Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices

Hoang, Poul; De Haan, Jan Mark; Tan, Zheng Hua; Jensen, Jesper

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2022.3145294

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

Multichannel Speech Enhancement with Own Voice-Based Interfering Speech Suppression for Hearing Assistive Devices

Poul Hoang, *Student member, IEEE,* Jan Mark de Haan, Zheng-Hua Tan, *Senior Member, IEEE,* and Jesper Jensen

***Abstract*—Enhancement of a desired speech signal in the presence of competing or interfering speech remains an unsolved problem, as it can be hard to determine which of the speech signals is the one of interest. In this paper, we propose a multichannel noise reduction algorithm which uses the presence of the user's own voice signal, e.g. during conversations with the target speaker, as an asset to efficiently identify interfering speech and noise. Specifically, following the typical speech pattern in natural conversations, the presence of an own voice may indicate the absence of the target speech, hence undesired speech and noise can be identified and estimated during own voice presence.**

In contrast to conventional noise reduction systems, the proposed noise reduction systems use the user's own voice to identify interfering speech that otherwise could be confused with the target speech. We demonstrate the performance of the proposed noise reduction systems in a comparison against state-of-the-art noise reduction systems in terms of beamforming performance for hearing assistive devices. The results show that the proposed beamforming scheme in particular outperforms state-of-the-art methods in terms of ESTOI and PESQ in situations with a target speaker and a strong interfering speaker.

***Index Terms*—Speech Enhancement, beamforming, maximum likelihood, turn-taking, speech behavior.**

I. Introduction

Spoken language is for most people their primary way of communicating in many social situations. Speech, however, may become challenging to understand, when the acoustic environment becomes increasingly noisy. Especially, when the acoustic environment is contaminated with many competing speakers or interferers, speech intelligibility is often poor.

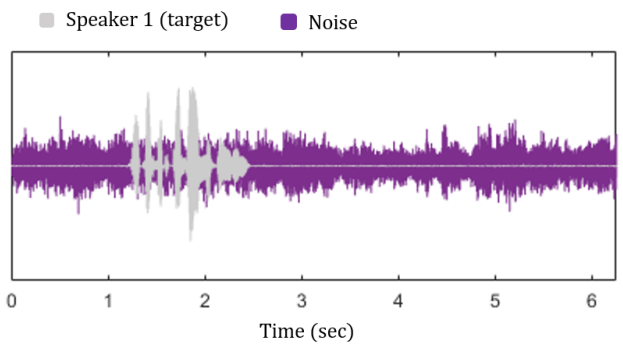
One of the purposes of hearing assistive devices (HADs), e.g. hearing aids (HAs), is to increase speech intelligibility and quality by reducing the background noise. This is commonly achieved with the use of noise reduction algorithms such as beamformers, when multiple microphones are accessible [1]–[3]. Examples of well-known beamformers are the minimum-variance distortion-less response (MVDR), the multichannel Wiener filter (MWF) and the linear constrained minimum variance (LCMV) beamformers [2]–[4]. Implementation of these beamformers is often done in the time-frequency (TF) domain and the parameters required are typically noise statistics, e.g. the noise cross power spectral density (CPSD) matrix [3] and the relative acoustic transfer function (RATF) vector of the target source [4]. These parameters are, however, rarely known in real-world situations and therefore have to be estimated.

One approach to estimate the noise CPSD matrix is to use noise dominant TF tiles to update the noise CPSD matrix, and use the resulting estimate during speech presence, e.g. [5]. Detecting noise dominant TF tiles requires a voice activity detector (VAD) or, more generally, speech presence probabilities (SPPs) estimated from the noisy microphone signals. Multichannel methods for estimating the speech presence probability have been proposed in [6]–[11]. These methods update the noise CPSD matrix in a soft-decision manner using a multichannel extension of the minima controlled recursive average procedure in [12,13]. These methods may perform less well if only few noise dominant TF tiles can be identified or if the noise is highly non-stationary during speech dominated TF tiles. To overcome this issue, several methods have been proposed to update the noise statistics using speech dominant TF tiles as well. For example, methods presented in [14]–[21] are maximum likelihood estimators (MLEs) of the noise CPSD matrix under the assumption that the spatial coherence of the noise field remains fixed during speech presence. As a consequence, these methods may perform less well when the spatial properties of the noise field change during speech presence. An example where this occurs, is when a non-stationary interfering source, e.g. a competing speaker emerges in the noise field. In [22], an MLE of the interference-plus-noise CPSD matrix was proposed to handle situations with strong interfering speech and noise. However, the method requires that the target RATF vector is known in advance.

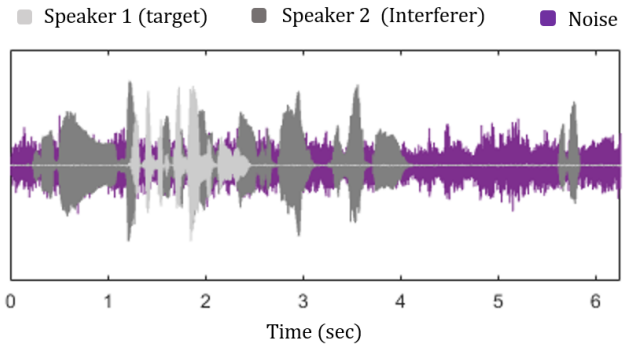
Accurate target localization and target RATF vector estimation are crucial for beamformers to steer the acoustic beam towards the target speaker [4]. In acoustic scenarios with interfering speakers, target RATF vector estimation can be particularly difficult. The problem of identifying a target speaker amongst a set of interfering speakers and background noise is essentially ill-posed: without any additional information, it is very difficult to single out the target speaker from the set of active speakers. Hence, in order to identify the target speaker, existing methods have applied various prior knowledge. For example, the widely used steered-response methods [3, ch. 8] implicitly rely on the assumption that the target source is closer in distance, and hence more powerful, than other sound sources. These methods identify the target source by

P. Hoang is with the Department of Electronic Systems, Aalborg University, Aalborg Øst 9220, Denmark, and Oticon A/S, Smørum 2750, Denmark.
J. M. de Haan is with Oticon A/S, 2750, Smørum, Denmark.
Z.-H. Tan is with the Department of Electronic Systems, Aalborg University, Aalborg Øst 9220, Denmark, and the Pioneer Centre for AI, Denmark.
J. Jensen is with the Department of Electronic Systems, Aalborg University, Aalborg Øst 9220, Denmark, and Oticon A/S, Smørum 2750, Denmark.

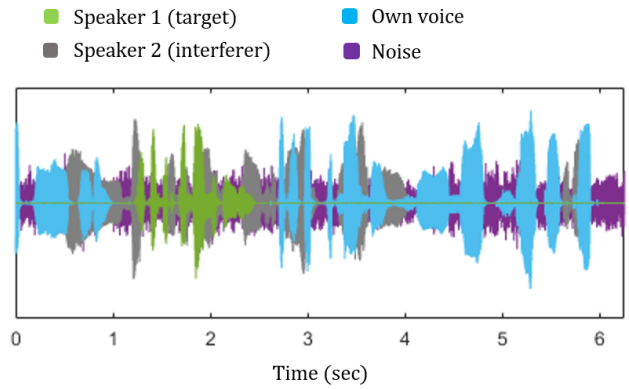
Manuscript received xx xx, xxxx; revised xx xx, xxxx.



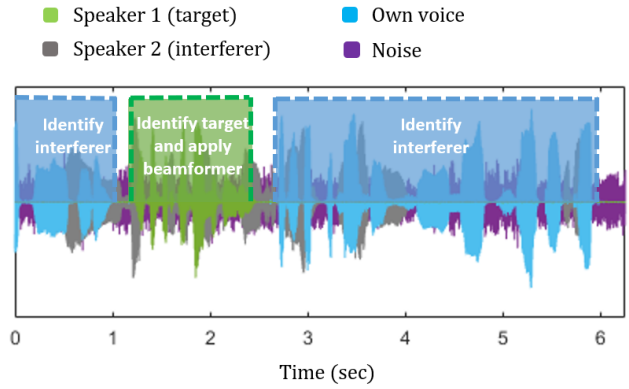
(a) Target speech in noise.



(b) Target speech and interfering speech in noise.



(c) Target speech and own voice in an environment with an interferer and noise.



(d) The proposed systems identify the interferer during own voice presence and identify the target speech during target presence.

Fig. 1: Fig. 1a depicts a simple acoustic situation with a target speaker in background noise but without interfering speakers. Fig. 1b extends the acoustic situation to include an interferer, in addition to the target speech and noise. In Fig. 1c, an own voice is conversing with the target speech in an acoustic environment with an interferer and noise. Fig. 1d shows the basic idea of the proposed noise reduction systems. During own voice presence, the interferer is identified and during target presence the target is identified.

directing beamformers to all possible directions, and selecting the beamformer with the highest output power. However, in many practical situations the target speech need not to be loudest, and systems based on this assumption will fail. Other methods rely on prior assumptions of the target location e.g. the methods presented in [23]–[25]. In HAD applications, the target location is often assumed frontal relative to the user [15,25,26]. This assumption is motivated by the observation that for face-to-face conversation, where the HAD-user uses eye-contact and lip-reading, the target source is often located in the frontal half-plane with respect to the user. However, also this assumption is not always valid, e.g., in situations, where the HAD user is unable to look at the target (e.g., when driving a car). Finally, other RATF vector estimation methods, e.g. [10,27,28], can perform well in simple situations where the target source is present in background noise, but where no interfering speakers are present, cf. Fig. 1a[1].

Unfortunately, in more complex acoustic situations, where one or more interfering speakers are simultaneously present, cf. Fig. 1b, estimation of the noise CPSD matrix and the target RATF vector can be difficult tasks. The presence of interferers can make it difficult to determine the target speaker, particularly when the interferer is voice-like. A voice-like interferer can make voice activity detection difficult as it is hard to distinguish between desired and interfering speech. This can result in interference and noise statistics being captured poorly and degrade the noise reduction performance significantly. Recent proposed methods can potentially help identifying the target speaker by decoding the direction of the user's auditory attention [30] or the user's eye-gaze direction [31,32] with the use of EEG signals or eye-trackers. However, these methods require the use of additional sensors which may not be available for the speech enhancement system. Other situations that can be particularly difficult for existing noise reduction systems to handle are conversations between the user and a target speaker. The situation is further complicated, if an interfering speaker is present during the conversation between the user and a target speaker, cf. Fig. 1c. The presence of the own voice signal will leave few instances of noise dominant TF tiles making the SPP-based methods ineffective.

In this paper, we propose a method which solves these problems by using the presence of the user's own voice signal as an asset. Specifically, we use the fact that the presence of own voice signal often indicates the absence of the target

[1]Speech signals used in the figures are from the speech database in [29]

signal due to the avoidance of speech overlap between the user and the target speaker [33]–[35]. Additionally, the absence of own voice may indicate the presence of a target signals.

The proposed method relies on the assumption that any sound source during own voice presence is of no interest to the user, and can hence be regarded as interfering signals. Therefore, statistics related to the interference and noise can be updated during own voice activity as shown in Fig. 1d. To demonstrate the idea, we consider the situation where only a single interfering speech source may be present. This problem is already very challenging to solve with state-of-the-art methods, as it is difficult to decide which of sound sources is target and which is the interfering speaker. However, the proposed method can in principle be extended to handle multiple interfering speakers such that any speaker during own voice presence is considered undesired. The acoustic situation, we specifically seek to solve in this paper, is the presence of a target speaker, the user's own voice, an interfering speaker, and noise. Such a situation can be regarded as particularly difficult to solve with the current state-of-the-art methods due to the interfering speaker and the very few instances, where noise dominates the noisy signal. As shown in Fig. 1d, the proposed systems identify the interferer during own voice, i.e. estimate the interferer RATF vector and use this estimate to support the implementation of a beamforming system during target speech presence. Specifically, the estimated interferer RATF vector from own voice presence is used during own voice absence (presumably target presence) to support the estimation of the interference-plus-noise CPSD matrix and target RATF vector. The estimated interference-plus-noise CPSD matrix and target RATF vector are then used in an MWF beamformer to suppress the interferer and noise.

