Taseska and Habets *EURASIP Journal on Advances in Signal Processing* (2017) 2017:60
DOI 10.1186/s13634-017-0495-7

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

DOA-informed source extraction in the presence of competing talkers and background noise

Maja Taseska* and Emanuël A. P. Habets

Abstract

A desired speech signal in hands-free communication systems is often degraded by noise and interfering speech. Even though the number and locations of the interferers are often unknown in practice, it is justified to assume in certain applications that the direction-of-arrival (DOA) of the desired source is approximately known. Using the known DOA, fixed spatial filters such as the delay-and-sum beamformer can be steered to extract the desired source. However, it is well-known that fixed data-independent spatial filters do not provide sufficient reduction of directional interferers. Instead, the DOA information can be used to estimate the statistics of the desired and the undesired signals and to compute optimal data-dependent spatial filters. One way the DOA is exploited for optimal spatial filtering in the literature, is by designing DOA-based narrowband detectors to determine whether a desired or an undesired signal is dominant at each time-frequency (TF) bin. Subsequently, the statistics of the desired and the undesired signals can be estimated during the TF bins where the respective signal is dominant. In a similar manner, a Gaussian signal model-based detector which does not incorporate DOA information has been used in scenarios where the undesired signal consists of stationary background noise. However, when the undesired signal is non-stationary, resulting for example from interfering speakers, such a Gaussian signal model-based detector is unable to robustly distinguish desired from undesired speech. To this end, we propose a DOA model-based detector to determine the dominant source at each TF bin and estimate the desired and undesired signal statistics. We demonstrate that data-dependent spatial filters that use the statistics estimated by the proposed framework achieve very good undesired signal reduction, even when using only three microphones.

Keywords: Spatial filtering, Speech enhancement, PSD matrix estimation, RTF estimation, Signal detection

1 Introduction

In applications that require hands-free capture of speech, the desired speech signal is often corrupted by background noise and interfering speech signals. Such applications involve human-to-human and human-to-machine communication, where speech enhancement is crucial: in the former, to improve the communication comfort, and in the latter, to ensure low error rate of speech recognisers. In this work, we address scenarios where the desired speaker has a known DOA with respect to the microphones, such as in-car applications, or voice-controlled devices where the source of interest is restricted to a pre-defined DOA. Given the DOA of the desired source and assuming anechoic propagation, fixed spatial filters such as the delay-and-sum beamformer (DSB) [1] or superdirective beamformers [2] can be used. However, these filters are suboptimal as they do not consider the spatio-temporal statistics of the signals, and often provide insufficient interference reduction. Moreover, propagation model mismatch due to reverberation and DOA errors further limit the performance. In this work, we focus on optimal data-dependent spatial filters [3, 4]. Two main paradigms can be distinguished which aim at optimal filtering given the source DOA: robust adaptive beamformers (RABs) and informed spatial filters (ISFs). While RABs seek to improve the robustness to errors in the DOAs and the signal propagation vectors, ISFs address the estimation of propagation vectors and signal statistics

*Correspondence: maja.taseska@audiolabs-erlangen.de
International Audio Laboratories Erlangen, Erlangen, Germany

from the microphone signals, and their usage for optimal spatial filtering.

RAB representatives include Bayesian beamformers [5, 6] and spatial filters with eigenvector constraints [7, 8], which are implemented using linearly constrained minimum variance (LCMV) filters. These filters seek to minimise the undesired signal power at the output, while imposing constraints to ensure that the desired signal from the DOA of interest is undistorted. Another approach is proposed in [9], where the desired signal power spectral density (PSD) matrix is computed by integrating the free field-based PSD matrices across the region of possible source DOAs. Note that the increased robustness in these approaches often comes at the cost of worse undesired signal reduction. RABs can also be implemented in a general sidelobe canceler (GSC) structure, where the robustness to DOA and propagation vector mismatches is ensured by using an adaptive blocking matrix [10, 11], and by imposing constraints to the adaptive noise cancellers [10]. The robust GSCs require a desired signal detector, as the noise cancellers need to be updated when the desired signal is absent, while the blocking matrix when the desired signal is present [12].

