A linear MMSE filter using delayed remote microphone signals for speech enhancement in hearing aid applications

Sathyapriyan, Vasudha; Pedersen, Michael Syskind; Østergaard, Jan; Brooks, Mike; Naylor, Patrick; Jensen, Jesper

Published in:
2022 International Workshop on Acoustic Signal Enhancement (IWAENC)

DOI (link to publication from Publisher):
10.1109/IWAENC53105.2022.9914711

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

A LINEAR MMSE FILTER USING DELAYED REMOTE MICROPHONE SIGNALS FOR SPEECH ENHANCEMENT IN HEARING AID APPLICATIONS

Vasudha Sathyapriyan[1,2], *Michael S. Pedersen*[2], *Jan Østergaard* [1], *Mike Brookes* [3], *Patrick A. Naylor* [3], *Jesper Jensen* [1,2]

[1]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[2]Demant A/S, Smørum, Denmark
[3]Department of Electrical and Electronic Engineering, Imperial College London, London, UK

ABSTRACT

Existing methods that use remote microphones with hearing aid (HA) noise reduction systems, assume the wireless transmission to be instantaneous. In practice, however, there exists a time difference of arrival (TDOA) between the wirelessly transmitted target signals and the acoustic signal arriving from the target source to the HA device, which degrades their noise reduction performance. As speech is correlated between consecutive time-frames in the short-time Fourier transform (STFT) domain, we propose a linear minimum mean-square error (MMSE) estimator to estimate the desired signal, by combining multiple HA microphone signals with multiple consecutive time-frames of the remote microphone signal. We derive closed form expressions for the resulting filter weights and interpret them in terms of existing multi-channel and multi-frame methods. The simulation results validate the interpretation and show that using a multi-frame method along with a multi-channel method is an advantage, in the presence of unknown, positive TDOA between the microphone signals.

***Index Terms*—** Multi-microphone speech enhancement, multi-frame speech enhancement, wireless acoustic sensor networks.

1. INTRODUCTION

Microphone arrays play a crucial role in noise reduction in hearing aid (HA) systems. In addition to the spectro-temporal information of the incoming acoustic signals, exploiting the spatial characteristics has benefited both noise reduction and spatial scene preservation in HAs [1]. However, the noise reduction performance of HA systems is limited, due to space constraints, by the number of microphones that can be placed discreetly [2].

Wireless communication has enabled the HA systems to connect to remote microphones placed in a wireless acoustic sensor network (WASN) [3]. Using spatially distributed remote microphones, helps to acquire the spatial characteristics of the entire acoustic environment, as opposed to the spatial information around the head of the HA user, obtained from the HA microphones alone [4]. Moreover, clinical studies have shown that using remote microphones located close to the target source significantly enhance the noise reduction and speech intelligibility in hearing assistive devices (HADs) (e.g., [5, 6]). Existing literature, thus focuses on the inclusion of remote microphones with HA microphones using available beam-forming algorithms (e.g., [7, 8, 9]. However, it ignores the time difference of arrival (TDOA) between the target signal acoustically transmitted to the HA microphones and the target signal which is wirelessly transmitted by the remote microphones to the HA system. In other words, it assumes instantaneous wireless transmission of the remote microphone signals to the HA system. This TDOA depends on the transmission delay introduced by the wireless communication protocol used, the position of the target source w.r.t the microphones and/or the distance between the microphones [3]. For example, under optimum conditions, the Bluetooth LE standard [10] can achieve a wireless latency of around 20 ms. In the presence of positive TDOAs, the multi-microphone beam-formers in existing literature, show poor noise reduction performance in a HA system [4, 11].

In this paper, we use the fact that speech is highly correlated across time and frequency in the discrete STFT domain, to overcome this performance loss. We propose a linear MMSE filter, that combines the multi-channel HA signals and multiple frames of a remote microphone signal, to help overcome the effects of the TDOA between the local HA microphone signals and the remote microphone signal. We derive closed form expressions of the proposed MMSE estimator and show that it is a linear combination of the traditional multi-channel Wiener filter (MWF) [2] and the multi-frame Wiener filter (MFWF) filter [12, 13]. Furthermore, we verify the derived expressions through simulations using oracle signal statistics and show that, even in the presence of an unknown, positive TDOA, the proposed method is able to enhance the signal-to-noise ratio (SNR) of the output, over using the MWF [2] on the HA microphone signals.

2. SIGNAL MODEL AND NOTATION

Consider a multi-channel noise reduction system composed of a HA system connected to remote microphones. We consider a monaural HA system, composed of a microphone array with $M \geq 2$ microphones, worn by the HA user, and a single remote microphone placed close to the mouth of the target source (e.g., a microphone clip worn on the chest of the target source). The complex STFT coefficients of the m^{th} HA microphone signal can be represented by

$$Y_m(k,l) = X_m(k,l) + V_m(k,l), \tag{1}$$

where $Y(k,l)$, $X(k,l)$ and $V(k,l)$ are the noisy signal, target speech and the additive noise, respectively. The frequency bin and the time frame index are denoted by k and l, respectively. As we assume independent sub-band processing, we omit the frequency bin index k for clarity.

