Speech enhancement using binary estimator selection applied to hearing aids with a remote microphone

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

Published in:
2023 8th International Conference on Frontiers of Signal Processing (ICFSP 2023)

DOI (link to publication from Publisher):
10.1109/ICFSP59764.2023.10372902

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

Speech enhancement using binary estimator selection applied to hearing aids with a remote microphone

Vasudha Sathyapriyan
Demant A/S, Aalborg University
Smørum, Denmark

Michael Syskind Pedersen
Demant A/S
Smørum, Denmark

Mike Brookes
Dept. of Electrical and Electronic Engineering
Imperial College London
London, United Kingdom

Jan Østergaard
Department of Electronic Systems
Aalborg University
Aalborg, Denmark

Patrick A. Naylor
Dept. of Electrical and Electronic Engineering
Imperial College London
London, United Kingdom

Jesper Jensen
Demant A/S, Aalborg University
Smørum, Denmark

***Abstract*—This paper introduces a speech enhancement algorithm for hearing assistive devices, e.g., hearing aids, connected to a remote microphone. Remote microphones are especially beneficial to hearing aid users when they are present in environments with low signal-to-noise ratios. The transmission of the acoustic data from the remote microphone to the hearing aid unit, however, happens through a wireless channel that is prone to network delays. Such delays, that occur in any real-world application, make the remote microphone signal less valuable, in contrast to when the transmission is assumed to be error-free and instantaneous, as is often done in literature. To make use of the remote microphone signal, despite the delay, we propose an estimator selection method that selects between the minimum mean-square error estimate of the desired signal, made using the hearing aid signals and the delayed remote microphone signal, respectively. This binary selection is made by comparing the normalized mean-square errors of the two desired signal estimates. We show that the proposed method provides a benefit in estimated speech intelligibility, for delays in transmission up to 30 ms at a signal-to-noise ratio of 0 dB, in comparison to the minimum mean-square error estimate made using only the hearing aid microphone signals.**

***Index Terms*—Wireless transmission delay, hearing aids, speech processing, multi-channel.**

I. Introduction

Speech enhancement methods using microphone arrays can improve speech quality and intelligibility, particularly for spatially separated sound sources [1]. Such multi-channel methods are beneficial for a wide range of applications, including hearing aid (HA) applications [2]. People who are hard of hearing, however, continue facing difficulties in following conversations, especially in noisy environments. In such difficult listening environments, by expanding the HA system with an additional remote microphone (RM), that is closer to the target source, could benefit HA users [3].

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.

Multi-channel noise reduction methods, that use HA microphones in combination with RMs have shown performance benefits under instantaneous and error-free wireless transmission assumptions [4], [5]. In reality, however, wireless transmission protocols introduce audio latency, due to encoding, transmission and decoding delay, that depend on the codec and buffer size used in the transmission pipeline, e.g., 20 ms to hundreds of milliseconds for Bluetooth LE with the LC3 codec [6]. Fig. 1 depicts the delays introduced during the acoustic and wireless transmission of the target source signal, $s(n+d_{a_{\text{ref}}})$, from the source to the HA user. The time difference of arrival (TDOA), τ [s], between the RM signal and the HA reference microphone signal, received at the HA, is given by the difference between the acoustic propagation delay, τ_a [s], and the wireless transmission delay, τ_w [s], i.e., $\tau = \tau_w - \tau_a$, as shown in Fig. 1. The TDOA, τ, will depend on the wireless protocol chosen, the location of the microphones and the distance between the target source and the microphones [3]. The TDOA, τ, can either be positive, i.e., the RM signal arrives later than the acoustic signal at the HA microphone, or negative, i.e., the acoustic signal at the HA microphone arrives later than the RM signal. When $\tau<0$, the RM signal can be stored in a buffer and used when the acoustic signal arrives at the HA microphones, thereby making it a trivial case. However, for $\tau>0$, the benefit of the RM signal reduces with increasing τ [7], [8].

One use of the RM in the HA application would be to simply playback the RM signal for the HA user, when received at the HA. However, this solution is limited to when $0 \leq \tau \leq 10$ ms, to avoid undesirable audio effects, e.g., echoes [9], [10]. Table I presents the benefits and drawbacks of using the HA microphone signals and the RM signal independently. From Table I, we infer that the benefit of one microphone may complement the weakness of the other, emphasizing the need to jointly process the HA and the RM signals.