The paper is structured as follows. In Sec. II, the signal model of the microphone signals is presented. In Sec. III, the MLEs of the interference and noise PSDs, and interference and target RATFs, are presented respectively. Sec. IV presents the simulation setup and evaluates the proposed noise reduction algorithm in simulation experiments. Finally, in Sec. V, a conclusion of the results is given.

II. Multi-Microphone Signal Model

We consider a HAD with M microphones placed in an arbitrary array geometry. The considered acoustic situation is depicted in Fig. 2. Each microphone picks up sound from the acoustic environment, and the signals are then sampled into a discrete-time sequence $x_m(n)$ for $m = 1, .., M$. The acoustic scene consists of an own voice signal, $s'_o(n)$, a target signal, $s'_t(n)$, interfering speech signal $s'_q(n)$, and noise denoted as $v(n)$. We assume, for simplicity, the presence of a single interferer per TF tile.

Let $h_{o,m}(n)$, $h_{t,m}(n)$, and $h_{q,m}(n)$ denote the acoustic impulse response (AIR) from the own voice, target, and interferer respectively to the m'th microphone. The signal model of the observed noisy signal is then

$$x_m(n) = \sum_{j\in\{t,o,q\}} s'_j(n) * h_{j,m}(n) + v_m(n), \tag{1}$$

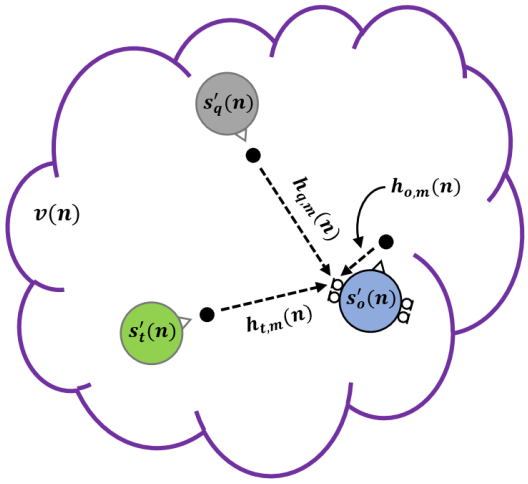


Fig. 2: Example of an acoustic scene with an own voice $s'_o(n)$, target $s'_t(n)$, interference $s'_q(n)$, and noise $v(n)$ where the microphones are mounted on the user's head. The acoustic impulse response from the j'th source ($j \in \{t, o, q\}$) to the m'th microphone is denoted as $h_{j,m}(n)$.

where $*$ denotes the linear convolution operator. The proposed noise reduction algorithm is derived and implemented in the TF domain using the short-time Fourier transform (STFT) with window function $\psi(n)$, window size N_{win}, and overlap N_{ov}. The STFT of the noisy signal is [1,36]

$$x_m(k,l) = \sum_{n=0}^{N_{\text{win}}-1} x_m(n + lN_{\text{ov}})\psi(n)e^{-2\pi ik\frac{n}{N_{\text{win}}}}, \tag{2}$$

where $i{=}\sqrt{-1}$, k and l denote the frequency bin and frame index, respectively. We define $\boldsymbol{x}(k,l) = [x_1(k,l), ..., x_M(k,l)]^T$ as an $M \times 1$ complex vector containing the noisy TF observations for all M microphones. In the TF domain, the signal model becomes

$$\boldsymbol{x}(k,l) = \sum_{j\in\{t,o,q\}} s'_j(k,l)\boldsymbol{h}_j(k,l) + \boldsymbol{v}(k,l), \tag{3}$$

where $\boldsymbol{h}_j(k,l)$ and $\boldsymbol{v}(k,l)$ are the stacked acoustic transfer functions (ATFs) and noise, respectively. We assume that the AIRs are shorter than the STFT analysis window $\psi(n)$ [37]. The signals $s'_j(k,l)$ for $j \in \{t, o, q\}$ denote the speech signals of the target, own voice, and interferer at their respective locations. Let m^* denote a pre-selected reference microphone, then we may normalize the ATFs with respect to the reference microphone such that

$$\boldsymbol{x}(k,l) = \sum_{j\in\{t,o,q\}} s_j(k,l)\boldsymbol{d}_j(k,l) + \boldsymbol{v}(k,l), \tag{4}$$

where

$$\boldsymbol{d}_j(k,l) = \left[\frac{h_{1,j}(k,l)}{h_{m^*,j}(k,l)}, ..., \frac{h_{M,j}(k,l)}{h_{m^*,j}(k,l)}\right]^T, \tag{5}$$

is the RATF vector [38] and $s_j(k,l)$ is the j'th signal as captured at the reference microphone. We assume that the presence of the own voice signal and the target signal are mutually

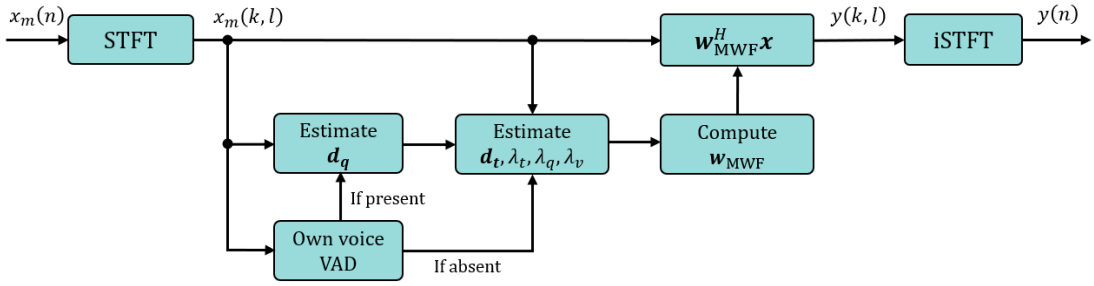


Fig. 3: Overview of the proposed noise reduction systems, where the interferer is identified during own voice presence, and the target is identified during own voice absence.

exclusive. This assumption is based on the conversational model in [34], where interlocutors in conversations avoid speech overlaps and pauses. This assumption is supported by results found in human experiments in [33,35,39]. These results suggest that interlocutors during conversations avoid speech overlap and pauses in noisy environment. Hence, the signal model may be divided to reflect two situations, namely, 1) when own voice is present and target is absent,

$$\boldsymbol{x}(k,l) = s_o(k,l)\boldsymbol{d}_o(k,l) + s_q(k,l)\boldsymbol{d}_q(k,l) + \boldsymbol{v}(k,l), \quad (6)$$

and 2) when the target is present, but own voice is absent

$$\boldsymbol{x}(k,l) = s_t(k,l)\boldsymbol{d}_t(k,l) + s_q(k,l)\boldsymbol{d}_q(k,l) + \boldsymbol{v}(k,l). \quad (7)$$

In the sequel, we omit the frequency bin and frame index, e.g. $\boldsymbol{x} \triangleq \boldsymbol{x}(k,l)$, for brevity.

A. Multichannel Wiener filter beamforming

The task of the beamformer is to retrieve the target speech s_t, while suppressing the interference and noise. The output of a linear beamformer is given by [3]

$$y = \boldsymbol{w}^H \boldsymbol{x}, \quad (8)$$

where $\boldsymbol{w}$ is the vector of beamformer weights. The multichannel Wiener filter (MWF) is the linear minimum mean square error (LMMSE) estimator of the target signal with beamformer weights $\boldsymbol{w}_{\text{MWF}}$ which are found by solving the following optimization problem [3]:

$$\boldsymbol{w}_{\text{MWF}} = \arg\min_{\boldsymbol{w}} \mathbb{E}\left[|s_t - \boldsymbol{w}^H \boldsymbol{x}|^2\right], \quad (9)$$

where H is the Hermitian transpose. Assuming that s_t, s_q and $\boldsymbol{v}$ are uncorrelated random variables, the MWF can be shown [3] to be dependent on the target RATF $\boldsymbol{d}_t$, the target power spectral density (PSD) $\lambda_t = \mathbb{E}\left[|s_t|^2\right]$, and the interference-plus-noise CPSD matrix, $\mathbf{C}_{qv}$. The interference-plus-noise CPSD matrix is defined to be

$$\begin{aligned} \mathbf{C}_{qv} &= \mathbb{E}\left[(s_q\boldsymbol{d}_q + \boldsymbol{v})(s_q\boldsymbol{d}_q + \boldsymbol{v})^H\right] \\ &= \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \mathbf{C}_v, \end{aligned} \quad (10)$$

where $\lambda_q = \mathbb{E}\left[|s_q|^2\right]$ is the interference PSD and $\mathbf{C}_v = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^H\right]$ is the noise CPSD matrix. Then the MWF beamformer can be expressed as [3]

$$\boldsymbol{w}_{\text{MWF}} = \frac{\mathbf{C}_{qv}^{-1}\boldsymbol{d}_t}{\boldsymbol{d}_t^H \mathbf{C}_{qv}^{-1}\boldsymbol{d}_t} \cdot \frac{\lambda_t}{\lambda_t + (\boldsymbol{d}_t^H \mathbf{C}_{qv}^{-1}\boldsymbol{d}_t)^{-1}}, \quad (11)$$

where the first factor is known as the minimum variance distortion-less response (MVDR) beamformer and the second factor is known as the single-channel post Wiener filter. We see that the MWF beamformer in the form of (11) requires $\boldsymbol{d}_t$, $\boldsymbol{d}_q$, λ_t, λ_q, and $\mathbf{C}_v$ to be known and in the following, we propose methods to estimate these parameters for each time-frequency tile by exploiting the own voice of the user. For simplification, we assume that the noise, $\boldsymbol{v}$, is a time-varying random process and that its CPSD matrix can be expressed as $\mathbf{C}_v = \lambda_v \mathbf{\Gamma}_v$. Here, $\mathbf{\Gamma}_v$ is a known noise CPSD matrix which is normalized with respect to the reference microphone and obtained from the most recent noise-only observation [14].

B. Target and interference identification

During own voice presence, we estimate $\boldsymbol{d}_q$. Following (6), the noisy CPSD matrix during own voice is modeled as

$$\mathbf{C}_x = \lambda_o \boldsymbol{d}_o \boldsymbol{d}_o^H + \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \lambda_v \mathbf{\Gamma}_v, \quad (12)$$

and likewise, during own voice absence, we assume that target is present cf. (7), such that the noisy CPSD matrix is modeled as

$$\mathbf{C}_x = \lambda_t \boldsymbol{d}_t \boldsymbol{d}_t^H + \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \lambda_v \mathbf{\Gamma}_v. \quad (13)$$

In applications such as HADs, the microphone array is commonly mounted in a fixed position on the user's head. Therefore, the acoustic transfer function, $\boldsymbol{d}_o$, from the user's mouth to the microphones can be considered approximately time-invariant. This allows offline estimation of the own voice RATF vector $\boldsymbol{d}_o$, which can be used during online deployment of the noise reduction algorithm. Additionally, the microphones are placed close to the user's mouth hence the own voice signal can be considerably louder than the target and interference speech signals, especially at lower frequencies, when the own voice is active [40, p. 251]. For these reasons, we consider the own voice RATF vector, $\boldsymbol{d}_o$, as known and assume that an own voice activity detector (OVAD) is available. The estimated $\boldsymbol{d}_q$ is then used during own voice absence (but target presence) to estimate the remaining parameters $\boldsymbol{d}_t$, λ_t, λ_q, and λ_v per TF tile and the resulting MWF beamformer can then be applied.

In practice, it may occur that the signal models in (12) or (13) are violated, for example due to speech overlap and gaps. A worst case example is speech overlap between the user

Algorithm 1: MWF beamformer with proposed target and interference identification.