Assuming an acoustic scene with one target speech source, the noisy measurements of the HA microphones located locally w.r.t. the HA user, can be stacked into a multi-channel vector as

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956369.

$$\begin{aligned}\mathbf{y}_{\mathrm{HA}}(l) &= \mathbf{a}(\theta_S, l)S(l + \left\lfloor \frac{\tau_a}{T} \right\rfloor) + \mathbf{v}_{\mathrm{HA}}(l) \\ &= \mathbf{d}(\theta_S, l)X_r(l) + \mathbf{v}_{\mathrm{HA}}(l) \\ &= \mathbf{x}_{\mathrm{HA}}(l) + \mathbf{v}_{\mathrm{HA}}(l) \in \mathbb{C}^{M\times 1},\end{aligned} \tag{2}$$

where $S(l+\left\lfloor \frac{\tau_a}{T} \right\rfloor)$ is the speech at the target source located along the azimuthal direction θ_S w.r.t the HA user, $\mathbf{a}$ is the acoustic transfer function (ATF) from the target source location to HA microphones, $\mathbf{d}$ is the corresponding relative acoustic transfer function (RATF), X_{r} is the target signal at a pre-selected HA reference microphone, τ_a is the acoustic propagation delay from the target source to X_{r} (in ms) and T is the STFT frame hop (in ms). Furthermore, $\mathbf{x}_{\mathrm{HA}}$ and $\mathbf{v}_{\mathrm{HA}}$ represent respectively, the speech and noise signal vectors at the HA microphones, which we assume to be uncorrelated.

Consider a remote microphone, placed close to the mouth of the target source. The wirelessly transmitted remote microphone signal, received at the HA system at the l^{th} time-frame, can be written as

$$Y_{\mathrm{E}}(l) = S(l - \left\lfloor \frac{\tau}{T} \right\rfloor) + N_{\mathrm{E}}(l - \left\lfloor \frac{\tau}{T} \right\rfloor), \tag{3}$$

where $Y_{\mathrm{E}}(l)$, $N_{\mathrm{E}}(l)$ are the STFT coefficients of the noisy signal and the additive noise in the remote microphone, and τ is the TDOA (in ms) of the remote microphone signal transmitted to the HA system, w.r.t to the acoustic signal X_r at the HA reference microphone. The TDOA between spatially separated microphone signals can be estimated as in [14, 15]. The TDOAs, $\tau \in \mathbb{R}$, however, in this paper we consider only $\tau \geq 0$. With the current and previous $L-1$ frames of the noisy delayed remote microphone signal, the L^{th} order multi-frame remote microphone vector is given by

$$\mathbf{y}_{\mathrm{E}}(l) = \begin{bmatrix} Y_{\mathrm{E}}(l) \\ \vdots \\ Y_{\mathrm{E}}(l-L+1) \end{bmatrix} = \mathbf{s}_{\mathrm{E}}(l) + \mathbf{n}_{\mathrm{E}}(l) \in \mathbb{C}^{L\times 1}, \tag{4}$$

where, $\mathbf{s}_{\mathrm{E}}(l)$ and $\mathbf{n}_{\mathrm{E}}(l)$ are the speech and noise component vectors in the remote microphone signal, defined corresponding to (3). The speech component of the delayed remote microphone signal is generally correlated to the speech component in the HA reference microphone signal, i.e., $\mathbb{E}[S(l - \left\lfloor \frac{\tau}{T} \right\rfloor)X_r^*(l)] \neq 0$, for a finite τ. Using the model in [12], the multi-frame speech component vector $\mathbf{s}_{\mathrm{E}}$ can decomposed into two mutually uncorrelated components as

$$\mathbf{s}_{\mathrm{E}}(l) = \boldsymbol{\rho}_X(l)X_r(l) + \mathbf{x}_i(l) \in \mathbb{C}^{L\times 1}, \tag{5}$$

where

$$\boldsymbol{\rho}_X(l) \triangleq \frac{\mathbb{E}[\mathbf{s}_{\mathrm{E}}(l)X_r^*(l)]}{\mathbb{E}[X_r(l)X_r^*(l)]} \tag{6}$$

is the normalised inter-frame correlation (IFC) coefficient vector and $\mathbf{x}_i(l)$ is uncorrelated with $X_r(l)$ [12]. $\mathbb{E}\{\cdot\}$ is the expectation operator and * is the conjugate operator. Using (5) in (4), we get