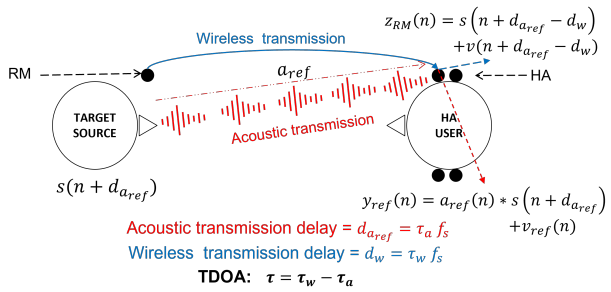

Fig. 1: Acoustic and wireless signal transmission model.

TABLE I: Trade-off between using only hearing aid microphones versus using only remote microphones

Hearing aid (HA)	Remote microphone (RM)
Low SNR in noisy environments	Potentially high SNR in noisy environments
Airborne acoustic channel latency	Wireless channel latency
Head and torso cues present	No head and torso cues

In contrast to existing work [4], [5], that assume instantaneous and error-free RM signal transmission, in [8], a method that uses both HA and RM signals under realistic delay assumptions, was proposed. It was found that the method provides a benefit for $\tau \leq 32$ ms, however, using *oracle second-order statistics (SOS)*. In particular, [8], shows that the minimum mean-square error (MMSE) estimator of the target signal at the HA, estimated using HA and RM signals, is a linear combination of a multi-channel Wiener filter (MWF) applied to the HA microphone signals and a linear prediction filter (LPF) applied to the delayed RM signal. However, implementing this method, by directly replacing the oracle SOS with practically estimated SOS, e.g., [11], due to highly time-varying speech statistics, led to unsatisfactory results.

Thus, we simplify the method in [8], where for each time-frequency (TF) tile, we choose the better of the two desired signal estimates. We demonstrate that the proposed method, which is implemented in practice without oracle information, improves the desired signal estimate, in the presence of TDOAs, $0 \leq \tau \leq 30$ ms at a signal-to-noise ratio (SNR) of 0 dB. The simplicity of the proposed method enables HAs to utilize a high-SNR RM signal in the presence of wireless transmission delays in real-world applications.

II. Signal Model and Notation

We consider an acoustic system that is composed of a hearing aid (HA) with $M \geq 2$ microphones and a remote microphone (RM), cf. Fig. 1. Let $a_m(n)$ denote the hearing aid head related impulse response (HAHRIR) from the target source to the m^{th} HA microphone, and let d_{a_m} [samples] denote the acoustic propagation delay, i.e., the number of 0's in $a_m(n)$ preceding the first impulse. Furthermore, let $a_{\text{ref}}(n)$ and $d_{a_{\text{ref}}}(n)$ denote the HAHRIR and its respective acoustic propagation delay at the HA reference microphone, and let $s(n + d_{a_{\text{ref}}})$ denote the target source signal. Then the noisy signal at the m^{th} HA microphone is given by

$$y_m(n) = \underbrace{a_m(n) * s(n + d_{a_{\text{ref}}})}_{x_m(n)} + v_m(n),\ m = 1 \dots M \tag{1}$$

where $y_m(n), x_m(n)$ and $v_m(n)$ are the noisy signal, speech and noise components present in the m^{th} HA microphone signal, respectively. The noise in every microphone in the acoustic system is assumed to be uncorrelated to its corresponding speech component. The M-channel stacked noisy multi-channel HA microphone signal vector, $\underline{y}_{\text{HA}}(n) \in \mathbb{R}^{M\times 1}$, can be written as

$$\underline{y}_{\text{HA}}(n) = [y_1(n), \dots, y_M(n)]^T = \underline{x}_{\text{HA}}(n) + \underline{v}_{\text{HA}}(n), \tag{2}$$

where, $\underline{x}_{\text{HA}}(n)$ and $\underline{v}_{\text{HA}}(n)$ are the multi-channel speech and the noise HA microphone signal vectors, defined like $\underline{y}_{\text{HA}}(n)$.