ISFs, in contrast to RABs, estimate the desired signal propagation vector and the undesired signal statistics from the data, and substitute them in optimal filter expressions such as the minimum variance distortionless response (MVDR) or the multichannel Wiener filter (MWF) [13–16]. As the ISFs are estimated and implemented in the frequency domain, the relevant statistics correspond to the PSD matrices of the desired and the undesired signals at each frequency. The advantage of estimating the propagation vectors from the data, rather than using anechoic propagation models, is well-known since the development of the transfer function-GSC [17] and the relative transfer function (RTF)-GSC [18]. However, a less often addressed question is how to design narrowband signal detectors which are required to estimate the propagation vectors and the PSD matrices, or perform the filter adaptations in the adaptive GSCs. Signal detection in the presence of non-stationary interferers is a very challenging problem [19]. The Gaussian model-based detectors used in state-of-the-art systems [15, 20, 21] assume that the noise is significantly more stationary than speech, which is not true for speech interferers.

The question addressed in this paper is how to design a robust narrowband signal detector, by using the microphone signals, narrowband DOA estimates extracted from the signals, and the information about the desired source DOA. Narrowband DOAs have been previously used for desired speech detection in the literature. For instance, in [22], the authors use narrowband DOAs to control the a priori desired speech presence probability (DSPP) in a Gaussian signal model, while in [23] a Gaussian DOA model is used to compute a DSPP and apply it as a single-channel gain to the output of a spatial filter. We propose a different statistical model for the narrowband DOA estimates which is used for desired signal detection and estimation of the propagation vectors and the PSD matrices in an ISF framework. Initial results obtained using the proposed framework were presented in [24]. In this paper, we provide a more detailed description of the system, further discussions and comparison to the state-of-the-art approaches, as well as an extended set of experiments to evaluate the performance of the narrowband signal detector and the quality of the extracted source signal at the ISF output.

2 Signal model and problem formulation

A compact array of M microphones captures the signal of a desired speaker, unknown number of competing speakers, and background noise. The STFT-domain signal of the mth microphone is given by

$$Y_m(t,k) = S_m(t,k) + I_m(t,k) + V_m(t,k), \tag{1}$$

where S_m, I_m, and V_m are the signals of the desired speaker, the sum of competing speakers, and the noise, and t and k are time and frequency indices. Let the $M\times 1$ vectors $\boldsymbol{y}(t,k)$, $\boldsymbol{s}(t,k)$, $\boldsymbol{i}(t,k)$, and $\boldsymbol{v}(t,k)$ contain the respective signals from all M microphones. The corresponding PSD matrices are given by $\boldsymbol{\Phi}_{\boldsymbol{s}} = \mathrm{E}\left[\boldsymbol{s}\,\boldsymbol{s}^{\mathrm{H}}\right]$, $\boldsymbol{\Phi}_{\boldsymbol{i}} = \mathrm{E}\left[\boldsymbol{i}\,\boldsymbol{i}^{\mathrm{H}}\right]$, and $\boldsymbol{\Phi}_{\boldsymbol{v}} = \mathrm{E}\left[\boldsymbol{v}\,\boldsymbol{v}^{\mathrm{H}}\right]$, where $\mathrm{E}[\cdot]$ denotes statistical expectation. As the signals are zero-mean and mutually uncorrelated, the following holds

$$\boldsymbol{\Phi}_{\boldsymbol{y}}(t,k) = \boldsymbol{\Phi}_{\boldsymbol{s}}(t,k) + \boldsymbol{\Phi}\boldsymbol{i}(t,k) + \boldsymbol{\Phi}_{\boldsymbol{v}}(t,k). \tag{2}$$