$$\mathbf{y}_{\mathrm{E}}(l) = \mathbf{x}_{\mathrm{E}}(l) + \mathbf{v}_{\mathrm{E}}(l) \in \mathbb{C}^{L\times 1}, \tag{7}$$

where, $\mathbf{x}_{\mathrm{E}}(l) \triangleq \boldsymbol{\rho}_X(l)X_r(l)$ and $\mathbf{v}_{\mathrm{E}}(l) \triangleq \mathbf{x}_i(l) + \mathbf{n}_{\mathrm{E}}(l)$ such that $\mathbf{x}_{\mathrm{E}}$ and $\mathbf{v}_{\mathrm{E}}$ are uncorrelated. As we assume each time-frequency (TF) tile to be processed independently, we omit the time index l for clarity, in the rest of the paper.

In this paper, we apply a linear filter of order $N = M + L$, $\mathbf{w}_{\mathrm{EMWF}} = \begin{bmatrix} \mathbf{w}_{\mathrm{HA}}^H & \mathbf{w}_{\mathrm{E}}^H \end{bmatrix}^H$, to the noisy measurements $\mathbf{y} = \begin{bmatrix} \mathbf{y}_{\mathrm{HA}}^H & \mathbf{y}_{\mathrm{E}}^H \end{bmatrix}^H \in \mathbb{C}^{N\times 1}$ to estimate the desired signal, i.e., the target signal at the HA reference microphone,

$$\hat{X}_{\mathrm{r}} = \mathbf{w}_{\mathrm{EMWF}}^H \mathbf{y}, \tag{8}$$

where the superscript ^{H} denotes the conjugate transpose operator.

Let the $N \times N$ cross power spectral density matrices (CPSDMs) of the noisy, speech and noise vectors of the HA and remote microphone vector be defined as

$$\mathbf{C}_{\mathbf{yy}} = \mathbb{E}\left[\mathbf{y}\mathbf{y}^H\right], \mathbf{C}_{\mathbf{xx}} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right], \mathbf{C}_{\mathbf{vv}} = \mathbb{E}\left[\mathbf{v}\mathbf{v}^H\right], \tag{9}$$

respectively. The CPSDMs for the HA microphone vector and the remote microphone vector can be defined similarly for $\{\mathbf{C}_{\mathbf{yy},\mathrm{HA}}, \mathbf{C}_{\mathbf{xx},\mathrm{HA}}, \mathbf{C}_{\mathbf{vv},\mathrm{HA}}\} \in \mathbb{C}^{M\times M}$ and $\{\mathbf{C}_{\mathbf{yy},\mathrm{E}}, \mathbf{C}_{\mathbf{xx},\mathrm{E}}, \mathbf{C}_{\mathbf{vv},\mathrm{E}}\} \in \mathbb{C}^{L\times L}$, respectively. Under the assumptions that, (i) the speech and noise processes are uncorrelated, (ii) the remote microphone is sufficiently far from the HA, that the noise component in the remote microphone signal is uncorrelated with the noise component in the HA microphone signals and using (2) and (7), we obtain

$$\mathbf{C}_{\mathbf{vv}} = \left[\begin{array}{c|c} \mathbb{E}[\mathbf{v}_{\mathrm{HA}}\mathbf{v}_{\mathrm{HA}}^H] & \mathbf{0} \\ \hline \mathbf{0} & \mathbb{E}[\mathbf{v}_{\mathrm{E}}\mathbf{v}_{\mathrm{E}}^H] \end{array}\right], \tag{10}$$

$$\mathbf{C}_{\mathbf{xx}} = \phi_{X_{\mathrm{r}}} \left[\begin{array}{c|c} \mathbf{d}\mathbf{d}^H & \mathbf{d}\boldsymbol{\rho}_X^H \\ \hline \boldsymbol{\rho}_X\mathbf{d}^H & \boldsymbol{\rho}_X\boldsymbol{\rho}_X^H \end{array}\right], \tag{11}$$

$$\mathbf{C}_{\mathbf{yy}} = \mathbf{C}_{\mathbf{xx}} + \mathbf{C}_{\mathbf{vv}}, \tag{12}$$

where, $\phi_{X_r} = \mathbb{E}[X_r X_r^*]$ is the power spectral density (PSD) of the clean signal component at the reference HA microphone.

3. LINEAR FILTERING

We propose a linear MMSE estimator, to estimate the desired signal using the HA microphone signals and the multi-frame remote microphone signal. We first briefly review the independent, linear MMSE estimators: MWF for the HA microphones and MFWF for the remote microphone. We then describe the proposed joint linear MMSE estimator using both local and remote microphones. Lastly, we show that the resulting extended multi-channel Wiener filter (EMWF) estimate can be decomposed into a linear combination of the MWF estimate and the MFWF estimate of the desired signal, i.e., $\hat{X}_{\mathrm{r,EMWF}} = \alpha\hat{X}_{\mathrm{r,MWF}} + \beta\hat{X}_{\mathrm{r,MFWF}}$, where the linear multipliers, $\alpha, \beta \in \mathbb{C}$ can be expressed in closed-form, as a function of the output SNRs of the MWF and MFWF estimates.