The signal recorded by the RM, that is received at the HA after wireless transmission can be written as

$$z_{\text{RM}}(n) = s(n + d_{a_{\text{ref}}} - d_w) + v(n + d_{a_{\text{ref}}} - d_w), \tag{3}$$

where, $z_{\text{RM}}(n)$ is the noisy signal from the RM, when received at the HA, $s(n + d_{a_{\text{ref}}} - d_w)$ and $v(n + d_{a_{\text{ref}}} - d_w)$ are the speech and noise signals at the RM, which are delayed by d_w [samples] due to the wireless transmission delay. For convenience, we assume there is no coloration of the signal picked up by the RM microphone and that the codec introduces negligible waveform reproduction error. The speech component, $x_{\text{ref}}(n)$, at the HA reference microphone is desirable to the HA user, as this signal contains head and torso reflections which are beneficial for sound localisation and intelligibility [12]. Thus, to impose these reflections onto the RM signal in (3), we convolve it with the HAHRIR, $a_{\text{ref}}(n)$, from the target source to the HA reference microphone. The noisy RM signal received at the HA, then becomes

$$\begin{aligned} y_{\text{RM}}(n) = &\underbrace{a_{\text{ref}}(n) * s(n + d_{a_{\text{ref}}} - d_w)}_{x_{\text{RM}}(n)} + \\ &\underbrace{a_{\text{ref}}(n) * v(n + d_{a_{\text{ref}}} - d_w)}_{v_{\text{RM}}(n)}. \end{aligned} \tag{4}$$

Now, $\underline{y}_{\text{HA}}(n)$ and $y_{\text{RM}}(n)$ are signals at the HA. Due to the convolution in (4), the TDOA, τ, between the HA reference microphone signal, $y_{\text{ref}}(n)$, and the RM signal received at the HA, $y_{\text{RM}}(n)$, is now $\tau = \tau_w$.

The signals are processed in the short-time Fourier transform (STFT) domain, thus the noisy multi-channel HA microphone signal vector, $\mathbf{y}_{\text{HA}}(l,k) \in \mathbb{C}^{M\times 1}$ becomes

$$\begin{aligned} \mathbf{y}_{\text{HA}}(l,k) &= \mathbf{d}(k)X_{\text{ref}}(l,k) + \mathbf{v}_{\text{HA}}(l,k) \\ &= \mathbf{x}_{\text{HA}}(l,k) + \mathbf{v}_{\text{HA}}(l,k), \end{aligned} \tag{5}$$

where, $X_{\text{ref}}(l,k)$ is the STFT coefficient of the speech component at the HA reference microphone signal, $\mathbf{d}(l,k)$ is the relative acoustic transfer function (RATF) vector from the target source to the HA microphones, $\mathbf{x}_{\text{HA}}(l,k)$ and $\mathbf{v}_{\text{HA}}(l,k)$ are the speech and noise STFT coefficient vectors of the HA

microphone signals, respectively, and l, k are the time-frame and frequency bin indices.

As in [8], the multi-frame noisy RM signal vector, $\mathbf{y}_{\mathrm{RM}}(l,k) \in \mathbb{C}^{L\times 1}$, can be written as

$$\begin{aligned}\mathbf{y}_{\mathrm{RM}}(l,k) &= [Y_{\mathrm{RM}}(l,k), \ldots, Y_{\mathrm{RM}}(l-L+1,k)]^T \\ &= \mathbf{s}_{\mathrm{RM}}(l,k) + \mathbf{v}_{\mathrm{RM}}(l,k),\end{aligned} \tag{6}$$

where, $Y_{\mathrm{RM}}(l,k)$ is the STFT coefficient of the noisy RM signal in (4), $\mathbf{s}_{\mathrm{RM}}(l,k)$, $\mathbf{v}_{\mathrm{RM}}(l,k)$ are the multi-frame speech and noise component vectors of the RM signal respectively, defined similar to $\mathbf{y}_{\mathrm{RM}}(l,k)$, and L is the number of past frames considered. In the rest of the paper, we ignore l, k indices for writing convenience.