Input: $\boldsymbol{d}_o$.
1: **if** own voice is present **then**
2: Estimate $\boldsymbol{d}_q$.
3: **else if** own voice is absent **then**
4: Estimate $\boldsymbol{d}_t$, λ_t, λ_q, and λ_v given $\boldsymbol{d}_q$.
5: Form the interference-plus-noise CPSD matrix $\mathbf{C}_{qv} = \lambda_q \boldsymbol{d}_q \boldsymbol{d}_q^H + \lambda_v \boldsymbol{\Gamma}_v$.
6: Compute the MWF beamformer weights $\boldsymbol{w}_{\text{MWF}}$ in (11).
7: Apply the beamformer $y = \boldsymbol{w}_{\text{MWF}}^H \boldsymbol{x}$.
8: **end if**

and the target speaker. Such situations can potentially lead to suppression of the target, as the target might be identified as the interfering speaker. One potential solution is to use several seconds of noisy observations during own voice presence. Since speech overlaps between the user and the target are often short and brief (e.g. 250 ms) [33,34,41], increasing the number of observations from own voice presence can reduce the likelihood of the target being identified as interference.

Furthermore, the use of own voice to identify an interfering speaker can be generalized to multiple interfering speakers. Specifically, any speakers that are present for minimum duration during own voice presence can be considered undesired. A potential procedure could for example involve a model-order selection algorithm e.g. minimum description length, Akaike or Bayesian information criterion to first determine the number of interfering speakers [42,43]. This is then followed by an estimation procedure of the RATF vectors for all the interfering speakers and finally estimation of the interference-plus-noise CPSD matrix.

The proposed noise reduction scheme is summarized in Fig. 3 and as pseudo-code in Algorithm 1.

III. Maximum Likelihood Estimation

In order to implement MWF beamformers for the considered acoustic situation, the parameters $\boldsymbol{d}_q$, $\boldsymbol{d}_t$, λ_t, λ_q, and λ_v must be estimated. In the following, we present several MLE-based schemes for estimation of the parameters of interest.

It is widely known that MLEs of the RATF vectors and PSDs perform well when used in a beamforming context e.g. in [10,14,18,22]. Comparative and theoretical performance of these estimators, e.g. in terms of Cramer-Rao bounds, have been derived and presented in [15,17,18,44]. Let us first note that the signal model in (12), where $\boldsymbol{d}_o$ and $\boldsymbol{\Gamma}_v$ are assumed known, and the signal model in (13) where $\boldsymbol{d}_q$ and $\boldsymbol{\Gamma}_v$ are assumed known, both can be written in the following general form

$$\mathbf{C} \triangleq \mathbf{C}(\lambda_1, \lambda_2, \phi, \boldsymbol{d}_1) = \lambda_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \phi \boldsymbol{\Gamma}. \tag{14}$$

Hence, finding estimates of the parameters of interest for both (12) and (13), corresponds to finding MLEs of λ_1, λ_2, ϕ, and $\boldsymbol{d}_1$ in (14). In particular, let $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\phi}$, and $\hat{\boldsymbol{d}}_1$ denote the MLEs of these parameters. Then for (12), we estimate the interference RATF vector as $\hat{\boldsymbol{d}}_q = \hat{\boldsymbol{d}}_1$, (the MLEs $\hat{\lambda}_q = \hat{\lambda}_1$, $\hat{\lambda}_o = \hat{\lambda}_2$, and $\hat{\lambda}_v = \hat{\phi}$ are nuisance parameters and, hence, not used in the subsequent steps). Similarly, when own voice is absent, the parameters of (13) are given by $\hat{\boldsymbol{d}}_t = \hat{\boldsymbol{d}}_1$, $\hat{\lambda}_t = \hat{\lambda}_1$, $\hat{\lambda}_q = \hat{\lambda}_2$, and $\hat{\lambda}_v = \hat{\phi}$.

We assume that the noisy observations, $\boldsymbol{x}$, are complex Gaussian distributed [10,16,18] such that the likelihood for N observations of $\boldsymbol{x}$, $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$, is

$$f(\boldsymbol{X}; \lambda_1, \lambda_2, \phi, \boldsymbol{d}_1) = \frac{\exp\left(-N \text{tr}(\mathbf{C}^{-1}\mathbf{R})\right)}{\pi^{NM} |\mathbf{C}|^N}, \tag{15}$$

where $\mathbf{R} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^H$, $|\cdot|$ is the determinant operator, and $\text{tr}(\cdot)$ denotes the trace operator. Furthermore, we assume that $\boldsymbol{d}_1$ is an element of a pre-defined dictionary $\mathcal{D} = \{\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, ..., \boldsymbol{d}^{(N_D)}\}$, where N_D is the dictionary size [10]. The MLEs can be found by solving the optimization problem

$$\arg \max_{\lambda_1, \lambda_2, \phi, \boldsymbol{d}_1 \in \mathcal{D}} \log f(\boldsymbol{X}; \lambda_1, \lambda_2, \phi, \boldsymbol{d}_1). \tag{16}$$

Closed-form solutions for this optimization problem seem not to exist [44]. Instead, we use a numerical approach to solve (16) in Sec. III-A which involves a two-dimensional search. In Secs. III-B and III-C, we adapt MLEs from [22] and [19] to estimate λ_1, λ_2, ϕ, and $\boldsymbol{d}_1$. The estimators in [22] and [19] are not strictly MLEs of the problem posed in (16). However, they are computationally much less expensive as they only involve a one-dimensional search and – as we show – to perform essentially on par with the true computational comples MLEs of (16) in terms of speech enhancement performance.

A. Joint ML using grid search

Let us rewrite equation (14) as

$$\begin{aligned} \mathbf{C} &= \lambda_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \phi \left(\frac{\lambda_2}{\phi} \boldsymbol{d}_2 \boldsymbol{d}_2^H + \boldsymbol{\Gamma} \right), \\ &= \lambda_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \phi \boldsymbol{\Phi}(\psi(\phi)), \end{aligned} \tag{17}$$

where $\psi(\phi) = \frac{\lambda_2}{\phi}$, and $\boldsymbol{\Phi}(\psi) \triangleq \psi \boldsymbol{d}_2 \boldsymbol{d}_2^H + \boldsymbol{\Gamma}$. For notational convenience, we define $\psi \triangleq \psi(\phi)$. For a given value of ψ, closed-form MLEs of λ_1 and ϕ exist, while conditioned on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ [15,19]. Hence, conditioned on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$, estimating the remaining parameters λ_1, ϕ, and ψ involves a one-dimensional search procedure over ψ or implicitly λ_2. In principle any numerical solver, e.g. a grid-search or gradient ascent method, can be used to solve the optimization problem. However, as a proof of concept, we use a grid-search based solver as the grid is only over ψ and $\boldsymbol{d}_1$. The grid-search procedure can be simplified to be over the dictionary, $\Psi = \{\psi^{(1)}, ..., \psi^{(N_\psi)}\}$, where N_ψ denotes the cardinality of Ψ. Obviously, for a sufficiently fine grid, Ψ, the proposed approach will return estimates arbitrarily close to the true MLE.

The first step in the procedure is to obtain an MLE for ϕ for a particular grid point, $\psi \in \Psi$, while conditioned on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$. To do so, we define the MVDR beamformer with distortion-less constraint on $\boldsymbol{d}_1$,

$$\boldsymbol{w}(\psi, \boldsymbol{d}_1) = \boldsymbol{\Phi}^{-1}(\psi) \boldsymbol{d}_1 \left(\boldsymbol{d}_1^H \boldsymbol{\Phi}^{-1}(\psi) \boldsymbol{d}_1 \right)^{-1}. \tag{18}$$

Furthermore, let

$$\mathbf{Q}(\psi, \boldsymbol{d}_1) = \mathbf{I} - \boldsymbol{d}_1 \boldsymbol{w}^H(\psi, \boldsymbol{d}_1), \tag{19}$$

where $\mathbf{I}$ is the identity matrix. The MLE of ϕ is then given by [19]

$$\hat{\phi}(\psi, \boldsymbol{d}_1) = \frac{1}{M-1} \text{tr}\left(\mathbf{Q}(\psi, \boldsymbol{d}_1)\mathbf{R}\mathbf{\Phi}^{-1}(\psi)\right), \tag{20}$$

where $\mathbf{R} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^H$ is the sample noisy CPSD matrix, and the MLE of λ_1 is [19]

$$\hat{\lambda}_1(\psi, \boldsymbol{d}_1) = \boldsymbol{w}(\psi, \boldsymbol{d}_1)^H \left(\mathbf{R} - \hat{\phi}(\psi, \boldsymbol{d}_1)\mathbf{\Phi}(\psi)\right) \boldsymbol{w}(\psi, \boldsymbol{d}_1). \tag{21}$$

The MLEs, $\hat{\phi}$ and $\hat{\lambda}_1$, are then used to concentrate the log-likelihood in (16), such that the optimization problem is reduced to

$$\hat{\psi}, \hat{\boldsymbol{d}}_1 = \underset{\psi \in \Psi, \boldsymbol{d}_1 \in \mathcal{D}}{\arg\max} \ \log f(\boldsymbol{X}, \hat{\phi}, \hat{\lambda}_1; \psi, \boldsymbol{d}_1). \tag{22}$$

Given the MLE $\hat{\psi}$, the MLE of λ_2 can be found as

$$\hat{\lambda}_2(\hat{\psi}, \hat{\boldsymbol{d}}_1) = \hat{\phi}(\hat{\psi}, \hat{\boldsymbol{d}}_1) \cdot \hat{\psi}. \tag{23}$$

The whole procedure is summarized in Algorithm 2.

Algorithm 2: Joint ML using grid search

Input: $\Psi = \{\psi^{(1)}, ..., \psi^{(N_\psi)}\}$, $\mathcal{D} = \{\boldsymbol{d}_1^{(1)}, ..., \boldsymbol{d}_1^{(N_D)}\}$, $\boldsymbol{d}_2$, $\mathbf{\Gamma}$.

for $i = 1, 2, ..., N_D$ **do**
  for $j = 1, 2, ..., N_\psi$ **do**
    Compute $\boldsymbol{w}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (18).
    Compute $\mathbf{Q}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (19).
    Estimate $\hat{\phi}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (20).
    Estimate $\hat{\lambda}_1(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (21).
    Evaluate $\log f(\boldsymbol{X}, \hat{\phi}, \hat{\lambda}_1; \psi^{(j)}, \boldsymbol{d}_1^{(i)})$ using (22) and (15).
    Compute $\hat{\lambda}_2(\psi^{(j)}, \boldsymbol{d}_1^{(i)}) = \psi^{(j)} \cdot \hat{\phi}(\psi^{(j)}, \boldsymbol{d}_1^{(i)})$.
  end for
end for
Find i^* and j^* that maximize $\log f(\boldsymbol{X}, \hat{\phi}, \hat{\lambda}_1; \psi^{(j)}, \boldsymbol{d}_1^{(i)})$.
The joint MLEs are then $\hat{\phi}(\psi^{(j^*)}, \boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_1(\psi^{(j^*)}, \boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_2(\psi^{(j^*)}, \boldsymbol{d}_1^{(i^*)})$, and $\hat{\boldsymbol{d}}_1 := \boldsymbol{d}_1^{(i^*)}$.

B. ML in the blocked domain

As an alternative to the joint ML method, which requires a two-dimensional dictionary search, we propose in the following a simpler ML estimation procedure in the blocked domain [16,44]. Specifically, the MLEs are not guaranteed to be ML optimal for the problem posed in (16), but have been demonstrated to perform well in terms of beamforming performance in [22]. The ML estimation of the parameters in (14), i.e. λ_1, λ_2, ϕ, and $\boldsymbol{d}_1$, in the blocked domain is adapted from [22] and consists of two stages. The first stage is ML estimation of λ_1 and ϕ conditioned on $\boldsymbol{d}_1$ in the blocked domain of $\boldsymbol{d}_2\boldsymbol{d}_2^H$, i.e. the null-space of $\boldsymbol{d}_2\boldsymbol{d}_2^H$. The second stage is ML estimation of λ_2 where the MLEs of λ_1 and ϕ conditioned on $\boldsymbol{d}_1$ are used to concentrate the log-likelihood in (16). The rationale behind this ML estimation in the blocked domain, is to simplify the estimation problem by canceling one of the speech components with a blocking matrix $\mathbf{B}$. Specifically, the speech components λ_2 and $\boldsymbol{d}_2$ are eliminated in the first stage by projecting $\boldsymbol{x}$ to the null-space of $\boldsymbol{d}_2\boldsymbol{d}_2^H$. In the second stage, only λ_1, ϕ, and $\boldsymbol{d}_1$ remain and are estimated using the MLEs in [19].