3.1. Multi-channel Wiener filter (MWF)

The MWF estimates a linear filter that minimizes the mean-square error (MSE) between the filtered HA microphone signal and the desired signal X_r [1]. The optimization problem is given by

$$\min_{\mathbf{w}} \mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}\mathbf{y}_{\mathrm{HA}} - X_r\right|^2\right] \quad . \tag{13}$$

Taking the speech CPSDM to be rank-1, i.e., $\mathbf{C}_{\mathbf{xx},\mathrm{HA}} = \mathbb{E}[\mathbf{x}_{\mathrm{HA}}\mathbf{x}_{\mathrm{HA}}^H] = \phi_{X_r}\mathbf{d}\mathbf{d}^H$ and using the matrix inversion lemma [16], the MWF solution can be written as, e.g., [17, 18]

$$\mathbf{w}_{\mathrm{MWF}} = \frac{\mathbf{C}_{\mathbf{vv},\mathrm{HA}}^{-1}\mathbf{d}\phi_{X_r}}{1 + \phi_{X_r}\mathbf{d}^H\mathbf{C}_{\mathbf{vv},\mathrm{HA}}^{-1}\mathbf{d}}. \tag{14}$$

Using (14), the SNR of the MWF estimate can be written as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{MWF}} &\triangleq \frac{\mathbf{w}_{\mathrm{MWF}}^{\mathrm{H}}\,\mathbf{C}_{\mathbf{xx},\mathrm{HA}}\,\mathbf{w}_{\mathrm{MWF}}}{\mathbf{w}_{\mathrm{MWF}}^{\mathrm{H}}\,\mathbf{C}_{\mathbf{vv},\mathrm{HA}}\,\mathbf{w}_{\mathrm{MWF}}} \\ &= \phi_{X_r}\mathbf{d}^{\mathrm{H}}\mathbf{C}_{\mathbf{vv},\mathrm{HA}}^{-1}\,\mathbf{d}.\end{aligned} \tag{15}$$

3.2. Multi-frame Wiener filter (MFWF)

In [12, 19], based on the speech correlation between consecutive time-frames, a multi-frame single channel speech enhancement method was proposed. Similarly, we estimate a linear filter that minimizes the MSE between the filtered remote microphone signal vector and the desired signal X_r as

$$\min_{\mathbf{w}} \mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}\mathbf{y}_{\mathrm{E}} - X_r\right|^2\right]. \tag{16}$$

Using (7) and the matrix inversion lemma [16], the optimal linear filter weights are obtained as [12]

$$\mathbf{w}_{\mathrm{MFWF}} = \frac{\mathbf{C}_{\mathbf{vv},\mathrm{E}}^{-1}\boldsymbol{\rho}_X\phi_{X_r}}{1+\phi_{X_r}{\boldsymbol{\rho}_X}^H\mathbf{C}_{\mathbf{vv},\mathrm{E}}^{-1}\boldsymbol{\rho}_X}. \tag{17}$$

Using (17), the SNR of the MFWF estimate can be written as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{MFWF}} &\triangleq \frac{\mathbf{w}_{\mathrm{MFWF}}^{\mathrm{H}}\,\mathbf{C}_{\mathbf{xx},\mathrm{E}}\,\mathbf{w}_{\mathrm{MFWF}}}{\mathbf{w}_{\mathrm{MFWF}}^{\mathrm{H}}\,\mathbf{C}_{\mathbf{vv},\mathrm{E}}\,\mathbf{w}_{\mathrm{MFWF}}} \\ &= \phi_{X_r}\boldsymbol{\rho}_X^{\mathrm{H}}\mathbf{C}_{\mathbf{vv},\mathrm{E}}^{-1}\,\boldsymbol{\rho}_X.\end{aligned} \tag{18}$$

3.3. Extended multi-channel Wiener filter (EMWF)

We propose a linear MMSE estimator to estimate the desired signal X_r, using both HA and remote microphone signal vectors in (2) and (7), respectively. The linear optimization problem is given by

$$\min_{\mathbf{w}} \mathbb{E}\left[\left|\mathbf{w}^{\mathrm{H}}\mathbf{y} - X_r\right|^2\right]. \tag{19}$$