The cross power spectral density matrices (CPSDMs) of the noisy HA microphone vector, $\Sigma_{\mathbf{y}_{\mathrm{HA}}} \in \mathbb{C}^{M\times M}$, and the multi-frame noisy RM vector, $\Sigma_{\mathbf{y}_{\mathrm{RM}}} \in \mathbb{C}^{L\times L}$, are defined as

$$\begin{aligned}\Sigma_{\mathbf{y}_{\mathrm{HA}}} &\triangleq \mathbb{E}\left[\mathbf{y}_{\mathrm{HA}}\,\mathbf{y}_{\mathrm{HA}}^H\right] = \Sigma_{\mathbf{x}_{\mathrm{HA}}} + \Sigma_{\mathbf{v}_{\mathrm{HA}}}, \\ \Sigma_{\mathbf{y}_{\mathrm{RM}}} &\triangleq \mathbb{E}\left[\mathbf{y}_{\mathrm{RM}}\,\mathbf{y}_{\mathrm{RM}}^H\right] = \Sigma_{\mathbf{x}_{\mathrm{RM}}} + \Sigma_{\mathbf{v}_{\mathrm{RM}}},\end{aligned} \tag{7}$$

respectively, where $\Sigma_{\mathbf{x}_{\mathrm{HA}}}, \Sigma_{\mathbf{v}_{\mathrm{HA}}}$ and $\Sigma_{\mathbf{x}_{\mathrm{RM}}}, \Sigma_{\mathbf{v}_{\mathrm{RM}}}$ are the CPSDMs of the speech and noise components at the HA and RM vectors, respectively, defined like $\Sigma_{\mathbf{y}_{\mathrm{HA}}}, \Sigma_{\mathbf{y}_{\mathrm{RM}}}$. The power spectral density (PSD) of the desired signal, X_{ref} is $\sigma^2_{X_{\mathrm{ref}}} \triangleq \mathbb{E}\left[X_{\mathrm{ref}} X^*_{\mathrm{ref}}\right]$.

In [8], the desired signal was estimated as a linear combination of two MMSE estimators, an MWF and an LPF as

$$\hat{X}_{\mathrm{ref}} = \alpha\,\hat{X}_{\mathrm{MWF}} + \beta\,\hat{X}_{\mathrm{LPF}}, \tag{8}$$

with

$$\hat{X}_{\mathrm{MWF}} = \mathbf{h}_{\mathrm{MWF}}^H\,\mathbf{y}_{\mathrm{HA}}, \quad \hat{X}_{\mathrm{LPF}} = \mathbf{h}_{\mathrm{LPF}}^H\,\mathbf{y}_{\mathrm{RM}}, \tag{9}$$

where $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$ are the MWF and the LPF estimates of the desired signal, X_{ref}, respectively, $\mathbf{h}_{\mathrm{MWF}}^H$ and $\mathbf{h}_{\mathrm{LPF}}^H$ are the filter coefficients of the MWF and the LPF, respectively. Moreover, it was found that α, β are functions of the output SNRs of $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$, respectively. Motivated by this, in the next section, we propose a binary estimator selection method, that selects between $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$ to estimate the desired signal.

III. Proposed Method

First, we describe the proposed method and the threshold criterion used for the binary estimator selection. Next, we describe how to estimate the normalised mean-square errors (nMSEs) of $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$, that are required to make the binary estimator selection. The section concludes with a summary of the proposed method.

A. Binary Estimator Selection

Motivated by [8], we propose a binary estimator selection method, using practically estimated SOS, which compares the two estimates, $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$, based on their nMSEs in each TF tile and selects the better estimate of the two. The proposed estimate can be expressed as

$$\hat{X}_{\mathrm{prop}} = g\,\hat{X}_{\mathrm{MWF}} + (1-g)\,\hat{X}_{\mathrm{LPF}}, \tag{10}$$

where

$$g = \begin{cases} 0, & \text{if } \frac{\mathrm{nMSE}_{\mathrm{MWF}}}{\mathrm{nMSE}_{\mathrm{LPF}}} \geq 1 \\ 1, & \text{otherwise} \end{cases}. \tag{11}$$

From (10) and (11), we infer that for high τ, most often, $\frac{\mathrm{nMSE}_{\mathrm{MWF}}}{\mathrm{nMSE}_{\mathrm{LPF}}} < 1$, and $\hat{X}_{\mathrm{prop}}$ would tend to select $\hat{X}_{\mathrm{MWF}}$. Similarly, for low τ, then typically, $\frac{\mathrm{nMSE}_{\mathrm{MWF}}}{\mathrm{nMSE}_{\mathrm{LPF}}} \geq 1$, and $\hat{X}_{\mathrm{prop}}$ would tend to select $\hat{X}_{\mathrm{LPF}}$.