In addition, we define the undesired signal PSD matrix containing the speech interferers and the noise as $\boldsymbol{\Phi}_{\boldsymbol{u}}(t,k) = \boldsymbol{\Phi}\boldsymbol{i}(t,k) + \boldsymbol{\Phi}\boldsymbol{v}(t,k)$. We seek to estimate the desired signal $S_m(t,k)$ at the mth microphone, and without loss of generality, we consider the first microphone as a reference, i.e., $m = 1$. The desired signal estimate $\widehat{S}_1(t,k)$ is obtained by applying a time and frequency-dependent optimal linear filter $\boldsymbol{w}_{\mathrm{opt}}$ to the microphone signals as follows

$$\widehat{S}_1(t,k) = \boldsymbol{w}_{\mathrm{opt}}^{\mathrm{H}}(t,k)\,\boldsymbol{y}(t,k). \tag{3}$$

Two fundamental assumptions underlie the proposed framework for computing the optimal filter coefficients $\boldsymbol{w}_{\mathrm{opt}}^{\mathrm{H}}(t,k)$. The first assumption is used in many existing multi-channel speech enhancement approaches and states that the PSD matrix of the desired signal is a rank-one matrix given by

$$\boldsymbol{\Phi}_{\boldsymbol{s}}(t,k) = \phi_{S_1}(t,k)\,\boldsymbol{g}_1(k)\,\boldsymbol{g}_1^{\mathrm{H}}(k), \tag{4}$$

where $\boldsymbol{g}_1(k)$ is the RTF vector of the desired source with the first microphone as a reference, and $\phi_{S_1}(t,k) = \mathrm{E}\left[|S_1(t,k)|^2\right]$ is the PSD of the desired signal at the first

microphone. The second assumption is that the speech signals are sparse in the STFT domain, so that each TF bin can be associated to one of the following mutually exclusive hypotheses

$$\mathcal{H}_s : \boldsymbol{y}(t,k) \approx \boldsymbol{s}(t,k) + \boldsymbol{v}(t,k) \text{ desired signal is dominant,} \tag{5a}$$

$$\mathcal{H}_i : \boldsymbol{y}(t,k) \approx \boldsymbol{i}(t,k) + \boldsymbol{v}(t,k) \text{ speech interferer is dominant,} \tag{5b}$$

$$\mathcal{H}_v : \boldsymbol{y}(t,k) \approx \boldsymbol{v}(t,k) \text{ background noise is dominant.} \tag{5c}$$

In addition, we introduce the hypothesis $\mathcal{H}_u = \mathcal{H}_i \cup \mathcal{H}_v$ that undesired signal is dominant, regardless whether it is a competing speaker or background noise.

The objective in this work is to define likelihood models for the hypotheses in (5), design a detector that associates each TF bin to the correct hypothesis, estimate the PSD matrices and the RTF vector $\boldsymbol{g}_1$, and finally, compute the ISF coefficients $\boldsymbol{w}_{\text{opt}}$ required for source extraction in (3). Narrowband DOA estimates play a key role in the framework, and appropriate state-of-the art estimators are briefly discussed in the following section.

3 Narrowband DOA estimation

The most important criteria when choosing a DOA estimator for our framework is the ability to obtain nearly instantaneous narrowband DOA estimates without requiring temporally averaged covariance matrices as the subspace-based estimators [25], and a sufficiently low complexity suitable for real-time implementation. We briefly review two estimators which satisfy these requirements.

