &\triangleq \left[\, G_{z,1}(j\omega)\ G_{z,2}(j\omega)\ \cdots\ G_{z,M}(j\omega) \,\right]^T, \ z \in \{s, v\}, \end{aligned}$$

$(\cdot)^T$ denotes the transpose of a vector or a matrix. The microphone array spatial covariance matrix is then found as

$$\begin{aligned} \mathbf{R}_{xx}(j\omega) &= \sigma_s^2(\omega)\mathbf{P}_{\mathbf{g}_s}(j\omega) + \mathbf{R}_{\psi\psi}(j\omega) \\ &= \sigma_s^2(\omega)\mathbf{P}_{\mathbf{g}_s}(j\omega) + \sigma_v^2(\omega)\mathbf{P}_{\mathbf{g}_v}(j\omega) + \mathbf{R}_{uu}(j\omega) + \mathbf{R}_{ww}(j\omega), \end{aligned} \tag{5}$$

[1] The proposed post filter theoretically can deal with multiple point interfering sources. But for clarity of presentation, only one point interferer is assumed. The generalization is left to the reader.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

where the assumption of mutually uncorrelated signals has been exploited,

$$\begin{aligned}\mathbf{R}_{zz}(j\omega) &\triangleq E\left\{\mathbf{z}(j\omega)\mathbf{z}^H(j\omega)\right\}, \ \ z \in \{x, \psi, u, w\},\\ \mathbf{P}_{\mathbf{g}_z}(j\omega) &\triangleq \mathbf{g}_z(j\omega)\mathbf{g}_z^H(j\omega), \ \ z \in \{s, v\},\\ \sigma_z^2(\omega) &\triangleq E\left\{Z(j\omega)Z^*(j\omega)\right\}, \ \ z \in \{s, v\},\end{aligned}$$

and $E\{\cdot\}$, $(\cdot)^H$, and $(\cdot)^*$ denote the mathematical expectation, the Hermitian transpose of a vector or matrix, and the conjugate of a complex variable, respectively.

A beamformer filters each microphone signal by a finite impulse response (FIR) filter $H_m(j\omega)$ $(m = 1, 2, \cdots, M)$ and sums the results up to produce a single-channel output

$$Y(j\omega) = \sum_{m=1}^{M} H_m^*(j\omega)X_m(j\omega) = \mathbf{h}^H(j\omega)\mathbf{x}(j\omega), \tag{6}$$

where

$$\mathbf{h}(j\omega) \triangleq \left[\begin{array}{cccc} H_1(j\omega) & H_2(j\omega) & \cdots & H_M(j\omega) \end{array}\right]^T.$$

3. MODELING NOISE COVARIANCE MATRICES

In (5), there are three interference/noise-related components that will be modeled as follows:

(1) **Point Interferer**: The covariance matrix $\mathbf{P}_{\mathbf{g}_v}(j\omega)$ due to the point interfering source $v(t)$ has rank 1. In general, when reverberation is present or the source is in the near field of the microphone array, the complex elements of the impulse response vector $\mathbf{g}_v$ may have different magnitudes. But if only the direct path is taken into account or if the point source is in the far field, we have

$$\mathbf{g}_v(j\omega) = \left[\begin{array}{cccc} e^{-j\omega\tau_{v,1}} & e^{-j\omega\tau_{v,2}} & \cdots & e^{-j\omega\tau_{v,M}} \end{array}\right]^T, \tag{7}$$

which incorporates only the interferer's time differences of arrival at the multiple microphones $\tau_{v,m}$ $(m = 1, 2, \cdots, M)$ with respect to a common reference point.