B. Estimation of nMSE in Multi-channel Wiener Filter

The optimisation problem of MWF [13], using HA microphone signals is

$$\mathbf{h}_{\mathrm{MWF}} = \arg\min_{\mathbf{h}}\ \mathbb{E}\left[\left|\mathbf{h}^{\mathrm{H}}\mathbf{y}_{\mathrm{HA}} - X_{\mathrm{ref}}\right|^2\right], \tag{12}$$

where $\mathbf{h}_{\mathrm{MWF}}$ can be decomposed into a minimum variance distortion-less response (MVDR) filter and a single-channel post filter [14] as

$$\mathbf{h}_{\mathrm{MWF}} = \underbrace{\frac{\Sigma^{-1}_{\mathbf{v}_{\mathrm{HA}}}\mathbf{d}}{\mathbf{d}^H\Sigma^{-1}_{\mathbf{v}_{\mathrm{HA}}}\mathbf{d}}}_{\mathbf{h}_{\mathrm{MVDR}}}\left(\frac{\xi}{1+\xi}\right), \tag{13}$$

where $\xi \triangleq \sigma^2_{X_{\mathrm{ref}}}\mathbf{d}^H\Sigma^{-1}_{\mathbf{v}_{\mathrm{HA}}}\mathbf{d}$, is the SNR at the output of the MVDR beamformer. Using (12) and (13), the MMSE of $\hat{X}_{\mathrm{MWF}}$ is

$$\mathrm{MMSE}_{\mathrm{MWF}} = \sigma^2_{X_{\mathrm{ref}}}\left[1 - \sigma^2_{X_{\mathrm{ref}}}\mathbf{d}^H\Sigma^{-1}_{\mathbf{y}_{\mathrm{HA}}}\mathbf{d}\right]. \tag{14}$$

Taking the CPSDM of the speech component to be rank-1, i.e., $\Sigma_{\mathbf{x}_{\mathrm{HA}}} = \sigma^2_{X_{\mathrm{ref}}}\mathbf{d}\mathbf{d}^H$, for a single target source, and by using the matrix inversion lemma [15], the nMSE of $\hat{X}_{\mathrm{MWF}}$ (normalised by $\sigma^2_{X_{\mathrm{ref}}}$) may be expressed as

$$\mathrm{nMSE}_{\mathrm{MWF}} = \left[\frac{1}{1+\sigma^2_{X_{\mathrm{ref}}}\mathbf{d}^H\Sigma^{-1}_{\mathbf{v}_{\mathrm{HA}}}\mathbf{d}}\right] = \left(\frac{1}{1+\xi}\right). \tag{15}$$

C. Estimation of nMSE in Linear Prediction Filter

In order to derive nMSE of $\hat{X}_{\mathrm{LPF}}$, we rely on [8], where multiple past time-frames of the delayed RM signal were used to estimate the desired speech at the current STFT and therby reduce the TDOA. The LPF optimisation problem is

$$\mathbf{h}_{\mathrm{LPF}} = \arg\min_{\mathbf{h}}\ \mathbb{E}\left[\left|\mathbf{h}^{\mathrm{H}}\mathbf{y}_{\mathrm{RM}} - X_{\mathrm{ref}}\right|^2\right], \tag{16}$$

we get

$$\mathbf{h}_{\mathrm{LPF}} = \Sigma^{-1}_{\mathbf{y}_{\mathrm{RM}}}\Sigma_{\mathbf{y}_{\mathrm{RM}}X_{\mathrm{ref}}}, \tag{17}$$

where $\Sigma_{\mathbf{y}_{\mathrm{RM}}X_{\mathrm{ref}}} \triangleq \mathbb{E}\left[\mathbf{y}_{\mathrm{RM}}X^*_{\mathrm{ref}}\right]$. Using (16) and (17), the MMSE of $\hat{X}_{\mathrm{LPF}}$ is

$$\mathrm{MMSE}_{\mathrm{LPF}} = \sigma^2_{X_{\mathrm{ref}}} - \Sigma^H_{\mathbf{y}_{\mathrm{RM}}X_{\mathrm{ref}}}\mathbf{h}_{\mathrm{LPF}}, \tag{18}$$

and the nMSE (normalised by $\sigma^2_{X_{\mathrm{ref}}}$) may be expressed as

$$\mathrm{nMSE}_{\mathrm{LPF}} = 1 - \frac{\Sigma^H_{\mathbf{y}_{\mathrm{RM}}X_{\mathrm{ref}}}\mathbf{h}_{\mathrm{LPF}}}{\sigma^2_{X_{\mathrm{ref}}}}. \tag{19}$$