Using (9) it can be shown that, the resulting linear filter is given by

$$\mathbf{w}_{\mathrm{EMWF}} = \mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\mathbf{e}_r = \mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{xx}}\mathbf{e}_r, \tag{20}$$

where $\mathbf{C}_{\mathbf{yx}} \triangleq \mathbb{E}[\mathbf{y}\mathbf{x}^H]$ and $\mathbf{e}_r$ is a microphone selection vector, with 1 at the HA reference microphone index and 0 elsewhere. Inserting (10) and (11) in (12) and applying the matrix inversion lemma [16], it can be shown that, (20) may be expressed as

$$\mathbf{w}_{\mathrm{EMWF}} = \begin{bmatrix} \alpha\mathbf{w}_{\mathrm{MWF}} \\ \beta\mathbf{w}_{\mathrm{MFWF}} \end{bmatrix} \tag{21}$$

with

$$\alpha = \frac{1+\phi_{X_r}\mathbf{d}^H\mathbf{C}_{\mathbf{vv},\mathrm{HA}}^{-1}\mathbf{d}}{1+\phi_{X_r}\mathbf{d}^H\mathbf{C}_{\mathbf{vv},\mathrm{HA}}^{-1}\,\mathbf{d}+\phi_{X_r}\boldsymbol{\rho}_X^H\mathbf{C}_{\mathbf{vv},\mathrm{E}}^{-1}\,\boldsymbol{\rho}_X}, \tag{22}$$

and

$$\beta = \frac{1+\phi_{X_r}\boldsymbol{\rho}_X^H\mathbf{C}_{\mathbf{vv},\mathrm{E}}^{-1}\boldsymbol{\rho}_X}{1+\phi_{X_r}\mathbf{d}^H\,\mathbf{C}_{\mathbf{vv},\mathrm{HA}}^{-1}\,\mathbf{d}+\phi_{X_r}\boldsymbol{\rho}_X^H\,\mathbf{C}_{\mathbf{vv},\mathrm{E}}^{-1}\,\boldsymbol{\rho}_X}. \tag{23}$$

Using (15) and (18), we can now re-write (22) and (23) as

$$\alpha = \frac{1+\mathrm{SNR}_{\mathrm{MWF}}}{1+\mathrm{SNR}_{\mathrm{MWF}}+\mathrm{SNR}_{\mathrm{MFWF}}}, \tag{24}$$

$$\beta = \frac{1+\mathrm{SNR}_{\mathrm{MFWF}}}{1+\mathrm{SNR}_{\mathrm{MWF}}+\mathrm{SNR}_{\mathrm{MFWF}}}. \tag{25}$$

Therefore, from (21), EMWF is a linear combination of MWF applied to the HA microphone signals and MFWF applied to the multi-frame remote microphone vector, i.e., $\hat{X}_{\mathrm{r,\,EMWF}} = \alpha\hat{X}_{\mathrm{r,\,MWF}} + \beta\hat{X}_{\mathrm{r,\,MFWF}}$. Moreover, the influence of the HA and the remote microphones are completely described by $\mathrm{SNR}_{\mathrm{MWF}}$ and $\mathrm{SNR}_{\mathrm{MFWF}}$. This can be realized well in the following two cases, that are verified experimentally in Sec. 4:

1. When $\mathrm{SNR}_{\mathrm{MWF}} \gg \mathrm{SNR}_{\mathrm{MFWF}}$, e.g., the TDOA is high (i.e., $\mathrm{SNR}_{\mathrm{MFWF}} \to 0$) and/or the target source is close to the HA user, (i.e., high $\mathrm{SNR}_{\mathrm{MWF}}$), then, $\alpha \to 1$ and $\beta \to 0$, i.e., the EMWF estimate is dominated by the MWF applied to the HA microphone signals, while the contribution of the remote microphone signal becomes insignificant.
2. When $\mathrm{SNR}_{\mathrm{MWF}} \ll \mathrm{SNR}_{\mathrm{MFWF}}$, e.g., the TDOA is low (i.e., high $\mathrm{SNR}_{\mathrm{MFWF}}$) and/or HA microphone signals are very noisy, (i.e., $\mathrm{SNR}_{\mathrm{MWF}} \to 0$), then, $\alpha \to 0$ and $\beta \to 1$, i.e., the EMWF estimate is dominated by the MFWF applied to the remote microphone signal frames, while the contribution of the HA microphone signals becomes insignificant.

4. RESULTS

In this section, we evaluate the performance of the proposed algorithm through simulation experiments using the knowledge of the clean and noisy signals. We use segmental signal-to-noise ratio (segSNR) [20] and short-term objective intelligibility (STOI) [21] to estimate the noise reduction and the speech intelligibility performance, respectively. The gain in the performance metric is calculated w.r.t the corresponding metric for the unprocessed noisy input signal at the HA reference microphone.