3.1 Least squares (LS)-fitting of instantaneous phase differences [26]

Assuming that a single source is dominant at each TF bin, its DOA can be estimated using the phase differences between the microphone signals at each TF bin. Denoting the 2D microphone locations as $\boldsymbol{d}_1, \dots, \boldsymbol{d}_M$, the frequency corresponding to the index k by f_k (in Hertz), the DOA of the dominant source as θ_{tk} (in radians), and the corresponding DOA vector as $\boldsymbol{q}(t,k) = [\cos(\theta_{tk}), \sin(\theta_{tk})]$, the anechoic RTF vector with respect to the first microphone reads

$$\boldsymbol{g}_1(t,k) = \left[1, \mathrm{e}^{j\frac{2\pi f_k}{c}(\boldsymbol{d}_2-\boldsymbol{d}_1)^{\mathrm{T}}\boldsymbol{q}(t,k)}, \dots, \mathrm{e}^{j\frac{2\pi f_k}{c}(\boldsymbol{d}_M-\boldsymbol{d}_1)^{\mathrm{T}}\boldsymbol{q}(t,k)}\right]. \tag{6}$$

Due to the relation $\boldsymbol{s}(t,k) = \boldsymbol{g}_1(t,k)\, S_1(t,k)$, the phase differences of the source signal at each microphone with respect to the first microphone (provided that $S_1(t,k) \neq 0$) are given by

$$\angle\frac{\boldsymbol{s}(t,k)}{S_1(t,k)} = \left[0, \frac{2\pi f_k}{c}(\boldsymbol{d}_2 - \boldsymbol{d}_1)^{\mathrm{T}}\,\boldsymbol{q}(t,k), \dots, \frac{2\pi f_k}{c}(\boldsymbol{d}_M - \boldsymbol{d}_1)^{\mathrm{T}}\,\boldsymbol{q}(t,k)\right]. \tag{7}$$

If we introduce the $(M-1)\times 1$ vector $\bar{\boldsymbol{s}} = \left[\angle\frac{S_2}{S_1}, \angle\frac{S_3}{S_1}, \dots, \angle\frac{S_M}{S_1}\right]$, and the $(M-1)\times 2$ matrix $\boldsymbol{D}$ containing $(\boldsymbol{d}_i - \boldsymbol{d}_j)^{\mathrm{T}}$ as rows for $i \in [2, M]$, Eq. (7) can be rewritten as

$$\bar{\boldsymbol{s}} = \frac{2\pi f_k}{c}\boldsymbol{D}\,\boldsymbol{q}(t,k), \tag{8}$$

and can be solved for the DOA vector $\boldsymbol{q}(t,k)$. However, in practice, the signal $\bar{\boldsymbol{s}}$ is not observable. Instead, the noisy phase vector $\bar{\boldsymbol{y}} = \left[\angle\frac{Y_2}{Y_1}, \angle\frac{Y_3}{Y_1}, \dots, \angle\frac{Y_M}{Y_1}\right]$, is used in (8) to obtain an estimate of the DOA vector by solving the LS problem

$$\hat{\boldsymbol{q}}(t,k) = \arg\min_{\boldsymbol{q}} \|\bar{\boldsymbol{y}}(t,k) - \frac{2\pi f_k}{c}\boldsymbol{D}\,\boldsymbol{q}\|_2^2 = \frac{c}{2\pi f_k}\boldsymbol{D}^{+}\bar{\boldsymbol{y}}(t,k), \tag{9}$$

where $()^{+}$ denotes Moore-Penrose pseudoinverse of a matrix.

3.2 LS-fitting of cross PSD phase differences [27]

Instead of instantaneous phase differences, the authors in [27] use phase differences between the short-term cross PSDs to estimate the DOA. According to the model in (6), the cross PSD between the mth and nth microphone is given by

$$\phi_{S,mn}(t,k) = \mathrm{E}\left[S_m(t,k)\, S_n(t,k)^*\right] = \mathrm{E}\left[|S_m(t,k)|^2\right] \mathrm{e}^{j\frac{2\pi f_k}{c}(\boldsymbol{d}_m-\boldsymbol{d}_n)^{\mathrm{T}}\boldsymbol{q}(t,k)}. \tag{10}$$