(2) **Diffuse Noise**: A diffuse noise field is considered spherically or cylindrically isotropic, i.e., it is characterized by uncorrelated noise signals of *equal* power propagating in all directions simultaneously. Its covariance matrix is given by (e.g., [5])

$$\mathbf{R}_{uu}(j\omega) = \sigma_u^2(\omega)\mathbf{\Gamma}_{uu}(\omega), \tag{8}$$

where the (p, q)th element of $\mathbf{\Gamma}_{uu}(\omega)$ is

$$[\mathbf{\Gamma}_{uu}(\omega)]_{p,q} = \begin{cases} \text{sinc}\left(\dfrac{\omega \cdot d_{pq}}{c}\right), & \text{Spherically Isotropic} \\ J_0\left(\dfrac{\omega \cdot d_{pq}}{c}\right), & \text{Cylindrically Isotropic} \end{cases} \tag{9}$$

d_{pq} is the distance between the pth and qth microphones, c is the speed of sound, and $J_0(\cdot)$ is the zero-order Bessel function of the first kind.

(3) **White Noise**: The covariance matrix of additive white noise is simply a weighted identity matrix:

$$\mathbf{R}_{ww}(j\omega) = \sigma_w^2(\omega) \cdot \mathbf{I}_{M\times M}. \tag{10}$$

4. MULTICHANNEL WIENER FILTER, MVDR BEAMFORMING, AND POST-FILTERING

When a microphone array is used to capture a desired wideband sound signal (e.g., speech and/or music), the intention is to minimize the distance between $Y(j\omega)$ in (6) and $S(j\omega)$ for all ω's. The MCWF that is optimal in the MMSE sense can be decomposed into an MVDR beamformer followed by a single-channel Wiener filter (SCWF) [1]:

$$\mathbf{h}_{\text{MCWF}}(j\omega) = \underbrace{\frac{\mathbf{R}_{\psi\psi}^{-1}(j\omega)\mathbf{g}_s(j\omega)}{\mathbf{g}_s^H(j\omega)\mathbf{R}_{\psi\psi}^{-1}(j\omega)\mathbf{g}_s(j\omega)}}_{\triangleq\, \mathbf{h}_{\text{MVDR}}(j\omega)} \cdot \underbrace{\frac{\sigma_{s'}^2(\omega)}{\sigma_{s'}^2(\omega) + \sigma_{\psi'}^2(\omega)}}_{\triangleq\, \mathbf{h}_{\text{SCWF}}(\omega)}, \tag{11}$$

where

$$\begin{aligned}\sigma_{s'}^2(\omega) &\triangleq \sigma_s^2(\omega) \cdot \mathbf{h}_{\text{MVDR}}^H(j\omega)\mathbf{P}_{\mathbf{g}_s}(j\omega)\mathbf{h}_{\text{MVDR}}(j\omega),\\ \sigma_{\psi'}^2(\omega) &\triangleq \mathbf{h}_{\text{MVDR}}^H(j\omega)\mathbf{R}_{\psi\psi}(j\omega)\mathbf{h}_{\text{MVDR}}(j\omega)\end{aligned}$$

are the power of the desired signal and noise at the output of the MVDR beamformer, respectively. This decomposition leads to a widely used structure for microphone array speech acquisition: the SCWF is regarded as a post-filter after the MVDR beamformer.

5. POST-FILTER ESTIMATION

In order to implement the front-end MVDR beamformer and the SCWF as a post-processor given in (11), we need to estimate the signal and noise covariance matrices from the calculated covariance matrix of the microphone signals. The multichannel microphone signals are first windowed (by a weighted overlap-add analysis window) in frames and then transformed by an FFT to get $\mathbf{x}(j\omega, i)$, where i is the frame index. The estimate of the microphone signals' covariance matrix is recursively updated by

$$\hat{\mathbf{R}}_{xx}(j\omega, i) = \lambda\hat{\mathbf{R}}_{xx}(j\omega, i-1) + (1-\lambda)\mathbf{x}(j\omega, i)\mathbf{x}^H(j\omega, i), \tag{12}$$

where $0 < \lambda < 1$ is a forgetting factor.

Again let us ignore the reverberation and hence similar to (7) we have

$$\mathbf{g}_s(j\omega) = \left[\begin{array}{cccc} e^{-j\omega\tau_{s,1}} & e^{-j\omega\tau_{s,2}} & \cdots & e^{-j\omega\tau_{s,M}} \end{array}\right]^T, \tag{13}$$

where $\tau_{s,m}$ is the desired signal's time difference of arrival for the mth microphone with respect to the common reference point.