The acoustic scene consists of a monaural HA user, with $M = 2$ microphones and a remote microphone placed close to the mouth of the target source. The listener is in a cylindrically isotropic noise field, while the remote microphone, due to its proximity to the source, is assumed to be nearly noise-free, with only microphone noise. To study the performance of the proposed algorithm for stationary and non-stationary target sources, we use sustained vowel signals and normal speech signals, respectively. The speech signals spoken by 10 subjects, are obtained from the TIMIT corpus [22]. Each signal is 20s long with 2s of initial silence, containing multiple sentences. The sustained phonation of vowels (i.e., /a/, /e/, /i/, /o/, /u/) were recorded by 3 subjects. Each vowel utterance is 10 s long with 2 s of initial silence. All the signals are sampled at a sampling rate of $f_s = 16\mathrm{kHz}$. The hearing aid head related impulse responses (HAHRIRs), measured on human subjects [23], are used to simulate the HA microphone signals. Since the HAHRIRs are measured with loudspeakers distributed uniformly around a ring of radius 1.9 m, centered at, and at the same height as the HA user's head, we consider the target source, and thereby the remote microphone to be placed on the same ring, along the direction $\theta_S = 0^\circ$. The cylindrically isotropic noise field at the HA microphones is simulated by convolving independent realisations of stationary speech shaped noise (SSN) with the HAHRIRs from the 48 directions available in the database. The microphone noise in the remote microphone is taken to be white Gaussian noise (WGN).

For the spectral analysis and synthesis of the signals, we use a square-root Hanning window of 8 ms frame length ($N = 128$ samples) and a frame hop of $T = 2$ ms. For simplicity, we choose the TDOA, $\tau = nT$ ms, where $n \in \mathbb{Z}$, to simulate multiple TDOA conditions. To focus on the validation and interpretation of the proposed algorithm, to keep the estimation errors low, we use the oracle second order statistics estimated using

$$\begin{aligned}\mathbf{C}_{\underline{\mathbf{y}}\underline{\mathbf{y}}}(l) &= \alpha_{\mathrm{yy}}\mathbf{C}_{\underline{\mathbf{y}}\underline{\mathbf{y}}}(l-1) + (1-\alpha_{\mathrm{yy}})\,\underline{\mathbf{y}}(l)\,\underline{\mathbf{y}}^H(l), \\ \mathbf{C}_{\underline{\mathbf{y}}\underline{\mathbf{x}}}(l) &= \alpha_{\mathrm{yx}}\mathbf{C}_{\underline{\mathbf{y}}\,\underline{\mathbf{x}}}(l-1) + (1-\alpha_{\mathrm{yx}})\,\underline{\mathbf{y}}(l)\,\underline{\mathbf{x}}^H(l),\end{aligned} \tag{26}$$

where α_{yy} and α_{yx} are smoothing coefficients. We take $\alpha_{\mathrm{yy}} = \alpha_{\mathrm{yx}}$, for simplicity. To capture the short-term variations of the non-

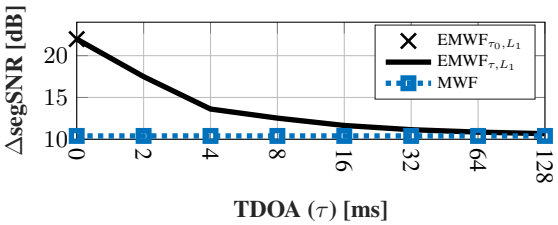

Fig. 1: Comparing the performance of EMWF-$_{\tau,L_1}$ with MWF as a function of TDOA (τ).

stationary speech in noise signal, we choose α_{yy} corresponding to a time constant of 95 ms, for both MWF and EMWF. The experiments are conducted at input SNRs of $\text{SNR}_{\text{in,HA}} = \{-20, -10, 0, 10, 20\}$ dB at the HA reference microphone, and at $\text{SNR}_{\text{in,E}} = 50$ dB at the remote microphone.

In the following discussion, the choice of the filter order (L) used for EMWF and MFWF algorithms, in the presence of varying TDOAs (τ), are denoted as subscripts, e.g., EMWF$_{\tau_0,L_1}$ denotes using $L = 1$ with EMWF, when the TDOA is $\tau = 0$ ms. Existing methods (e.g., [7]), use only the current frame of the remote microphone signal under the ideal assumption of zero TDOA, which is equivalent to the case, EMWF$_{\tau_0,L_1}$. Fig. 1 compares ΔsegSNR achieved by MWF and EMWF$_{\tau,L_1}$, as a function of TDOA (τ), at $\text{SNR}_{\text{in,HA}} = 0$ dB. The ΔsegSNR achieved by EMWF$_{\tau_0,L_1}$ (indicated by $\times$ in Fig. 1), is over 10 dB higher than the gain achieved by MWF. With positive TDOAs, i.e., $\tau > 0$, ΔsegSNR of MWF stays constant due to its independence to the remote microphone signal, while the performance of EMWF$_{\tau,L_1}$ tends to the performance of MWF. In other words, the benefit of using a single-frame remote microphone is maximum when the TDOA is zero, however, when there is a positive TDOA, which is typical in practical wireless transmission of signals, the benefit of using the remote microphone declines.