1) ML estimation of λ_1 and ϕ: To map the noisy observations into the blocked domain, we form a blocking matrix, which cancels the $\lambda_2\boldsymbol{d}_2\boldsymbol{d}_2^H$ term from (14). The blocking matrix, $\mathbf{B}$, is given as [22]

$$\mathbf{B} = \left(\mathbf{I}_{M\times M} - \frac{\boldsymbol{d}_2\boldsymbol{d}_2^H}{\boldsymbol{d}_2^H\boldsymbol{d}_2}\right)\mathbf{I}_{M\times M-1}. \tag{24}$$

where $\mathbf{I}_{M\times M}$ is an $M \times M$ identity matrix and $\mathbf{I}_{M\times M-1}$ is the first $M-1$ column vectors of $\mathbf{I}_{M\times M}$. Applying the blocking matrix to the input vector $\mathbf{B}^H\boldsymbol{x}$, the CPSD matrix in the blocked domain is

$$\tilde{\mathbf{C}} = \mathbf{B}^H\mathbf{C}\mathbf{B} = \tilde{\lambda}_1\tilde{\boldsymbol{d}}_1\tilde{\boldsymbol{d}}_1^H + \tilde{\phi}\tilde{\mathbf{\Gamma}}, \tag{25}$$

where $\mathbf{C}$ is the CPSD matrix from (14), $\tilde{\mathbf{\Gamma}} = \mathbf{B}^H\mathbf{\Gamma}\mathbf{B}$, and $\tilde{\boldsymbol{d}}_1 = \mathbf{B}^H\boldsymbol{d}_1$. The parameters to estimate in (25) are the blocked domain PSDs $\tilde{\lambda}_1$ and $\tilde{\phi}$, and the RATF vector $\tilde{\boldsymbol{d}}_1$. The CPSD matrix in (25) has a form that is identical to the CPSD matrix in (17). Therefore, estimating $\tilde{\lambda}_1$, $\tilde{\phi}$, and $\tilde{\boldsymbol{d}}_1$ follows a similar procedure as found in Sec. III-A. In the first stage, the likelihood function in the blocked domain is

$$f(\tilde{\boldsymbol{X}}; \tilde{\lambda}_1, \tilde{\phi}|\tilde{\boldsymbol{d}}_1) = \frac{\exp\left(-N\text{tr}(\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{R}})\right)}{\pi^{NM}|\tilde{\mathbf{C}}|^N}, \tag{26}$$

while conditioned on $\tilde{\boldsymbol{d}}_1$, and $\tilde{\boldsymbol{X}} = \mathbf{B}^H\boldsymbol{X}$ and $\tilde{\mathbf{R}} = \mathbf{B}^H\mathbf{R}\mathbf{B}$. The optimization problem is

$$\underset{\tilde{\lambda}_1, \tilde{\phi}}{\arg\max} \ \log f(\tilde{\boldsymbol{X}}; \tilde{\lambda}_1, \tilde{\phi}|\tilde{\boldsymbol{d}}_1). \tag{27}$$

In the following, the MLEs of $\tilde{\lambda}_1$ and $\tilde{\phi}$ are adaptations of the MLEs derived in [19]. The ML estimate of $\tilde{\phi}$ can be shown to be a function of an MVDR beamformer in the blocked domain with a distortion-less constraint on $\tilde{\boldsymbol{d}}_1$ [19] i.e.

$$\tilde{\boldsymbol{w}}_1(\tilde{\boldsymbol{d}}_1) = \tilde{\mathbf{\Gamma}}^{-1}\tilde{\boldsymbol{d}}_1\left(\tilde{\boldsymbol{d}}_1{}^H\tilde{\mathbf{\Gamma}}^{-1}\tilde{\boldsymbol{d}}_1\right)^{-1}, \tag{28}$$

and

$$\tilde{\mathbf{Q}}_1(\tilde{\boldsymbol{d}}_1) = \mathbf{I}_{M-1\times M-1} - \tilde{\boldsymbol{d}}_1\tilde{\boldsymbol{w}}_1^H(\tilde{\boldsymbol{d}}_1), \tag{29}$$

where $\mathbf{I}_{M-1\times M-1}$ is an $M-1 \times M-1$ identity matrix. The MLE of ϕ in the blocked domain is [19]

$$\hat{\tilde{\phi}}(\tilde{\boldsymbol{d}}_1) = \frac{1}{M-2}\text{tr}\left(\tilde{\mathbf{Q}}_1(\tilde{\boldsymbol{d}}_1)\tilde{\mathbf{R}}\tilde{\mathbf{\Gamma}}^{-1}\right), \tag{30}$$

where $\tilde{\mathbf{R}} = \mathbf{B}^H\mathbf{R}\mathbf{B}$ and the MLE of λ_1 is [19,22]

$$\hat{\tilde{\lambda}}_1(\tilde{\boldsymbol{d}}_1) = \tilde{\boldsymbol{w}}_1^H(\tilde{\boldsymbol{d}}_1)\left(\tilde{\mathbf{R}} - \hat{\tilde{\phi}}(\tilde{\boldsymbol{d}}_1)\tilde{\mathbf{\Gamma}}\right)\tilde{\boldsymbol{w}}_1(\tilde{\boldsymbol{d}}_1). \tag{31}$$

Algorithm 3: ML in the blocked domain

Input: $\mathcal{D} = \{\boldsymbol{d}_1^{(1)}, ..., \boldsymbol{d}_1^{(N_D)}\}$, $\boldsymbol{d}_2$, $\boldsymbol{\Gamma}$.
1: Obtain the blocking matrix $\mathbf{B}$ from (24)
2: Compute blocked domain $\boldsymbol{\Gamma}$ as $\tilde{\boldsymbol{\Gamma}} = \mathbf{B}^H \boldsymbol{\Gamma} \mathbf{B}$
3: **for** $i = 1, 2, ..., N_D$ **do**
4: $\quad \tilde{\boldsymbol{d}}_1^{(i)} = \mathbf{B}^H \boldsymbol{d}_1^{(i)}$
5: $\quad$ Compute $\tilde{\boldsymbol{w}}_1(\tilde{\boldsymbol{d}}_1^{(i)})$ in (28).
6: $\quad$ Compute $\tilde{\mathbf{Q}}_1(\tilde{\boldsymbol{d}}_1^{(i)})$ in (29).
7: $\quad$ Estimate $\hat{\phi}(\tilde{\boldsymbol{d}}_1^{(i)})$ using (30).
8: $\quad$ Estimate $\hat{\lambda}_1(\tilde{\boldsymbol{d}}_1^{(i)})$ using (31).
9: $\quad$ Set $\hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1^{(i)}) := \hat{\lambda}_1 \boldsymbol{d}_1^{(i)} (\boldsymbol{d}_1^{(i)})^H + \hat{\phi} \boldsymbol{\Gamma}$
10: $\quad$ Compute $\boldsymbol{w}_2(\boldsymbol{d}_1^{(i)})$ in (35).
11: $\quad$ Compute $\mathbf{Q}_2(\boldsymbol{d}_1^{(i)})$ in (36).
12: $\quad$ Estimate $\hat{\gamma}(\boldsymbol{d}_1^{(i)})$ using (37).
13: $\quad$ Estimate $\hat{\lambda}_2(\boldsymbol{d}_1^{(i)})$ using (38).
14: $\quad$ Evaluate $\log f(\boldsymbol{X}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi}; \boldsymbol{d}_1^{(i)})$ in (41) using (15)
15: **end for**
16: Find $i^* = \arg\max_i \log f(\boldsymbol{X}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi}; \boldsymbol{d}_1^{(i)})$
17: The joint MLEs are then $\hat{\phi}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_1(\boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_2(\boldsymbol{d}_1^{(i^*)})$, $\hat{\boldsymbol{d}}_1 := \boldsymbol{d}_1^{(i^*)}$.

2) ML estimation of λ_2: Given $\hat{\tilde{\lambda}}_1(\tilde{\boldsymbol{d}}_1)$ and $\hat{\tilde{\phi}}(\tilde{\boldsymbol{d}}_1)$, these may be inserted into the noisy CPSD matrix in (14) such that it becomes

$$\begin{aligned} \mathbf{C}(\lambda_2, \boldsymbol{d}_1) &= \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \left(\hat{\tilde{\lambda}}_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \hat{\tilde{\phi}} \boldsymbol{\Gamma} \right) \\ &= \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1), \end{aligned} \tag{32}$$

where $\hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1) = \hat{\tilde{\lambda}}_1 \boldsymbol{d}_1 \boldsymbol{d}_1^H + \hat{\tilde{\phi}} \boldsymbol{\Gamma}$. For ML estimation of the remaining parameter, λ_2, we introduce the parameter γ such that the noisy CPSD matrix is

$$\mathbf{C}(\lambda_2, \boldsymbol{d}_1, \gamma) = \lambda_2 \boldsymbol{d}_2 \boldsymbol{d}_2^H + \gamma \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1), \tag{33}$$

which ensures that (33) has a form identical to (25), and, hence, the MLEs of γ and λ_2 can be found similarly. The optimization problem is

$$\arg\max_{\lambda_2, \gamma} \log f(\boldsymbol{X}; \lambda_2, \gamma | \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1)), \tag{34}$$

where the likelihood function is conditioned on $\hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1)$, and has the form as in (15). To estimate λ_2 and γ, first we form the MVDR beamformer with distortion-less constraint on $\boldsymbol{d}_2$

$$\boldsymbol{w}_2(\boldsymbol{d}_1) = \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1)^{-1} \boldsymbol{d}_2 \left(\boldsymbol{d}_2{}^H \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1)^{-1} \boldsymbol{d}_2 \right)^{-1}, \tag{35}$$

such that

$$\mathbf{Q}_2(\boldsymbol{d}_1) = \mathbf{I}_{M \times M} - \boldsymbol{d}_2 \boldsymbol{w}_2^H(\boldsymbol{d}_1). \tag{36}$$

Then the MLE of γ is [19]

$$\hat{\gamma}(\boldsymbol{d}_1) = \frac{1}{M-1} \text{tr} \left(\mathbf{Q}_2(\boldsymbol{d}_1) \mathbf{R} \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1)^{-1} \right), \tag{37}$$

and the MLE of λ_2 is [19]

$$\hat{\lambda}_2(\boldsymbol{d}_1) = \boldsymbol{w}_2^H(\boldsymbol{d}_1) \left(\mathbf{R} - \hat{\gamma}(\boldsymbol{d}_1) \hat{\boldsymbol{\Phi}}(\boldsymbol{d}_1) \right) \boldsymbol{w}_2(\boldsymbol{d}_1). \tag{38}$$

The introduction of the variable $\hat{\gamma}(\boldsymbol{d}_1)$, means that the MLE of λ_1 and ϕ becomes

$$\hat{\lambda}_1(\boldsymbol{d}_1) = \hat{\tilde{\lambda}}_1(\boldsymbol{d}_1) \cdot \hat{\gamma}(\boldsymbol{d}_1), \tag{39}$$

and

$$\hat{\phi}(\boldsymbol{d}_1) = \hat{\tilde{\phi}}(\boldsymbol{d}_1) \cdot \hat{\gamma}(\boldsymbol{d}_1). \tag{40}$$

Finally, the MLE of $\boldsymbol{d}_1$ is found by evaluating the log-likelihood for each $\boldsymbol{d}_1 \in \mathcal{D}$, and choose the one that maximizes the log-likelihood i.e.

$$\hat{\boldsymbol{d}}_1 = \arg\max_{\boldsymbol{d}_1 \in \mathcal{D}} \log f(\boldsymbol{X}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi}; \boldsymbol{d}_1). \tag{41}$$

The ML procedure in the blocked domain is summarized in Algorithm 3.