Algorithm 1 Proposed method of binary estimator selection

Require: τ, voice activity detector (VAD).

for k bins **do**
  Compute $\mathbf{h}_{\mathrm{LP}}(k)$ and $\mathrm{nMSE}_{\mathrm{LP}}(k)$ using (16) and (19) when speech is present.
  for l frames **do**
    Compute $\Sigma_{\mathbf{v}_{\mathrm{HA}}}$ when speech is absent, as

$$\Sigma_{\mathbf{v}_{\mathrm{HA}}}(l,k) = \alpha_{\mathrm{v}} \Sigma_{\mathbf{v}_{\mathrm{HA}}}(l-1,k) + (1-\alpha_{\mathrm{v}}) \mathbf{y}_{\mathrm{HA}}(l,k) \mathbf{y}_{\mathrm{HA}}^{H}(l,k)$$

    Compute $\Sigma_{\mathbf{y}_{\mathrm{HA}}}$ when speech is present, as

$$\Sigma_{\mathbf{y}_{\mathrm{HA}}}(l,k) = \alpha_{\mathrm{y}} \Sigma_{\mathbf{y}_{\mathrm{HA}}}(l-1,k) + (1-\alpha_{\mathrm{y}}) \mathbf{y}_{\mathrm{HA}}(l,k) \mathbf{y}_{\mathrm{HA}}^{H}(l,k)$$

    Compute $\mathbf{d}(l,k)$, e.g., using method proposed in [16].
    Compute $X_{\mathrm{MWF}}(l,k)$ and $\mathrm{nMSE}_{\mathrm{MWF}}(l,k)$ using (13) and (15).
    Compute $g(l,k)$ using (11), (15) and (19).
    To avoid musical artifacts, smooth the binary estimator selector, $g(l,k)$ by

$$\overline{g}(l,k) = 0.8\, \overline{g}(l-1,k) + 0.2\, g(l,k)$$

    Compute X_{prop} using

$$X_{\mathrm{prop}}(l,k) = \overline{g}(l,k) \hat{X}_{\mathrm{MWF}}(l,k) + (1-\overline{g}(l,k)) \hat{X}_{\mathrm{LP}}(l,k)$$

  end for
end for

D. Algorithm summary

The proposed method is summarized in Algorithm 1.

IV. Simulation and results

To demonstrate the benefit of the proposed method, over traditional methods used in HAs, that do not use RMs, we evaluate the performance of the MWF, LPF and the binary estimator selection method in simulated acoustic scenes similar to Fig. 1.

A. Experiment setup and implementation

To simulate the acoustic scene in Fig. 1, we consider a single target speech source, $s(n + d_{a_{\mathrm{ref}}})$, located in front of a HA user in ambient noise. A left monaural HA with $M = 2$ microphones is worn by the HA user and we assume the RM is worn by the target source, thus picking up the target speech at a higher SNR than the HA microphones. For the target speech source, we use 30 s of 5 male and 5 female speech signals, from the CSTR VCTK Corpus [17]. The measured HAHRIRs [18], are used to simulate the speech signals at the HA microphones. The distance between the HA user and the target source is 1.9 m. We use cylindrical isotropic speech shaped noise (SSN) and ambient cafeteria noise [18], as the noise types in the HA microphones and the RM. To simulate the RM signal, $z_{\mathrm{RM}}(n)$, received at the HA, we add uncorrelated noise to the target source signal, $s(n + d_{a_{\mathrm{ref}}})$, and delay it by $\tau_w \in [0, 40]$ ms, to simulate the wireless transmission delay. The simulations are conducted with the HA reference microphones at SNRs $\in \{-10, -5, 0, 5, 10\}$ dB and the RM signals at 20 dB SNR, due to its proximity to the target source. Different SNRs at the HA were used to study the trade-off between the SNRs and the TDOAs.