Fig. 2 shows the performance of the proposed EMWF compared to MWF as a function of the filter order (L), for different positive TDOAs (τ), at $\text{SNR}_{\text{in,HA}} = 0$ dB. We use the performance of EMWF$_{\tau_0,L_1}$ (indicated by $\times$ in Fig. 2), as the benchmark for the performance analysis of EMWF. For the sustained vowel signals, Fig. 2a shows that the degradation of ΔsegSNR due to positive TDOAs, as seen in Fig. 1, is overcome by increasing the filter order (L) in EMWF. For $\tau \leq 8$ ms, $\Delta\text{segSNR}_{\text{EMWF}}$ matches $\Delta\text{segSNR}_{\text{EMWF}_{\tau_0,L_1}}$, i.e., EMWF overcomes the performance loss due to the TDOA. However, at higher TDOAs, the target source in the remote microphone signal is less correlated to the target source in the HA reference microphone signal, causing only a small gain in noise reduction at higher filter orders (L). Moreover, with $L \geq 10$, $\Delta\text{segSNR}_{\text{EMWF}}$ saturates, as sustained vowels are relatively stationary, and using more past frames does not improve the estimate made using few highly correlated neighboring frames. The results obtained using speech signals in Fig. 2b, show a similar improvement as that of sustained vowel signals. From Figs. (2a) and (2b), we infer that for small TDOAs such as $\tau \leq 8$ ms, filter orders $L \leq 10$ lead to a significant gain in noise reduction and speech intelligibility, while at TDOAs, e.g., $\tau \geq 32$ ms, this performance can not be reached even with higher filter orders $L \geq 10$.

To validate the weight interpretation discussed in Sec. 3.3, Fig. 3 shows segSNR for EMWF$_{\tau_{16},L_5}$, MWF and MFWF$_{\tau_{16},L_5}$ as a function of $\text{SNR}_{\text{in,HA}}$. We consider $\tau = 16$ ms and $L = 5$ here, as $\Delta\text{segSNR}_{\text{EMWF}}$ does not show a significant improvement for $\tau \geq 16$ ms and $L \geq 10$, as seen in Fig. 2b. At $\text{SNR}_{\text{in,HA}} = -20$ dB, the performance of EMWF$_{\tau_{16},L_5}$ is nearly identical to the performance of MFWF$_{\tau_{16},L_5}$. With an increase in $\text{SNR}_{\text{in,HA}}$, the performance of EMWF$_{\tau_{16},L_5}$ closely follows the improvement in the performance of MWF. Thus, when $\text{SNR}_{\text{in,HA}}$ is low, the remote microphone signal is more valuable. Conversely, when the $\text{SNR}_{\text{in,HA}}$ is sufficiently high, the HA microphone signals become more valuable than a delayed remote microphone signal.

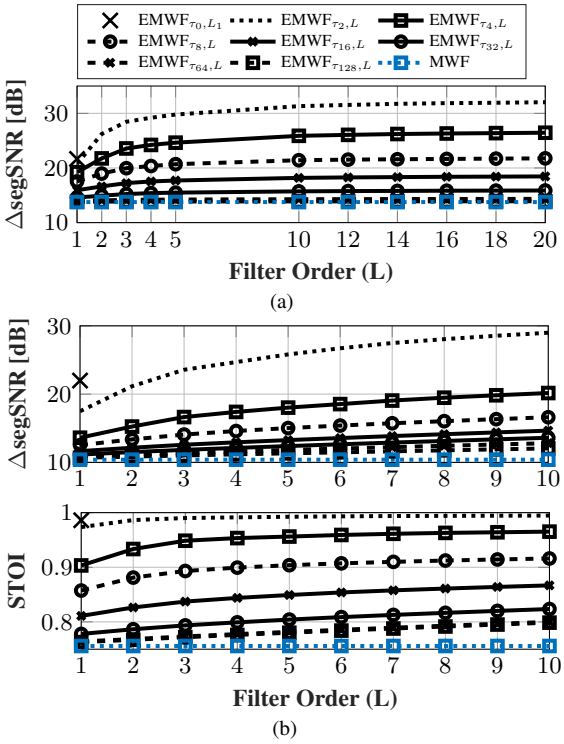

Fig. 2: Performance of EMWF for different TDOA (τ) as a function of filter order (L) for (a) sustained vowel and (b) speech signals.