C. Unconstrained Joint ML

Let $\mathbf{D} \triangleq [\boldsymbol{d}_1 \, \boldsymbol{d}_2]$ and $\boldsymbol{\Lambda}(\lambda_1, \lambda_2) = \text{diag}(\lambda_1, \lambda_2)$. Then the CPSD matrix in (14) can be written as

$$\mathbf{C}(\boldsymbol{d}_1, \boldsymbol{\Lambda}) = \mathbf{D} \boldsymbol{\Lambda} \mathbf{D}^H + \phi \boldsymbol{\Gamma}. \tag{42}$$

Note that the signal model is only identical to the one in (14) if $\boldsymbol{\Lambda}$ is a diagonal matrix. For known matrices $\mathbf{D}$ and $\boldsymbol{\Gamma}$, MLEs of $\boldsymbol{\Lambda}$ and ϕ were derived in [19]. However, the MLEs presented in [19] do not guarantee that the estimate of $\boldsymbol{\Lambda}$ is a diagonal matrix. The MLEs in [19], therefore, are not necessary maximum likelihood for the problem posed in (16). Nevertheless, as demonstrated in the simulation experiment in Appendix A, using the diagonal elements of the MLE of [19] works essentially as good as the joint ML method from Sec. III-A.

In [19], the MLE for ϕ is derived by first defining the linearly constrained minimum variance (LCMV) beamformer with distortion-less constraints on $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$,

$$\mathbf{W}(\boldsymbol{d}_1) = \boldsymbol{\Gamma}^{-1} \mathbf{D} \left(\mathbf{D}^H \boldsymbol{\Gamma}^{-1} \mathbf{D} \right)^{-1}, \tag{43}$$

where $\mathbf{W}(\boldsymbol{d}_1) \in \mathbb{C}^{M \times 2}$ and

$$\mathbf{Q}(\boldsymbol{d}_1) = \mathbf{I} - \mathbf{D} \mathbf{W}^H(\boldsymbol{d}_1). \tag{44}$$

Then the MLE of the noise PSD is given as [19]

$$\hat{\phi}(\boldsymbol{d}_1) = \frac{1}{M-2} \text{tr} \left(\mathbf{Q}(\boldsymbol{d}_1) \mathbf{R} \boldsymbol{\Gamma}^{-1} \right), \tag{45}$$

and the ML estimate of $\boldsymbol{\Lambda}$ is then [19]

$$\hat{\boldsymbol{\Lambda}}(\boldsymbol{d}_1) = \mathbf{W}^H(\boldsymbol{d}_1) \left(\mathbf{R} - \hat{\phi}(\boldsymbol{d}_1) \boldsymbol{\Gamma} \right) \mathbf{W}(\boldsymbol{d}_1). \tag{46}$$

We propose to find, the estimates of λ_1 and λ_2 as the main diagonal of $\hat{\boldsymbol{\Lambda}}(\boldsymbol{d}_1)$, i.e. $\hat{\lambda}_1 = \hat{\Lambda}_{1,1}$ and $\hat{\lambda}_2 = \hat{\Lambda}_{2,2}$. Finally, in order to estimate $\boldsymbol{d}_1$, we concentrate the log-likelihood with the MLEs of $\boldsymbol{\Lambda}(\boldsymbol{d}_1)$ and $\phi(\boldsymbol{d}_1)$, and search over the dictionary $\mathcal{D}$ until the element that returns the highest log-likelihood is found i.e.

$$\hat{\boldsymbol{d}}_1 = \arg\max_{\boldsymbol{d}_1 \in \mathcal{D}} \log f(\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \hat{\phi}; \boldsymbol{d}_1). \tag{47}$$

The unconstrained ML procedure is summarized in Algorithm 4. We have compared the three proposed algorithms in terms of speech enhancement performance in Appendix A. Our experiments demonstrate that the proposed algorithms essentially perform on par in terms of ESTOI and PESQ. 3 and Algorithm 4 perform marginally better than Algorithm 2 in terms of PESQ score, however slightly worse in terms of ESTOI score. This is possibly due to a slightly more aggressive noise reduction for Algorithm 3 and Algorithm 4 than Algorithm 2. This leads to higher PESQ scores but at the cost of more speech distortion and lower ESTOI scores. For this reason, we choose to leave Algorithm 2 out of the evaluation in Sec. IV-D, although Algorithm 3 and 4 do not solve the initial problem posed in (16). However, since Algorithm 2 requires a two-dimensional search, Algorithm 2 is likely much more computationally complex compared to Algorithm 3 and 4. The experiments in Sec. IV-D furthermore reveal that the real-time factor of Algorithm 4 is 9.46, Algorithm 3 is 4.98, and Algorithm 2 is 153.46. Although, we did not perform code optimization, the real-time factors give an indication of the computation complexity of the proposed methods and favors Algorithm 4 and 3 over Algorithm 2.

Algorithm 4: Unconstrained joint ML

Input: $\mathcal{D} = \{\boldsymbol{d}_1^{(1)}, ..., \boldsymbol{d}_1^{(N_D)}\}$, $\boldsymbol{d}_2$, $\boldsymbol{\Gamma}$.

1: **for** $i = 1, 2, ..., N_D$ **do**
2: Define $\mathbf{D}_i \triangleq [\boldsymbol{d}_1^{(i)}\ \boldsymbol{d}_2]$
3: Compute $\mathbf{W}(\boldsymbol{d}_1^{(i)})$ in (43).
4: Compute $\mathbf{Q}(\boldsymbol{d}_1^{(i)})$ in (44).
5: Estimate $\hat{\phi}(\boldsymbol{d}_1^{(i)})$ using (45).
6: Estimate $\hat{\boldsymbol{\Lambda}}(\boldsymbol{d}_1^{(i)})$ using (46).
7: Evaluate $\log f(\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \hat{\phi}; \boldsymbol{d}_1^{(i)})$ in (47)
8: **end for**
9: Find $i^* = \arg\max_i \ \log f(\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \hat{\phi}; \boldsymbol{d}_1^{(i)})$
10: The joint MLEs are then $\hat{\lambda}_1 = \hat{\Lambda}_{1,1}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\phi}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\lambda}_2 = \hat{\Lambda}_{2,2}(\boldsymbol{d}_1^{(i^*)})$, $\hat{\boldsymbol{d}}_1 := \boldsymbol{d}_1^{(i^*)}$.

D. Robust wideband estimation of the RATF vector

The proposed MLEs estimate the interferer and target RATF vectors independently over frequency bins. This approach allow multiple target and interferers in the acoustic scene, as long as a maximum of one target and interferer is present for a given TF tile. However, for acoustic sound sources, it is plausible to assume that the location of the target is identical across frequency. Therefore, in order to improve performance, estimation of the RATF vector can be done jointly over frequency hence made more robust [10,25]. The joint MLE of $\boldsymbol{d}_1$ is

$$\arg\max_{i\in\{1,2,...,|\mathcal{D}|\}} \sum_{k=1}^{K} \log f(\boldsymbol{X}(k), \hat{\lambda}_1(k), \hat{\lambda}_2(k), \hat{\phi}(k); \boldsymbol{d}_1^{(i)}(k)), \tag{48}$$

where $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$. Hence, the concentrated log-likelihood for a particular dictionary index i is added across frequency, where i corresponds to a location of the sound source.

IV. Performance Evaluation of Proposed Beamforming Systems

The proposed MLEs in Sec. III are evaluated in terms of beamforming performance when implemented into noise reduction systems. The beamforming performance of the proposed systems is found through simulation experiments where the task is to retrieve a target speech signal contaminated with interfering speech and noise. We compare the proposed methods against state-of-the-art methods which solve similar problems but do not explicitly model the presence of the own voice and interferer. The parameter estimation of the PSDs and RATF vectors used in the proposed noise reduction systems are based on Algorithm 3 and Algorithm 4 in Sec. III and used in an MWF beamformer, Algorithm 1, as shown in Fig. 3. We refer to the noise reduction systems based on Algorithm 3 as ML-BD and Algorithm 4 as UML.

A. Acoustic impulse response and sound databases

1) Acoustic impulse response database: Acoustic impulse functions (AIRs) are used to simulate the sound waves propagating from sound sources to the HA microphones. The AIRs were measured in an acoustic setup consisting of a circular loudspeaker array with a radius of 1.9 meters placed in an acoustically damped room [45]. A human HA user was seated in the center of the array during the measurements wearing two behind-the-ear (BTE) HAs; one placed on each ear. Each HA has a front and rear microphone separated by 1.3 cm. The AIRs mostly depend on the head and torso acoustics while reverberation has been removed by truncating the AIRs.

All M=4 microphones are used in a binaural HA configuration for the simulations. A direct implementation - as used in our simulations - of the MWF beamformers for a binaural HA configuration will result in a "noise collapse" [46]. In other words, all noise sources will sound as if they were originating from the target location. This is obviously important for a binaural HA application. However, several methods have been developed to mask or avoid this unwanted perceptual effect, e.g., [46]–[48]. Such methods are outside the scope of the present paper.

We assume instantaneous and error-free signal exchange between the left and right HAs. The AIRs were sampled at a horizontal resolution of 7.5 degrees with 0° defined as the frontal direction from the HA user's point of view, and the azimuth is counterclockwise rotating. Hence, the dictionary of AIRs contains AIRs from 48 different directions. The own voice AIRs were measured using a mouth reference microphone placed in front of the HA user's mouth. The HA user was asked to read a text up loud, and the AIRs from the own voice reference point to the HA microphones were measured [45].

In Sec. IV-E, AIR mismatches are simulated by using two different sets of AIR dictionaries measured on two different

human heads. One dictionary is used to simulate the acoustic scene, while the other is used as a dictionary in the noise reduction systems.

The RATF dictionary used for the proposed algorithms is obtained by transforming the AIR dictionary using (5). The frontal microphone of the left ear HA is used as the reference microphone.

2) Speech and noise databases: Speech signals used for the own voice, target, and interference, are obtained from the TIMIT database [49]. Speech pauses are removed with an energy-based VAD to minimize the influence of speech gaps in the evaluation. We do not simulate speech gaps caused by conversation pauses. However, the acoustic scene still include situations where neither the own voice nor the target speech are present in a TF tile due to speech being sparse in the TF domain. Hence, there are TF-tiles where own voice or target speech is absent even if they are detected present.

The noise database used in the simulation is recordings of noise found in realistic acoustic environments (e.g. a busy canteen and car cabin). The recordings of the noise are made with a spherical microphone array to accurately capture the noise field as measured at a reference point of the spherical microphone array. The captured noise is then transformed and convolved with the AIRs, such that the resulting noise field at the HA microphones in the simulation is identical to the one measured with the spherical microphone array [50].

B. Simulation of acoustic scenes

1) Target and noise levels: We define the input signal-to-interference-plus-noise ratio (SINR) as the ratio between the average target speech power and the average interference-plus-noise power. The target speech and interference-plus-noise power are computed prior to convolving the signals with the AIRs. The interference-to-noise ratio (INR) is defined similarly as the ratio between the average interfering speech power and the average noise power prior to convolving with the AIRs. The own voice and target speech are set to have equal power prior to convolving with the AIRs.

2) Target and interferer locations: The target RATF vector is randomly drawn from the dictionary of RATF vectors. Each RATF vector is associated with a direction and the RATF vectors are drawn from a uniform distribution where the set of possible outcomes is {-90°, -82.5°, ..., 90°}. Hence, the target is located in the frontal half-plane as the HA user in realistic situations is likely to be facing the target speaker [15,26]. The RATF vector for the interfering speech is randomly selected to be from the directions 75° or 225° and with this choice, the target speech and interfering speech are allowed to overlap in direction, when both the target speaker and interfering speaker are arriving from 75°.