Suppose that both $\tau_{s,m}$ and $\tau_{v,m}$ are known and do not change over time. So, according to (5) and by using (8) and (10), we have at the ith time frame

$$\begin{aligned}\hat{\mathbf{R}}_{xx}(j\omega, i) = \ &\sigma_s^2(\omega, i)\mathbf{P}_{\mathbf{g}_s}(j\omega) + \sigma_v^2(\omega, i)\mathbf{P}_{\mathbf{g}_v}(j\omega) + \\ &\sigma_u^2(\omega, i)\mathbf{\Gamma}_{uu}(\omega) + \sigma_w^2(\omega, i)\mathbf{I}_{M\times M}.\end{aligned} \tag{14}$$

This equality allows to define a criterion based on the Frobenius norm of the difference between the left and the right hand sides of (14). By minimizing such a criterion, an LS estimator for $\left\{\sigma_s^2(\omega, i), \sigma_v^2(\omega, i), \sigma_u^2(\omega, i), \sigma_w^2(\omega, i)\right\}$ can be deduced. Note that the matrices in (14) are Hermitian. We may not want to include redundant information in this formulation.

For an $M \times M$ Hermitian matrix $\mathbf{A} = [a_{pq}]$, we can define two vectors: one consists of its diagonal elements and the other is the off-diagonal half vectorization (odhv) of its lower triangular part

$$\text{diag}\{\mathbf{A}\} \triangleq \left[\begin{array}{cccc} a_{11} & a_{22} & \cdots & a_{MM} \end{array}\right]^T, \tag{15}$$

$$\text{odhv}\{\mathbf{A}\} \triangleq \left[a_{21}\cdots a_{M1} \ a_{32}\cdots a_{M2} \ \cdots \ a_{M(M-1)}\right]^T. \tag{16}$$

For a plurality of (say N) Hermitian matrices of the same size, we define

$$\text{diag}\{\mathbf{A}_1, \cdots, \mathbf{A}_N\} \triangleq \left[\, \text{diag}\{\mathbf{A}_1\} \;\cdots\; \text{diag}\{\mathbf{A}_N\} \,\right], \quad (17)$$

$$\text{odhv}\{\mathbf{A}_1, \cdots, \mathbf{A}_N\} \triangleq \left[\, \text{odhv}\{\mathbf{A}_1\} \;\cdots\; \text{odhv}\{\mathbf{A}_N\} \,\right]. \quad (18)$$

By using these notations, we re-organize (14) to get

$$\hat{\boldsymbol{\phi}}_{xx}(i) = \boldsymbol{\Theta}\boldsymbol{\chi}(i), \quad (19)$$

where we have omitted the parameter $j\omega$ for clarity of presentation, and

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_{xx}(i) &\triangleq \begin{bmatrix} \text{diag}\{\hat{\mathbf{R}}_{xx}(j\omega, i)\} \\ \text{odhv}\{\hat{\mathbf{R}}_{xx}(j\omega, i)\} \end{bmatrix}, \quad \boldsymbol{\Theta} \triangleq \begin{bmatrix} \mathbf{D}(j\omega) \\ \mathbf{C}(j\omega) \end{bmatrix}, \\
\mathbf{D}(j\omega) &\triangleq \text{diag}\left\{\mathbf{P}_{\mathbf{g}_s}(j\omega), \mathbf{P}_{\mathbf{g}_v}(j\omega), \boldsymbol{\Gamma}_{uu}(j\omega), \mathbf{I}_{M\times M}\right\}, \\
\mathbf{C}(j\omega) &\triangleq \text{odhv}\left\{\mathbf{P}_{\mathbf{g}_s}(j\omega), \mathbf{P}_{\mathbf{g}_v}(j\omega), \boldsymbol{\Gamma}_{uu}(j\omega), \mathbf{I}_{M\times M}\right\}, \\
\boldsymbol{\chi}(i) &\triangleq \left[\; \sigma_s^2(\omega, i) \;\; \sigma_v^2(\omega, i) \;\; \sigma_u^2(\omega, i) \;\; \sigma_w^2(\omega, i) \;\right]^T.
\end{aligned}$$

Here we have $M(M+1)/2$ equations and 4 unknowns. If $M \geq 3$, this is an overdetermined problem.