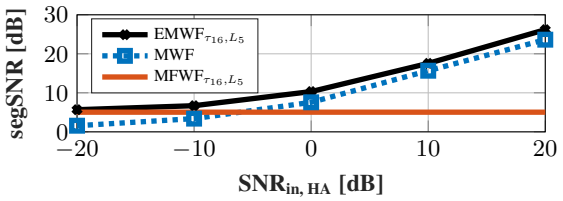

Fig. 3: segSNR as a function of $\text{SNR}_{\text{in,HA}}$ for the methods discussed, at $\tau = 16$ ms and $L = 5$, estimated for speech signals.

5. CONCLUSION

We proposed a beam-former using HA microphone signals and a remote microphone signal, to overcome the loss in performance caused by positive TDOAs between the wirelessly transmitted remote microphone signal and HA microphone signals. We show that the proposed method can be decomposed into existing MWF and MFWF methods, and the linear combination coefficients are a function of their respective output SNRs. Moreover, the proposed method was compared to existing methods, that ignore the TDOAs, using segSNR and STOI. The experimental results demonstrate that the performance, for positive TDOAs up to $\tau \leq 16$ ms, can be improved by using multiple past frames of the remote microphone signal along with the HA microphone signals. Further research includes performance assessment using estimated signal statistics.

6. REFERENCES

[1] Michael Brandstein, Darren Ward, Arild Lacroix, and Anastasios Venetsanopoulos, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[2] Simon Doclo, Walter Kellermann, Shoji Makino, and Sven Erik Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[3] Gerald R. Popelka, Brian C. J. Moore, Richard R. Fay, and Arthur N. Popper, Eds., *Hearing Aids*, vol. 56 of *Springer Handbook of Auditory Research*, Springer International Publishing, Cham, 2016.

[4] Alexander Bertrand and Marc Moonen, "Robust Distributed Noise Reduction in Hearing Aids with External Acoustic Sensor Nodes," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, Dec. 2009.

[5] Elizabeth M. Fitzpatrick, Christiane Séguin, David R. Schramm, Shelly Armstrong, and Josée Chénier, "The Benefits of Remote Microphone Technology for Adults with Cochlear Implants," *Ear & Hearing*, vol. 30, no. 5, pp. 590–599, Oct. 2009.

[6] Theresa Hnath Chisolm, Colleen M. Noe, Rachel McArdle, and Harvey Abrams, "Evidence for the Use of Hearing Assistive Technology by Adults: The Role of the FM System," *Trends in Amplification*, vol. 11, no. 2, pp. 73–89, June 2007.

[7] Nico Gößling, Daniel Marquardt, and Simon Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, 2017, pp. 1–5, IEEE.

[8] Randall Ali, Giuliano Bernardi, Toon van Waterschoot, and Marc Moonen, "Methods of Extending a Generalized Sidelobe Canceller With External Microphones," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 9, pp. 1349–1364, Sept. 2019.

[9] Sriram Srinivasan, "Using a remotewireless microphone for speech enhancement in non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5088–5091, IEEE.

[10] Bluetooth Special Interest Group (SIG), "Bluetooth core specification v5.2," December 2019.

[11] Asutosh Kar, Ankita Anand, Jan Østergaard, Søren Holdt Jensen, and M. N. S. Swamy, "Sound Quality Improvement for Hearing Aids in Presence of Multiple Inputs," *Circuits Syst Signal Process*, vol. 38, no. 8, pp. 3591–3615, Aug. 2019.

[12] Jacob Benesty and Yiteng Huang, "A single-channel noise reduction MVDR filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 273–276, IEEE.

[13] Kristian Timm Andersen and Marc Moonen, "Robust speech-distortion weighted interframe wiener filters for single-channel noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 97–107, 2018.

[14] Tsvi G. Dvorkind and Sharon Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[15] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[16] James W. Demmel, "Matrix Computations (Gene H. Golub And Charles F. van Loan)," *SIAM Rev.*, vol. 28, no. 2, pp. 252–255, June 1986.

[17] K. Uwe Simmer, Joerg Bitzer, and Claude Marro, *Post-Filtering Techniques*, pp. 39–60, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[18] Simon Doclo and Marc Moonen, *GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement*, pp. 111–132, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[19] Yiteng Arden Huang and Jacob Benesty, "A Multi-Frame Approach to the Frequency-Domain Single-Channel Noise Reduction Problem," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[20] Philipos C. Loizou, *Speech Enhancement*, CRC Press, June 2007.

[21] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[22] John S. Garofolo, Lori F. Lamel, William M. Fisher, David S. Pallett, Nancy L. Dahlgren, Victor Zue, and Jonathan G. Fiscus, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[23] Alastair H. Moore, Jan Mark de Haan, Michael Syskind Pedersen, Patrick A. Naylor, Mike Brookes, and Jesper Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 2971–2981, May 2019.