Introducing the $(M-1)\times 1$ vector $\bar{\boldsymbol{\phi}}(t,k) = \left[\angle\frac{\phi_{S,12}(t,k)}{\phi_{S,11}(t,k)}, \dots, \angle\frac{\phi_{S,1M}(t,k)}{\phi_{S,11}(t,k)}\right]$, and using the matrix $\boldsymbol{D}$ similarly as in (8) we obtain the relation

$$\bar{\boldsymbol{\phi}}_s = \frac{2\pi f_k}{c}\boldsymbol{D}\,\boldsymbol{q}(t,k). \tag{11}$$

As the signals $S_1(t,k), \dots S_M(t,k)$ are unobservable, the noisy cross-PSDs $\phi_{Y,mn}(t,k)$ can be used instead. By defining $\bar{\boldsymbol{\phi}}_Y(t,k) = \left[\angle\frac{\phi_{Y,12}(t,k)}{\phi_{Y,11}(t,k)}, \dots, \angle\frac{\phi_{Y,1M}(t,k)}{\phi_{Y,11}(t,k)}\right]$, the DOA vector estimate is obtained analogously to (9), as

$$\hat{\boldsymbol{q}}(t,k) = \frac{c}{2\pi f_k}\boldsymbol{D}^{+}\bar{\boldsymbol{\phi}}_Y(t,k). \tag{12}$$

The estimators given by (9) and (12) assume that for each microphone pair, the spatial aliasing frequency lies above $\frac{F_s}{2}$, where F_s is the sampling rate. Alternatively, frequency-dependent binary weights can be used to exclude microphone pairs at the frequency bins where spatial aliasing might occur for those pairs, as done in [27].

4 State-of-the-art DOA-informed source extraction

4.1 DSB and MPDR beamforming

If the signal propagation is modelled as a pure delay, the DSB is the simplest filter which can be applied for source extraction. However, the DSB offers suboptimal performance as it does not consider the signal statistics and the reverberation.

Data-dependent spatial filters such as the MVDR or the LCMV filter can be applied if the locations or the PSD matrices of the interfering sources are known. However, in the considered application, this information is unavailable. One possibility to employ data-dependent spatial filtering without requiring the undesired signal PSD matrix is by using a minimum power distortionless response (MPDR) beamformer [28], computed using the anechoic RTF vector (6) and the microphone signal PSD matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$. An MPDR filter that provides an estimate of the desired signal received at the first microphone is given by

$$\boldsymbol{w}_{\mathrm{mpdr}}(t,k) = \frac{\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}(t,k)\,\boldsymbol{g}_1(k)}{\boldsymbol{g}_1{}^{\mathrm{H}}(k)\,\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}(t,k)\boldsymbol{g}_1(k)}. \tag{13}$$

In contrast to the MVDR filter which is expressed in terms of the undesired signal PSD matrix $\boldsymbol{\Phi}_{\boldsymbol{u}}$, the MPDR filter is expressed in terms of $\boldsymbol{\Phi}_{\boldsymbol{y}}$, which contains the desired signal as well. Therefore, if the RTF vector is inaccurate due to the anechoic model mismatch in reverberant environments, or due to DOA errors, the MPDR filter causes severe distortion of the desired signal [28].

4.2 Informed spatial filtering

To extract a desired source from a given DOA, while reducing noise and directional interferers using ISFs, the narrowband detectors used for RTF and PSD matrix estimation need to distinguish TF bins where desired signal is dominant from TF bins where undesired signal is dominant. Such framework was developed in [22] for spatial filtering in the spherical harmonic domain, where the signal detector was obtained by estimating a Gaussian model-based DSPP. To estimate the DSPP, the likelihoods of the signal vector are modelled as (indices t and k omitted for brevity)

$$f\left(\boldsymbol{y}\,|\,\mathcal{H}_u\right) = \left(\pi^M \det[\,\boldsymbol{\Phi}_{\boldsymbol{u}}]\right)^{-1}\ \mathrm{e}^{-\boldsymbol{y}^{\mathrm{H}}\boldsymbol{\Phi}_{\boldsymbol{u}}^{-1}\boldsymbol{y}}, \tag{14a}$$

$$f\left(\boldsymbol{y}\,|\,\mathcal{H}_s\right) = \left(\pi^M \det\left[\boldsymbol{\Phi}_{\boldsymbol{s}} + \boldsymbol{\Phi}_{\boldsymbol{v}}\right]\right)^{-1}\ \mathrm{e}^{-\boldsymbol{y}^{\mathrm{H}}(\boldsymbol{\Phi}_{\boldsymbol{s}}+\boldsymbol{\Phi}_{\boldsymbol{v}})^{-1}\boldsymbol{y}}. \tag{14b}$$