The aforementioned error criterion is written as

$$J \triangleq \left\| \hat{\boldsymbol{\phi}}_{xx}(i) - \boldsymbol{\Theta}\boldsymbol{\chi}(i) \right\|^2. \quad (20)$$

Minimizing this criterion leads to

$$\hat{\boldsymbol{\chi}}_{\text{LS}}(i) = \Re\left\{ \left(\boldsymbol{\Theta}^H \boldsymbol{\Theta}\right)^{-1} \boldsymbol{\Theta}^H \hat{\boldsymbol{\phi}}_{xx}(i) \right\}, \quad (21)$$

$\Re\{\cdot\}$ denotes the real part of a complex number/vector. Presumably the estimation errors in $\hat{\boldsymbol{\phi}}_{xx}(i)$ are IID random variables. So the LS solution given in (21) is optimal in the MMSE sense. Substituting this estimate into (11) leads to what we refer to as an LS post-filter (LSPF).

So far the deduced LS solution has required that $M \geq 3$. This is due to the use of a more generalized acoustic-field model that consists of four types of sound signals. But if a better knowledge about the acoustic field is available such that some types of interfering signals can be ignored (e.g., no point interferer and/or merely white noise), then those columns in (19) that correspond to these ignorable sound sources can be removed and an LSPF can still be developed even with $M = 2$.

Now let us briefly review how existing post-filtering techniques solve this problem and explain why they are not optimal.

(a) **Zelinski's Post-Filter** [2] (ZPF) assumes:

1) no point interferer, i.e., $\sigma_v^2(\omega) = 0$,
2) no diffuse noise, i.e., $\sigma_u^2(\omega) = 0$, but
3) only additive *incoherent* white noise.

So (19) is simplified as follows

$$\begin{bmatrix} \text{diag}\{\hat{\mathbf{R}}_{xx}(i)\} \\ \text{odhv}\{\hat{\mathbf{R}}_{xx}(i)\} \end{bmatrix} = \begin{bmatrix} \text{diag}\{\mathbf{P}_{\mathbf{g}_s}\} & \mathbf{1}_{M\times 1} \\ \text{odhv}\{\mathbf{P}_{\mathbf{g}_s}\} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sigma_s^2(i) \\ \sigma_w^2(i) \end{bmatrix}. \quad (22)$$

Instead of calculating the optimal LS solution for $\sigma_s^2(i)$ using (21), the ZPF used only the bottom odhv-part of (22) to get

$$\hat{\sigma}^2_{s,\text{ZPF}}(i) = \frac{\sum_{p=1}^{M(M-1)/2} \Re\left\{\text{odhv}\{\hat{\mathbf{R}}_{xx}(i)\}\right\}_p}{\sum_{p=1}^{M(M-1)/2} \Re\left\{\text{odhv}\{\mathbf{P}_{\mathbf{g}_s}\}\right\}_p}. \quad (23)$$

Note from (13) that $\Re\left\{\text{odhv}\{\mathbf{P}_{\mathbf{g}_s}\}\right\}_p = 1$. So (23) becomes

$$\hat{\sigma}^2_{s,\text{ZPF}}(i) = \frac{\sum_{p=1}^{M(M-1)/2} \Re\left\{\text{odhv}\{\hat{\mathbf{R}}_{xx}(i)\}\right\}_p}{M(M-1)/2}. \quad (24)$$

If we use the same acoustic model for the LSPF as what the ZPF uses (i.e., only white noise), it can be shown that the ZPF and the LSPF are equivalent when $M = 2$, but fundamentally different when $M \geq 3$.

(b) **McCowan's Post-Filter** [3] (MPF) assumes:

1) no point interferer, i.e., $\sigma_v^2(\omega) = 0$,
2) no additive white noise, i.e., $\sigma_w^2(\omega) = 0$, but
3) only diffuse noise.