By assuming the noise components to be uncorrelated in the HA reference microphone signal, $y_{\mathrm{ref}}(n)$, and the delayed RM signal received at the HA, $z_{\mathrm{RM}}(n)$, the HAHRIR, $a_{\mathrm{ref}}(n)$, from the target source to the HA reference microphone, can be estimated recursively, e.g., using the least-mean-square (LMS) algorithm [19]. We use batch LMS algorithm, with 12 s of the practically available $y_{\mathrm{ref}}(n)$ and $z_{\mathrm{RM}}(n)$ signals, to estimate a 400 tap-length HAHRIR filter, $a_{\mathrm{ref}}(n)$. The signals are processed in the STFT domain with a Hann window of 8 ms and a 25% frame-shift. For the LPF method, we use $L = 3$, as found to be reasonable in [8]. The noise CPSDMs of the HA microphone signals are updated recursively with α_{v} corresponding to a time constant of 250 ms, during the noise only frames, detected using an ideal voice activity detector (VAD). For the LPF method, we perform an offline estimation of the SOS, due to the high estimation errors found with recursively estimated SOS, during onsets and offsets of speech. The RATF, $\mathbf{d}$, required in (13), is estimated using [16]. We assume the TDOA, τ, can be estimated from $a_{\mathrm{ref}}(n)$ and with the knowledge of the specific codec used.

B. Results

As mentioned, we simulated acoustic scenes at SNRs $\in \{-10, -5, 0, -5, 10\}$ dB. When the SNR at the HA microphones is low, the proposed method, $\hat{X}_{\mathrm{prop}}$, would tend to select $\hat{X}_{\mathrm{LPF}}$ across a greater number of TF tiles than $\hat{X}_{\mathrm{MWF}}$. This is because, the $\mathrm{nMSE}_{\mathrm{MWF}}$ would vary inversely with SNR. However, due to space limitations we only present the results obtained for the HA microphones at an SNR of 0 dB. Fig. 2 reports the performance in terms of mean-square error (MSE) [20], and estimated extended short-term objective intelligibility (ESTOI) [21], as a function of TDOA, τ for SSN (Figs. (2a) and (2c)) and cafeteria noise (Figs. (2b) and (2d)). The proposed method, $\hat{X}_{\mathrm{prop}}$, selects between $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$ at each TF tile. Thus, with ideal choices, we would expect the proposed method to perform either better or equivalent to either the $\hat{X}_{\mathrm{MWF}}$ or $\hat{X}_{\mathrm{LPF}}$. To validate this, we estimate the MSEs of $\hat{X}_{\mathrm{MWF}}$, $\hat{X}_{\mathrm{LPF}}$ and $\hat{X}_{\mathrm{prop}}$. The performance scores reported are averaged over the 10 talkers.

Generally, as expected, $\hat{X}_{\mathrm{MWF}}$ which only uses the HA microphone signals, improves over the noisy, unprocessed HA reference microphone signal, Y_{ref}. In Fig. 2, the performance of $\hat{X}_{\mathrm{MWF}}$ and Y_{ref} do not vary with τ, as they are independent of the RM signal. Focusing on $\hat{X}_{\mathrm{LPF}}$, which uses only the RM signal, Figs. (2c) and (2d) demonstrate the benefit of $\hat{X}_{\mathrm{LPF}}$ over $\hat{X}_{\mathrm{MWF}}$ in speech intelligibility, particularly for lower values of τ, where speech can be predicted well. However, for $\tau > 20$ ms in SSN and for $\tau > 28$ ms in cafeteria noise, using $\hat{X}_{\mathrm{LPF}}$ can be more harmful than using the noisy HA reference signal. Thus, by making a binary selection between $\hat{X}_{\mathrm{MWF}}$ and $\hat{X}_{\mathrm{LPF}}$, the proposed method, $\hat{X}_{\mathrm{prop}}$, overall for both SSN and cafeteria noise, uses the benefit of $\hat{X}_{\mathrm{LPF}}$ for