3) Simulation settings: The sampling frequency of the simulation is 16 kHz. We used (1) to simulate the noisy microphone signals. The STFT and inverse STFT are used to transform the microphone signals into the time-frequency domain. A square-root Hanning window with a window size of 256 samples is used as analysis and synthesis windows. The window overlap is 128 samples. All algorithms in the evaluation have access to an oracle generic VAD that is able to perfectly detect regions with speech absence (i.e. frames with neither own voice, target, nor interfering speech). Since the generic VAD does not require to distinguish between own voice, target, nor interfering speech, this significantly simplifies the task of designing a robust VAD. The generic VAD is used to initialize Γ_v from noise-only region before any speech activity. Furthermore, an oracle OVAD is used in the evaluation for the proposed algorithms. The OVAD can detect the presence of own voice per frame but not per TF-tile. When own voice is detected absent, the proposed algorithms assume the presence of target speech. The duration of an acoustic scene is 5 seconds and Γ_v is initialized in a no-speech region before the beginning of the acoustic scene. The own voice is active in the first 2.5 seconds, followed by 2.5 seconds of own voice absence where the target is active to simulate a conversation. The interfering speaker is active during the whole 5 second simulation. Each reported performance score is an average over 40 acoustic scenes.

The HA user may occasionally rotate the head during conversations [33]. However, such head rotation were not implemented in our simulations. In practice, one might use other sensors e.g. accelerometers on board the HAD to detect or measure head rotations. After such detection, the noise reduction system may then compensate for the head-rotations or resort to a simpler baseline algorithm such as the one presented in [14] to increase robustness. Moreover, the target and interferer locations are fixed during the simulations. The proposed algorithms can in principle handle situations with moving targets, since the target RATF vectors are estimated for each TF-tile independently. Similarly, the proposed algorithms can handle moving interferers, but only during own voice regions. Moving interferers during own voice absence can potentially cause issues, but robustness against such situations can be increased by hypothesis testing. Specifically, if the noisy observations during own voice absence poorly match the interference-plus-noise CPSD matrix estimated from own voice presence (due to a moving interferer), hypothesis testing can help detecting these and resort to a simpler signal model to increase robustness e.g. (49).

A summary of the different acoustic settings for the experiments is given in Table I with references to the figures where the results are reported.

Number of mics	4	4	4
Noise type	Canteen	Car noise	Canteen
AIR mismatch	No	No	Yes
Figure	Fig. 5	Fig. 6	Fig. 7

TABLE I: Simulation settings used in the evaluation.

C. Baseline noise reduction systems

We compare the proposed system variants to recent state-of-the-art methods used for beamforming in HADs. These methods solve the problem of enhancing a single-target in noise using an MWF beamformer [10,14]. More advanced techniques presented in [23,51] can handle multiple speakers but require additional information about the target location or

target speech activity. We do not assume that the noise reduction systems have access to such information and therefore these methods were not included in the evaluation. The state-of-the-art methods we have included in the evaluation are:

1) MWF beamformer with ML PSD estimation assuming frontal target: In the context of HADs, the target speaker is often assumed to be frontal with respect to the HA user [15, 26]. The MWF beamforming scheme presented in [14,15] is used as a baseline method for MWF beamformers that assume frontal targets. For this particular method, the noisy CPSD matrix is modeled as

$$\mathbf{C}_x = \lambda_t \boldsymbol{d}_t \boldsymbol{d}_t^H + \lambda_v \mathbf{\Gamma}_v, \tag{49}$$

where $\boldsymbol{d}_t$ is the RATF vector associated with the frontal direction. The PSDs λ_t and λ_v are replaced by ML estimates and used to implement an MWF beamformer as in Fig. 3 during the remaining 2.5 seconds of an acoustic scene with target presence. The method is referred to as ML-FRONTAL.

2) MWF beamformer with ML PSD and target RATF estimation: The method proposed in [10] generalizes the method in [14,15] by including ML estimation of the target RATF vector. The noisy CPSD matrix is modeled as in (49), but the frontal RATF vector $\boldsymbol{d}_t$ is replaced by an estimated RATF vector. The log-likelihood function is denoted as $\log f(\boldsymbol{X}; \lambda_t, \lambda_v, \boldsymbol{d}_t)$ and is parameterized by λ_t, λ_v, and $\boldsymbol{d}_t$. These parameters are estimated by solving:

$$\underset{\lambda_t, \lambda_v, \boldsymbol{d}_t \in \mathcal{D}}{\arg\max} \ \log f(\boldsymbol{X}; \lambda_t, \lambda_v, \boldsymbol{d}_t). \tag{50}$$

The estimated λ_t, λ_v, $\boldsymbol{d}_t$ are similarly used to implement an MWF beamformer. We refer to this method to as ML-DOA. The baseline methods have access to a generic VAD to detect speech absence where neither own voice, interfering, nor target speech are present. The generic VAD is used to initialize Γ_v. In contrast to the propose methods, the baseline methods do not exploit the own voice to assume the target absence during own voice presence.

As a reference for upper bound performance, an "oracle" MWF beamformer, where all parameters for the MWF beamformer are known, is included.

D. Simulation results for canteen and car noise

Beamforming performance is evaluated in terms of estimated speech intelligibility using ESTOI [52] and in terms of speech quality using PESQ [53]. Performance is reported as a function of SINR to compare the robustness towards different noise level and as a function of INR to compare the robustness against the presence of interfering speech. When evaluating the performance as a function of INR, the SINR is fixed to 0 dB to simulate a reasonable noisy acoustic scene. Similarly, when evaluating the performance as a function of SINR, the INR is chosen to be fixed at 6 dB to maintain the presence of a fairly strong interfering speaker.

The beamforming performance in canteen noise is shown in Fig. 5 and performance for car noise is shown in Fig. 6. By visual inspection, we see that the proposed methods i.e. ML-BD and UML perform well in the presence of an interfering speaker. At high INRs, the proposed methods outperform both ML-FRONTAL and ML-DOA significantly. This observation indicates that the proposed methods are more efficient at identifying and suppressing interfering speech due to the use of the user's own voice. Sample spectrograms of the beamformer outputs are also shown in Fig. 4 for a visual comparison between ML-FRONTAL and UML. We only show the baseline method ML-FRONTAL as the target is located at 0° which is the best case scenario for ML-FRONTAL. ML-BD is also omitted from Fig. 4 as it shows very similar patterns to UML. The spectrograms show that the proposed algorithms suppress the interfering speaker more efficiently than the baseline methods while preserving the target speech. This can be seen in Fig. 4e and Fig. 4f where a comparison reveals that the interfering speaker is almost completely canceled using UML in contrast to ML-FRONTAL.

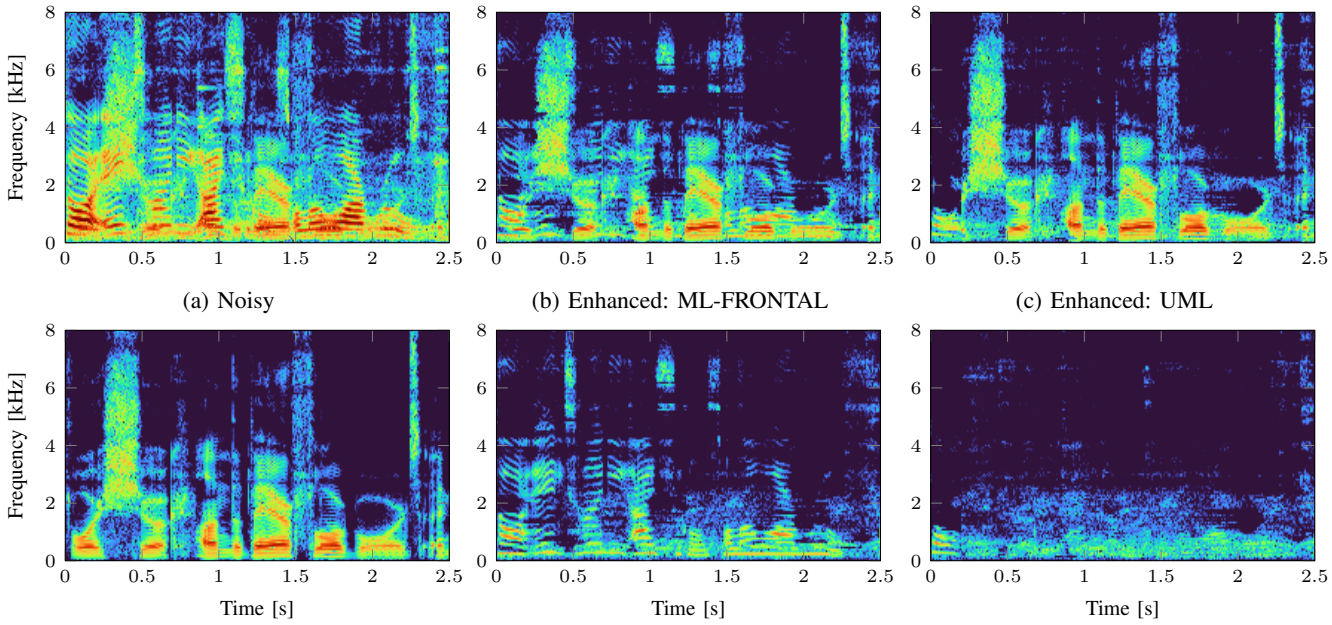
Another notable observation in Fig. 5 and Fig. 6 is that, the ML-DOA method return a very poor ESTOI and PESQ score when the INR is high. This is due to large amounts of target speech distortion as a consequence of the interfering speech mistakenly being identified as the target speech. In severe situations, e.g. when the INR is 12 dB, the performance of ML-DOA approaches the performance of the noisy signal.

ESTOI and PESQ scores of the proposed methods and the state-of-the-art methods, ML-DOA, are close at low INRs (see left panels in Fig. 5 and Fig. 6). To analyze these performance differences, we conduct pairwise t-tests [54] with Bonferroni corrected significance levels. The null-hypothesis is that the mean ESTOI score between two selected methods is identical for a given INR. We choose a significance level of $\alpha = 0.05$ before Bonferroni correction.

For canteen noise in Fig. 5, we compare ML-DOA with ML-BD and UML. The pairwise t-tests reveal no significant difference at -12 dB and -6 dB INR for ESTOI and -12 dB dB INR for PESQ. For car noise in Fig. 6, no significant difference is observed at -12 dB INR for ESTOI when comparing ML-DOA with ML-BD and UML. In terms of PESQ, the comparisons reveal that all pairwise comparisons for ML-DOA with ML-BD and UML are significant. The results for car noise, suggest that the proposed methods perform much better in comparison to the state-of-the-art methods, when the noise is approximately isotropic and stationary. A possible explanation is that detection and suppression of weak interferers at low INRs, is substantially easier in these noise fields for the proposed noise reduction systems.

We also examined the performance in situation where the target location is fixed to the front of the user (0°) and known to the beamforming systems. These situations may occur when the user steers the beamformer by rotating the head. However, we did not include these results as they lead to similar conclusions to the experiments with unknown target locations. This is due to the proposed algorithms being able to identify and suppress the interfering speaker more efficiently than the baseline methods.

In summary, the evaluation and statistical tests suggest that both proposed noise reduction systems, i.e. ML-BD and UML, have a significantly advantage over state-of-the-art methods in situations where a strong interferer is present. Additionally, we may conclude that the proposed systems, also perform on



(a) Noisy (b) Enhanced: ML-FRONTAL (c) Enhanced: UML

(d) Clean target speech (e) Residual noise: ML-FRONTAL (f) Residual noise: UML

Fig. 4: Spectrograms of the noisy, clean target, and processed signals from a single realization of the experiment in Sec. IV-D. The interferer is a competing speaker and the background noise is canteen noise. The INR is set to 12 dB, and the SINR is 0 dB. The target location is in the front (0°), and the interfering speaker is located at 75°. The figures show the spectrograms of the last 2.5 seconds of an acoustic scene with target presence. Fig. 4a and Fig. 4d show the noisy and clean target signals at the reference microphone, respectively. Fig. 4b and Fig. 4e show the output of the MWF beamformer using ML-FRONTAL, where Fig. 4b shows the processed signal and Fig. 4e shows the processed interference-plus-noise components only (i.e. without the target). Fig. 4c and Fig. 4f show the output of the MWF beamformer using UML, where Fig. 4c shows processed signal and Fig. 4f shows the processed interference-plus-noise components only.