Under these assumptions, (19) becomes

$$\begin{bmatrix} \text{diag}\{\hat{\mathbf{R}}_{xx}(i)\} \\ \text{odhv}\{\hat{\mathbf{R}}_{xx}(i)\} \end{bmatrix} = \begin{bmatrix} \text{diag}\{\mathbf{P}_{\mathbf{g}_s}\} & \text{diag}\{\boldsymbol{\Gamma}_{uu}\} \\ \text{odhv}\{\mathbf{P}_{\mathbf{g}_s}\} & \text{odhv}\{\boldsymbol{\Gamma}_{uu}\} \end{bmatrix} \begin{bmatrix} \sigma_s^2(i) \\ \sigma_u^2(i) \end{bmatrix}. \quad (25)$$

Note from (9) that $\text{diag}\{\boldsymbol{\Gamma}_{uu}\} = \mathbf{1}_{M\times 1}$.

Eq. (25) is an overdetermined system. Again, instead of finding a global LS solution by following (21), the MPF takes three equations from (25) that correspond to the pair of the pth and qth microphones to form a subsystem like the following

$$\begin{bmatrix} \hat{\sigma}^2_{x_p x_p} \\ \hat{\sigma}^2_{x_q x_q} \\ \hat{\phi}_{x_p x_q} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & \mathit{\Gamma}_{pq} \end{bmatrix} \begin{bmatrix} \sigma_s^2 \\ \sigma_u^2 \end{bmatrix}, \quad (26)$$

where we have denoted

$$\hat{\phi}_{x_p x_q} \triangleq \Re\left\{\hat{\mathbf{R}}_{xx}\right\}_{p,q}, \quad \mathit{\Gamma}_{pq} \triangleq \Re\left\{\boldsymbol{\Gamma}_{uu}\right\}_{p,q}.$$

The MPF algorithm solves (26) for σ_s^2 in a particular way as follows

$$\{\hat{\sigma}^2_{s,\text{MPF}}\}_{p,q} = \frac{(\hat{\sigma}^2_{x_p x_p} + \hat{\sigma}^2_{x_q x_q})/2 - \hat{\phi}_{x_p x_q}}{1 - \mathit{\Gamma}_{pq}}. \quad (27)$$

Since there are $M(M-1)/2$ different microphone pairs, the final MPF estimate is simply the average of all subsystems' results

$$\hat{\sigma}^2_{s,\text{MPF}} = \frac{\sum_{p=1}^{M-1} \sum_{q=p+1}^{M} \{\hat{\sigma}^2_{s,\text{MPF}}\}_{p,q}}{M(M-1)/2}. \quad (28)$$

The diffuse noise model is arguably more common in practice than the white noise model. The later can be regarded as a special case of the former when $\boldsymbol{\Gamma}_{uu} = \mathbf{I}_{M\times M}$. But the MPF's approach to solving (25) is heuristic and is unfortunately not optimal. Again, it can be shown that if the LSPF also uses a diffuse-noise-only model, it is equivalent to the MPF when $M = 2$, but fundamentally different when $M \geq 3$.

(c) **Leukimmiatis's Post-Filter** [4] follows the algorithm proposed in the MPF to estimate $\sigma_s^2(i)$. Leukimmiatis et al. simply fixed the bug in the Zelinski's and McCowan's post-filters that the denominator of the post-filter in (11) should be $\sigma^2_{s'}(\omega) + \sigma^2_{\psi'}(\omega)$ rather than $\sigma^2_s(\omega) + \sigma^2_\psi(\omega)$.