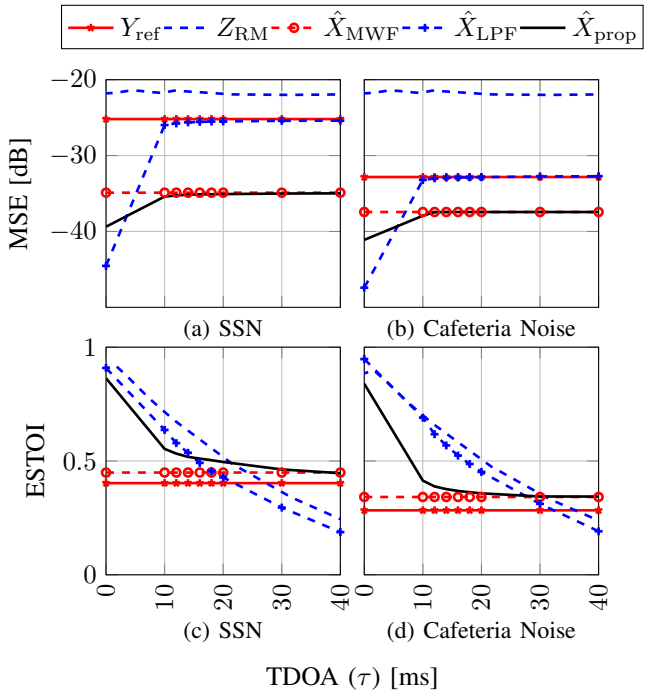

Fig. 2: Performance in terms of (a), (b) MSE and (c), (d) ESTOI for different TDOAs (τ), for SSN and Cafeteria Noise at 0 dB SNR.

$0 \leq \tau < 30$ ms and tends to select $\hat{X}_{\text{MWF}}$ for $\tau > 30$ ms. Therefore, particularly for the acoustic scene considered here, where the HA signals are at 0 dB SNR, the proposed method, that uses the RM signal, provides a benefit for $0 \leq \tau \leq 30$ ms, over MWF that uses only HA microphone signals. Audio examples of the methods discussed are provided here.

Presently, there exist no intrusive objective measures to suitably evaluate signals with a TDOA w.r.t the reference signal, making it difficult to analyze the performance of Z_{RM} in comparison to $\hat{X}_{\text{prop}}$. This can be observed in Fig. 2, where Z_{RM} appears to perform better than $\hat{X}_{\text{prop}}$ in terms of ESTOI, whereas it appears to perform worse in terms of MSE. Furthermore, since Z_{RM} is delayed w.r.t X_{ref}, i.e., $\tau \geq 0$ ms, cf. Sec. I, a direct playback of the Z_{RM} for $\tau > 10$ ms would cause undesirable perceptual effects [10]. These effects can only be evaluated through a listening test, which is out of scope of this work and will be a part of future research.

V. Conclusion

We proposed a binary estimator selection algorithm that is applied to a HA connected to a RM, without ignoring the realistic TDOA between the HA and RM signals. The TDOA, τ, between the HA microphone signal and the RM signal arriving at the HA, can introduce undesirable audio effects to the HA user when $\tau > 10$ ms. Therefore, we proposed a simple, HA friendly method that uses the signals from the HA microphone and RM, by making a binary selection between MMSE estimate of the desired signal, made using the HA signals and the RM signal, by comparing their nMSEs. The proposed method provides a benefit in terms of estimated speech intelligibility, for $0 \leq \tau \leq 30$ ms, over MWF, which uses only the HA microphone signals.

References

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[3] G. R. Popelka, B. C. Moore, R. R. Fay, and A. N. Popper, *Hearing aids*, ser. Springer Handbook of Auditory Research. Springer International Publishing, 2016, vol. 56.

[4] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field," in *IEEE Hands-Free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 1–5.

[5] R. Ali, G. Bernardi, T. Van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1349–1364, 2019.

[6] B. S. I. G. (SIG), "Bluetooth core specification v5.2," 2019.

[7] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on Advances in Signal Processing*, pp. 1–14, 2009.

[8] V. Sathyapriyan, M. S. Pedersen, J. Østergaard, M. Brookes, P. A. Naylor, and J. Jensen, "A linear MMSE filter using delayed remote microphone signals for speech enhancement in hearing aid applications," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[9] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[10] L. Bramsløw, "Preferred signal path delay and high-pass cut-off in open fittings," *International Journal of Audiology*, vol. 49, no. 9, pp. 634–644, 2010.

[11] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1355–1365, 2014.

[12] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1983.

[13] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on signal processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[14] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," *Microphone arrays: Signal processing techniques and applications*, pp. 39–60, 2001.

[15] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., ser. Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Press, 1996.

[16] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint maximum likelihood estimation of power spectral densities and relative acoustic transfer functions for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6119–6123.

[17] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[18] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on advances in signal processing*, vol. 2009, pp. 1–10, 2009.

[19] S. Haykin, *Adaptive Filter Theory*, 2nd ed. Prentice-Hall, Inc., 1991.

[20] P. C. Loizou, *Speech enhancement: theory and practice*, 2nd ed. CRC press, 2013.

[21] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.