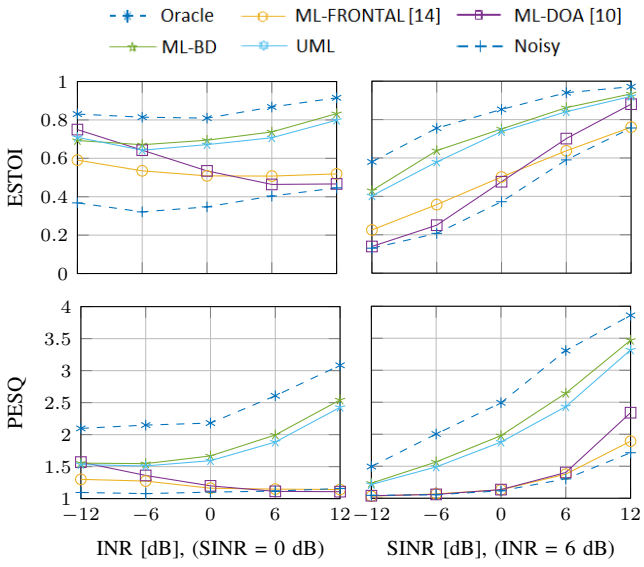


Fig. 5: Beamforming performance in *canteen noise*.

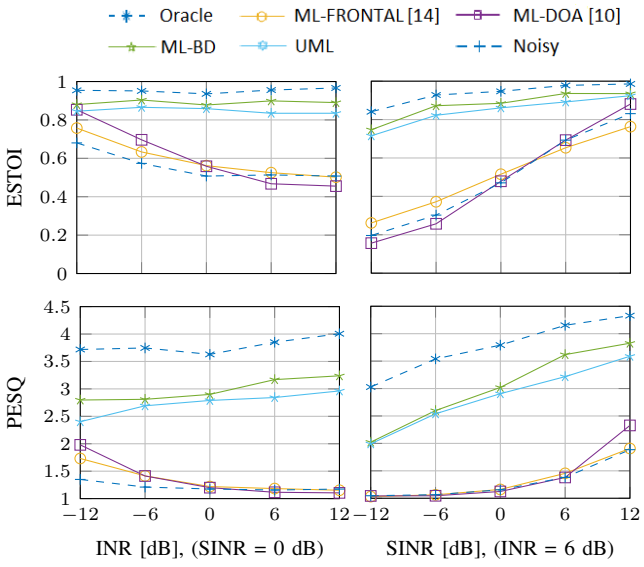


Fig. 6: Beamforming performance in *car noise*.

pair with state-of-the art in situations with weak interferers.

E. Simulation results with AIR mismatch and reverberation

In real-world scenarios, the AIRs of the RATF dictionary may not match the actual AIRs, and the microphone signals may be contaminated by reverberation in addition to noise. Both phenomena can potentially have a degrading impact on the beamforming performance of the proposed algorithms. To examine the robustness of the proposed algorithms, we therefore evaluate them against AIR mismatch and reverbera-

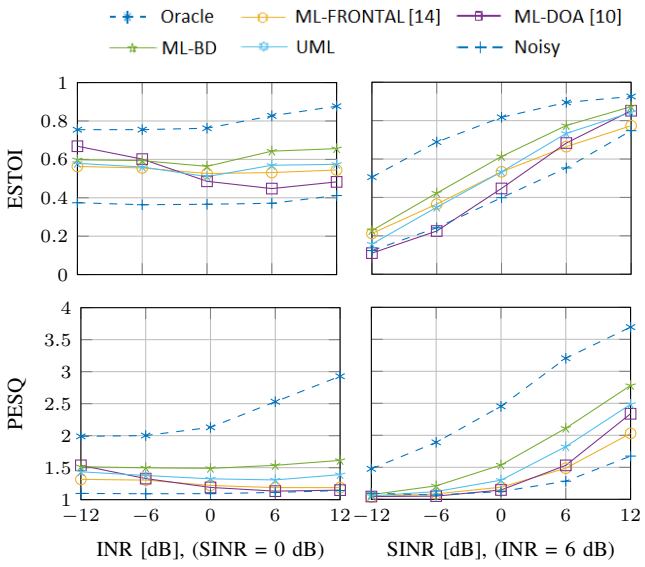


Fig. 7: Performance in *canteen noise* and *AIR mismatch*.

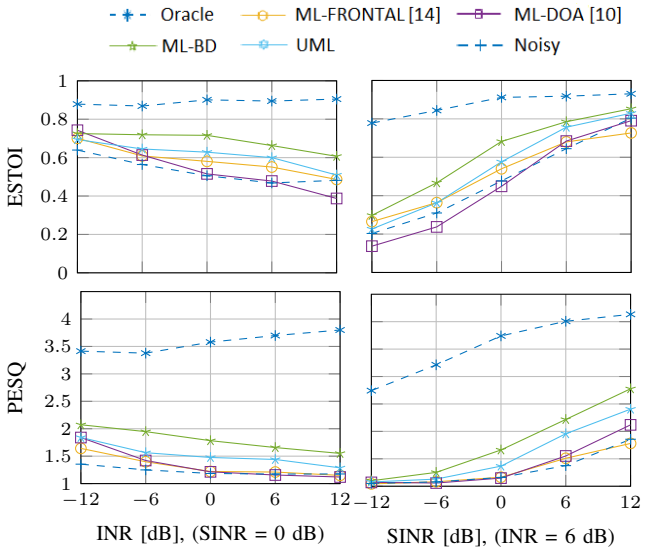


Fig. 8: Performance in *car noise* and *AIR mismatch*.

tion. We perform two experiments where the first experiment examines the robustness against AIR mismatch, and the second experiment examines the robustness against reverberation.

1) Simulation with AIR mismatch: In the first experiment, AIR mismatches can arise due to non-personalized RATF dictionaries. For example, the RATF dictionary may be measured on a different head than the HA user. To simulate such AIR mismatch, we use two sets of AIR databases fitted and measured on two arbitrarily chosen human heads. One is used to simulate the acoustic scene, and the other is used as a non-personalized RATF dictionary for the parameter estimation and MWF beamformer in the noise reduction systems.

Fig. 7 and Fig. 8 show the results for canteen noise and car noise, respectively. As expected, mismatches in the RATF dictionary cause performance degradations for all methods. Generally, the difference in mean ESTOI score between methods is smaller than the experiments in Sec. IV-D.

As in Sec. IV-D, we perform pairwise t-tests with Bonferroni corrected significance levels. We compare the proposed algorithms with ML-FRONTAL and ML-DOA. Statistically significant differences are primarily observed at high INRs, where ML-BD performs better than any of the baseline methods. At lower INRs, there are no strong indications that any of the noise reduction systems perform differently than the other.

2) Simulation with reverberation: In the second experiment, we use reverberant AIRs to simulate reverberation on the target and interference sources. The AIRs are measured in a listening room with physical dimensions L x W x H = 7.9 m x 6.0 m x 3.5 m. The reverberation time in the room is approximately $T_{60} = 150$ ms. The clean target and clean interference signals are convolved with reverberant AIRs to simulate the room reverberation. The canteen and car noise already contain natural reverberation from the environment they were measured in, hence we did not convolve the reverberant AIRs with the noise. We did not have access to a reverberant own voice transfer function, and therefore used the dry own voice transfer function in the simulation. We used an RATF dictionary obtained from the dry AIRs for the noise reduction systems.

Figs. 9 and 10 show performance results for canteen noise and car noise, respectively. Generally, all noise reduction systems suffer substantial performance degradation when the target and interference signals are reverberant. In canteen noise, we measure no strong statistical difference between the noise reduction systems except for ML-BD which performs better than ML-FRONTAL and ML-DOA at 12 dB INR. In terms of PESQ, the results are more decisive and seem to favor ML-BD which is statistically significant better than the baseline methods for INRs between 0 to 12 dB. In car noise, ML-BD performs significantly better than the baseline methods between -6 dB to 12 dB INR both in terms of ESTOI and PESQ. However, UML did not perform statistically significantly differently than any of the baseline methods.

Our evaluations with AIR mismatch and reverberation seem to favor ML-BD over UML and the baseline methods. However, despite reduced performance in these situations compared to the results in Sec. IV-D, it is worth pointing out that the overall conclusion remains: The proposed methods, in particular ML-BD, perform significantly better than the baseline methods in situations with a prominent interfering speaker and perform on par with the baseline methods in the absence of an interfering speaker.

V. Conclusion

In this paper, we propose multichannel noise reduction systems for hearing assistive devices (HADs). The proposed noise reduction systems can solve the problem of enhancing a target speech contaminated by noise and strong interfering speech, which is often considered difficult to solve for existing systems. We rely on the HAD user's own voice to identify interfering speech during own voice presence, but target absence. Furthermore, the multichannel Wiener filter (MWF) is used to retrieve the target speech and we propose three maximum likelihood estimation methods to estimate the target,

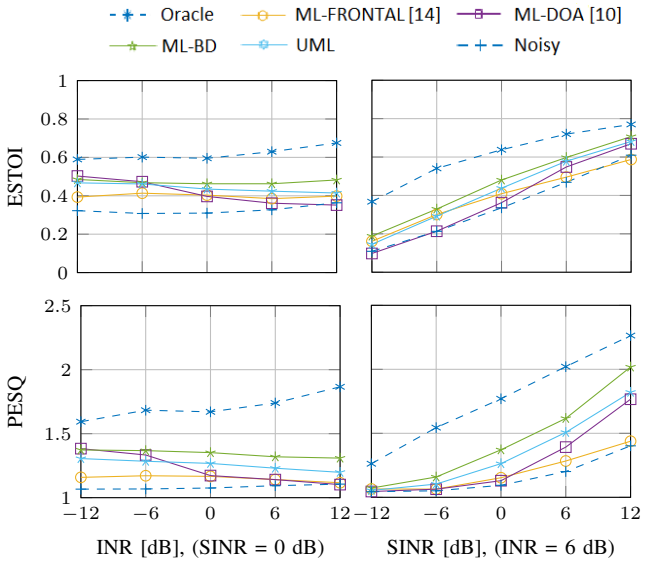


Fig. 9: Beamforming performance in *canteen noise* and *reverb*.

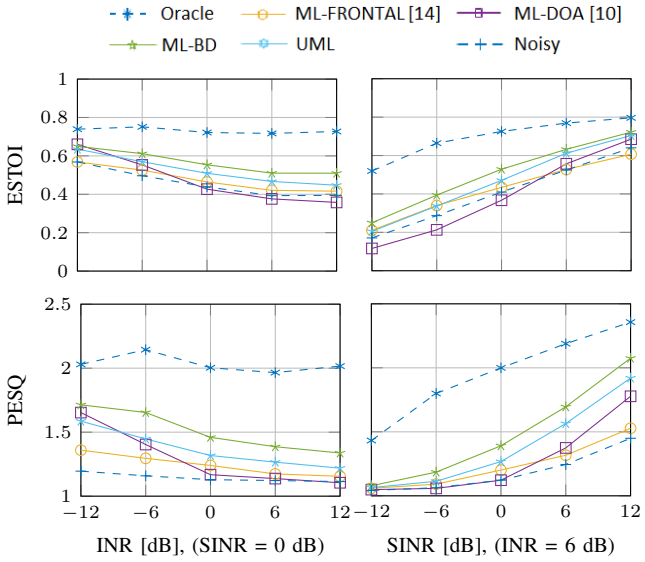


Fig. 10: Beamforming performance in car noise and *reverb*.

interfering speech, and noise statistics needed for the MWF beamformer.