6. EXPERIMENTAL RESULTS

This section details the speech enhancement experiments performed to validate the proposed LSPF technique. These experiments use the real measured multichannel impulse response database jointly developed by the Institute of Communication Systems and Data Processing (IND), RWTH Aachen University, Germany and the Acoustic Lab at Bar-Ilan University (BIU), Israel [6]. A detailed description of how the database was designed and built can be found in [7]. Presented here is a set of experiments that consider the first 4 microphones of their array whose spacing is 3 cm. The 60 dB reverberation time is 360 ms. The desired source is at the broadside (0°) of the array while the interfering source is at the 45° direction. Both are 2 m from the array. Clean, continuous, 16 kHz/16-bit speech signals are used for these point sound sources. The former is a female speaker and the latter is a male speaker. The voiced parts of the two signals have many overlaps. Accordingly, the impulse responses are resampled at 16 kHz and are truncated to 4096 samples. Spherically isotropic diffuse noise is generated using the method similar to what was presented in [8]. In our simulations, $72 \times 36 = 2592$ point sources distributed on a large sphere are used. All the signals are truncated to 20 s.

In this study, we have defined three *full-band* measures to characterize a sound field (subscript SF): namely, the signal-to-interference ratio (SIR), signal-to-noise ratio (SNR), and diffuse-to-white-noise ratio (DWR), as follows

$$\text{SIR}_\text{SF} \triangleq 10 \cdot \log_{10}\{\sigma_s^2/\sigma_v^2\}, \quad (29)$$

$$\text{SNR}_\text{SF} \triangleq 10 \cdot \log_{10}\{\sigma_s^2/(\sigma_u^2+\sigma_w^2)\}, \quad (30)$$

$$\text{DWR}_\text{SF} \triangleq 10 \cdot \log_{10}\{\sigma_u^2/\sigma_w^2\}, \quad (31)$$

where $\sigma_z^2 \triangleq E\{z^2(t)\}$ and $z \in \{s, v, u, w\}$.

For performance evaluation, we focus on two objective metrics: the signal-to-interference-and-noise ratio (SINR) and the perceptual evaluation speech quality (PESQ) [9]. We compute the SINR's and PESQ's at each microphone and then take their averages as the input SINR and PESQ, respectively. The output SINR and PESQ (denoted by SINR_o and PESQ_o, respectively) are similarly estimated. The difference between the input and output measures (i.e., the delta values) are of particular interest to be examined. To better see the amount of noise reduction and speech distortion at the output, we also calculate the interference and noise reduction (INR) and the desired-speech-only PESQ (dPESQ). For dPESQ's, we pass the processed desired speech and clean speech to the PESQ estimator. The output PESQ indicates the quality of the enhanced signal while the dPESQ value quantifies the amount of speech distortion introduced. The Hu & Loizou's Matlab codes for PESQ [10] are used in this study.

In order to avoid the well-known signal cancellation problem in the MVDR beamformer due to room reverberation, we choose to use the delay-and-sum (D&S) beamformer for front-end processing and compare four different post-filtering algorithms: namely, none, ZPF, MPF, and LSPF. The D&S-only implementation is just used as a benchmark here. For ZPF and MPF, Leukimmiatis's correction has been employed. Tests are conducted under the following three different setups:

1) White Noise ONLY: SIR_SF = 30 dB, SNR_SF = 5 dB, DWR_SF = -30 dB.
2) Diffuse Noise ONLY: SIR_SF = 30 dB, SNR_SF = 10 dB, DWR_SF = 30 dB.
3) Mixed Noise/Interferer: SIR_SF = 0 dB, SNR_SF = 10 dB, DWR_SF = 0 dB.

Table 1: Microphone array speech enhancement results.

Method	**INR** (dB)	**SINR**$_\text{o}$ / $\triangle$**SINR** (dB)	**PESQ**$_\text{o}$ / $\triangle$**PESQ**	**dPESQ**$_\text{o}$ / $\triangle$**dPESQ**
	White Noise Only			
D&S Only	5.978	14.201/ +5.667	1.795/+0.363	2.286/-0.019
D&S+ZPF	11.893	17.827/ +9.293	2.055/+0.623	2.351/+0.046
D&S+MPF	16.924	17.161/ +8.627	2.115/+0.683	2.130/-0.175
D&S+LSPF	13.858	21.460/+12.925	2.180/+0.748	2.299/-0.006
	Diffuse Noise Only			
D&S Only	3.735	16.915/ +3.423	1.852/+0.088	2.286/-0.019
D&S+ZPF	7.467	18.594/ +5.102	1.954/+0.190	2.311/+0.006
D&S+MPF	10.012	16.545/ +3.053	2.122/+0.358	2.427/+0.121
D&S+LSPF	12.236	17.699/ +4.207	2.254/+0.490	2.516/+0.211
	Mixed Noise/Interferer			
D&S Only	0.782	2.398/ +0.435	1.493/+0.122	2.286/-0.019
D&S+ZPF	2.879	2.424/ +0.461	1.563/+0.193	2.314/+0.009
D&S+MPF	9.470	4.211/ +2.248	1.791/+0.420	2.297/-0.008
D&S+LSPF	16.374	9.773/ +7.810	1.940/+0.569	2.336/+0.031

In these tests, the square-root Hamming window and 512-point FFT are used for the STFT analysis. Two neighboring windows have 50% overlapped samples. The weighted overlap-add method is used to reconstruct the processed signal.

The experimental results are summarized in Table 1. Let us first look at the results for the white-noise-only sound field. Since this is the type of sound field addressed by the ZPF method, the ZPF does a reasonably good job in suppressing noise and enhancing speech quality. But the proposed LSPF achieves more noise reduction and offers higher output PESQ although meanwhile it introduces more speech distortion with a slightly lower dPESQ. The MPF produces a deceptively high INR since its SINR gain is lower than that of the ZPF and LSPF. This means that the MPF significantly suppresses not only noise but also speech signals. Besides, its PESQ and dPESQ are all lower than that of the LSPF.

In the second sound field, it is as expected that the D&S beamformer is less effective to deal with diffuse noise and the ZPF's performance degrades too. In this case the MPF's performance is reasonally good while still the LSPF yields evidently best results.

The third sound field is apparently the most challenging case to tackle due to the presence of a time-varying interfering speech source. The LSPF outperforms the other methods in all metrics.

Finally, it is noteworthy that these purely objective performance evaluation results are consistent with subjective perception of the four techniques in informal listening tests carried out with a small number of our colleagues.

7. CONCLUSIONS

In this paper, we have presented a novel LS post-filtering algorithm for microphone array applications. Unlike the existing post-filtering techniques, the proposed method can deal with not only diffuse and white noise but also point interferers. Moreover it is a globally optimal solution that exploits the information collected by a microphone array more efficiently. The advantage of the proposed technique over existing methods was validated and quantified by simulations in various acoustic scenarios.

8. REFERENCES

[1] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 3, pp. 39–60. Springer-Verlag, Berlin, Germany, 2001.

[2] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” in *Proc. IEEE ICASSP*, Apr. 1988, vol. 5, pp. 2578–2581.

[3] I. A. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” *IEEE Trans. Speech Audio Proc.*, vol. 11, pp. 709–716, Nov. 2003.

[4] S. Leukimmiatis and P. Maragos, “Optimum post-filter estimation for noise reduction in multichannel speech processing,” in *Proc. EUSIPCO*, Sept. 2006, pp. 1–5.

[5] G. W. Elko, “Spatial coherence functions for differential microphones in isotropic noise fields,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 4, pp. 61–85. Springer-Verlag, Berlin, Germany, 2001.

[6] Rwth Aachen University (Germany) and Bar-Ilan University (Israel), “Multichannel impulse response database,” `http://http://www.ind.rwth-aachen.de/en/research/tools-downloads/multichannel-impulse-response-database/`, 2014, [Online; latest accessed 16-Sep-2015].

[7] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan les Pins, France, Sept. 2014.

[8] E. A. P. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *J. Acoust. Soc. Am.*, vol. 122, pp. 3464–3470, Dec. 2007.

[9] ITU-T Rec. P. 862, *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T, Geneva, Switzerland, Feb. 2001.

[10] Y. Hu and P. Loizou, “Matlab software for PESQ,” `http://ecs.utdallas.edu/loizou/speech/software.htm`, 2006, [Online; latest accessed 9-Sep-2015].