The proposed noise reduction systems are compared to state-of-the-art methods in terms ESTOI and PESQ to examine estimated speech intelligibility and speech quality. Simulation results indicate that the proposed noise reduction systems are able to outperform the state-of-the-art methods particularly in situations with a prominent interfering speaker.

Appendix A
Speech Enhancement Performance Evaluation of the Maximum Likelihood Estimators

In this appendix, we evaluate the speech enhancement performance of the MLEs presented in Sec. III. The purpose of this evaluation is to show that the performances of ML-BD (Algorithm 3, and UML (Algorithm 4, see Sec. III-B) are close

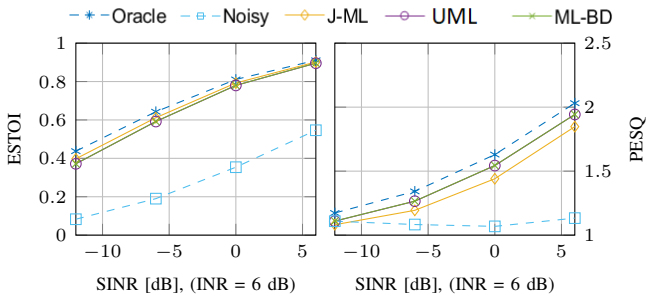


Fig. 11: Beamforming performance in situations with target, interferer, and noise with known target RATF vector.

to identical to the joint MLEs in III-A (J-ML). This result is of particular interest as the MLEs for ML-BD and UML only require a one-dimensional search over $\boldsymbol{d}_1$ and hence potentially a much lower computational complexity.

For simplicity, the target RATF vectors are assumed known and the evaluation is focused on speech enhancement performance. Although the target RATF vectors are assumed known for this evaluation, results can still give a sufficient indication on how ML-BD, UML, and J-ML compares. The comparison is made in terms of ESTOI and PESQ as a function of signal-to-interference-plus-noise ratio (SINR).

For this experiment, the setup is similar to the one presented in Sec. IV but with the target and interferer RATF vector known to the noise reduction systems.

Fig. 11 shows performance in terms of for ESTOI and PESQ as a function of SINR with the INR fixed to 6 dB. The unprocessed signal and the output of an oracle MWF with known target and noise statistics are also evaluated to indicate lower and upper performance bounds. Each performance score per SINR is averaged over 50 realization of acoustic scenes.

From Fig. 11, we see that J-ML, ML-BD, and UML perform almost identically without large differences. Furthermore, they perform close to the oracle MWF. We observe that J-ML performs slightly better than UML and ML-BD in terms of ESTOI but slightly worse in terms of PESQ. This is possibly due to UML and ML-BD having a slightly more aggressive noise suppression than J-ML which translates to marginally higher PESQ score at the cost of speech distortion and lower ESTOI score.

An evaluation of the runtime of the algorithms, showed that the real-time factors were 9.46 for UML, 4.98 for ML-BD, and 153.46 for J-ML. We see that J-ML has a significantly higher real-time factor compared to UML and ML-BD. This is partly due to the choice of the grid resolution used in J-ML as a high grid resolution will increase the real-time factor. Lowering the grid resolution will decrease the real-time factor but can result in performance degradation. It should also be noted that the implementation of the algorithms are not code optimized, but the real-time factors can still give a rough indication of computational complexity when comparing the proposed methods.

Because of the insignificant difference between J-ML, ML-BD and UML, we omit J-ML in the evaluation in Sec. IV due to its high computational complexity. However, if J-ML was

chosen to be included in the evaluation in Sec. IV, similar performance to ML-BD and UML would be expected.

REFERENCES

[1] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer handbook of speech processing*. Berlin ; London: Springer, 2008, oCLC: ocn190966783.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, ser. Springer topics in signal processing. Berlin: Springer, 2008, no. 1.

[3] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[4] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[5] X. Zhang and Y. Jia, "A Soft Decision Based Noise Cross Power Spectral Density Estimation for Two-Microphone Speech Enhancement Systems," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. 813–816.

[6] M. Souden, Jingdong Chen, J. Benesty, and S. Affes, "Gaussian Model-Based Multichannel Speech Presence Probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[7] M. Souden, J. Chen, J. Benesty, and S. Affes, "An Integrated Solution for Online Multichannel Noise Tracking and Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[8] M. Taseska and E. A. P. Habets, "Non-Stationary Noise PSD Matrix Estimation for Multichannel Blind Speech Extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2017.

[9] ——, *MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator*, International Workshop on Acoustic Signal Enhancement, Ed. Piscataway, N.J.: IEEE, 2012, oCLC: 835582169.

[10] M. Zohourian, G. Enzner, and R. Martin, "Binaural Speaker Localization Integrated Into an Adaptive Beamformer for Hearing Aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[11] 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 *2021 IEEE ICASSP*. IEEE, 2021.

[12] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[13] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003. [Online]. Available: http://ieeexplore.ieee.org/document/1223596/

[14] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 295–299.

[15] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[16] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - A theoretical and experimental comparison," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 91–95.

[17] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[18] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[19] Hao Ye and D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramer-Rao bounds for additive unknown colored noise," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, Apr. 1995.

[20] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 151–155.

[21] S. Braun and E. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–14, 2015.

[22] P. Hoang, Z.-H. Tan, T. Lunner, J. M. de Haan, and J. Jensen, "Maximum likelihood estimation of the interference-plus-noise cross power spectral density matrix for own voice retrieval," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020, p. xx.

[23] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," vol. 22, no. 12, pp. 2182–2196.

[24] S. Chakrabarty and E. A. P. Habets, "A Bayesian Approach to Informed Spatial Filtering With Robustness Against DOA Estimation Errors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 145–160, Jan. 2018.

[25] P. Hoang, Z.-H. Tan, J. M. de Haan, T. Lunner, and J. Jensen, "Robust Bayesian and Maximum a Posteriori Beamforming for Hearing Assistive Devices," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Ottawa, ON, Canada: IEEE, Nov. 2019, pp. 1–5.

[26] A. Kuklasinski and J. Jensen, "Multichannel Wiener Filters in Binaural and Bilateral Hearing Aids — Speech Intelligibility Improvement and Robustness to DoA Errors," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 8–16, Feb. 2017.

[27] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[28] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 544–548.

[29] A. J. Sørensen, M. Fereczkowski, and E. N. MacDonald, "Task Dialog By Native-Danish Talkers In Danish And English In Both Quiet And Noise," Mar. 2018, publisher: Zenodo. [Online]. Available: https://zenodo.org/record/1204951

[30] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066017, Dec. 2018.

[31] V. Best, E. Roverud, T. Streeter, C. R. Mason, and G. Kidd, "The Benefit of a Visually Guided Beamformer in a Dynamic Speech Task," *Trends in Hearing*, vol. 21, p. 233121651772230, Dec. 2017.

[32] A. Favre-Félix, C. Graversen, R. K. Hietkamp, T. Dau, and T. Lunner, "Improving Speech Intelligibility by Hearing Aid Eye-Gaze Steering: Conditions With Head Fixated in a Multitalker Environment," *Trends in Hearing*, vol. 22, p. 233121651881438, Jan. 2018.

[33] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Scientific Reports*, vol. 9, no. 1, p. 10451, Dec. 2019.

[34] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, p. 696, Dec. 1974.

[35] A. J. Munch Sørensen, M. Fereczkowski, and E. N. MacDonald, "Effects of noise and l2 on the timing of turn taking in conversation," in *7th International Symposium on Auditory and Audiological Research (ISAAR)*, Aug 2019.

[36] P. C. Loizou, *Speech enhancement: theory and practice*. Boca Raton, Fla.: CRC Press, 2013, oCLC: 958799095.

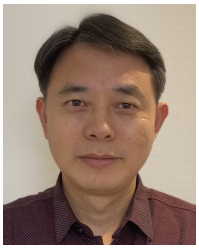
[37] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," vol. 14, no. 5, pp. 337–340. [Online]. Available: http://ieeexplore.ieee.org/document/4154719/

[38] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE*

Transactions on Signal Processing, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
[39] A. J. M. Sørensen, T. Lunner, and E. N. MacDonald, "Timing of turn taking between normal-hearing and hearing-impaired interlocutors," in *7th International Symposium on Auditory and Audiological Research (ISAAR)*, Aug 2019.
[40] H. Dillon, *Hearing aids*, 2nd ed. Sydney: Boomerang Press [u.a.], 2012.
[41] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, Jun. 2009.
[42] H. Akaike, "A new look at the statistical model identification," vol. 19, no. 6, pp. 716–723.
[43] P. Stoica and Y. Selen, "Model-order selection," vol. 21, no. 4, pp. 36–47. [Online]. Available: http://ieeexplore.ieee.org/document/1311138/
[44] Y. Laufer, B. Laufer-Goldshtein, and S. Gannot, "ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in Rank-Deficient Noise Field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 619–634, 2020.
[45] A. Moore, J. M. de Haan, M. Pedersen, D. Brookes, P. Naylor, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *Journal of the Acoustical Society of America*, vol. 145, pp. 2971–2981, 2019.
[46] D. Marquardt, S. Doclo, V. Hohmann, and R. Martin, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*, 1st ed. München: Verlag Dr. Hut, 2016.
[47] 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, Mar. 2015.
[48] Pei Chee Yong, S. Nordholm, and Hai Huyen Dam, "Effective Binaural Multi-Channel Processing Algorithm for Improved Environmental Presence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2012–2024, Dec. 2014.
[49] "TIMIT: acoustic-phonetic continuous speech corpus." Philadelphia, Pa., 1993.
[50] P. Minnaar, S. F. Albeck, C. Simonsen, B. Søndersted, S. Oakley, and J. Bennedbæk, "Reproducing real-life listening situations in the laboratory for testing hearing aids," *Journal of The Audio Engineering Society*, 2013.
[51] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," vol. 24, no. 3, pp. 543–558.
[52] 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, Nov. 2016.
[53] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2. Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.
[54] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*, 4th ed. Amsterdam ; Boston: Academic Press/Elsevier, 2009, oCLC: ocn255903275.

Poul Hoang received the M.Sc. degree in Electrical Engineering with specialization in Signal Processing and Computing from Aalborg University, Aalborg, Denmark, in 2018. He has since 2018 been pursuing his industrial Ph.D. at Oticon A/S in collaboration with Aalborg University, and partly funded by Innovation Fund Denmark. Since 2021, he has been employed as a DSP specialist at Oticon A/S. His main research interests include speech enhancement and acoustic signal processing.

Jan Mark de Haan received the M.Sc. degree in Electrical Engineering and the Ph.D. degree in Applied Signal Processing from Blekinge Institute of Technology, Karlskrona, Sweden, in 1998 and 2004, respectively. From 1999 to 2004 he was a Ph.D. student with the Department of Applied Signal Processing, Blekinge Institute of Technology. In 2003 he was a visiting researcher at the Western Australia Telecommunication Research Institute, Perth, Australia. Since 2004 he is employed at Oticon A/S, Copenhagen, Danmark. His main interest are in acoustic signal processing and signal processing applications in hearing aids.



Zheng-Hua Tan (M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 1999. He is a Professor in the Department of Electronic Systems and a Co-Head of the Centre for Acoustic Signal Processing Research at Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA, an Associate Professor at the Department of Electronic Engineering, SJTU, Shanghai, China, and a postdoctoral fellow at the AI Laboratory, KAIST, Daejeon, Korea. His research interests include machine learning, deep learning, pattern recognition, speech and speaker recognition, noise-robust speech processing, multimodal signal processing, and social robotics. He has (co)-authored over 200 refereed publications. He is the Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC). He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He has served as an Associate/Guest Editor for several other journals. He was the General Chair for IEEE MLSP 2018 and a TPC Co-Chair for IEEE SLT 2016.

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Principal Scientist with Oticon A/S, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, at Aalborg University. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.