Given an a priori DSPP $q_s = p(\mathcal{H}_s)$, the a posteriori DSPP follows from the Bayes theorem as follows

$$p\left(\mathcal{H}_s\,|\,\boldsymbol{y}\right) = \frac{q_s f\left(\boldsymbol{y}\,|\,\mathcal{H}_s\right)}{q_s f\left(\boldsymbol{y}\,|\,\mathcal{H}_s\right) + (1-q_s) f\left(\boldsymbol{y}\,|\,\mathcal{H}_u\right)}. \tag{15}$$

The authors in [22] incorporate the DOA information in the a priori DSPP q_s, to provide more robust discrimination between desired and undesired speakers. If $\Theta_{\theta,\hat{\theta}}(t,k)$ denotes the angle between the true DOA θ of the source of interest, and $\hat{\theta}_{tk}$ the DOA estimate at TF bin (t,k), the a priori DSPP in [22] is computed as $q_s(t,k) = w\left(\Theta_{\theta,\hat{\theta}}(t,k)\right)$, where $w(\Theta)$ is a Gaussian window centred at $\Theta = 0$.

5 Proposed DOA model-based signal detection

The Gaussian model-based DSPP is very sensitive to non-stationarity of the undesired signal, as the expression (15) requires an estimate of the PSD matrix $\boldsymbol{\Phi}_{\boldsymbol{u}}$. To estimate the DSPP at TF bin (t,k), the PSD matrix estimate $\widehat{\Phi}_{\boldsymbol{u}}(t-1,k)$ from the previous frame $t-1$ is used, which leads to estimation errors when the undesired signal changes in consecutive frames. The DOA-based a priori DSPP used in [22] in the spherical harmonic domain, seeks to reduce this sensitivity in scenarios with non-stationary interferers. Nevertheless, our experiments for a posteriori DSPP estimation in the traditional signal domain indicated that the DOA-based a priori DSPP is often insufficient to compensate for errors in the likelihoods (15) occurring due to erroneous $\widehat{\Phi}_{\boldsymbol{u}}$ estimates. This is our motivation to develop a different method to incorporate DOA information in the a posteriori DSPP estimation, by using a generative probabilistic model of the narrowband DOAs.

5.1 Likelihood model for the narrowband DOA estimates

To derive the a posteriori DSPP, we propose likelihood models for the DOA estimates under the hypotheses $\mathcal{H}_s$, $\mathcal{H}_i$, and $\mathcal{H}_v$. As the DOA estimates represent circular random variables, we propose to model $f\left(\hat{\theta}_{tk}\,|\,\mathcal{H}_s\right)$ by a von Mises distribution, which closely approximates a wrapped normal distribution on the circle [29]. The von Mises distribution is characterised by a mean $\tilde{\theta}$ and a concentration κ, and is given by

$$f\left(\hat{\theta}_{tk}\,|\,\mathcal{H}_s;\tilde{\theta},\kappa\right) = c_{\mathcal{M}}(\kappa)\,\mathrm{e}^{\kappa\cos\left(\hat{\theta}_{tk}-\tilde{\theta}\right)}. \tag{16}$$

The normalisation $c_{\mathcal{M}}(\kappa) = [\,2\pi I_0(\kappa)]^{-1}$ is derived in [30], where I_0 is the modified Bessel function of the first kind. If the DOA estimator is unbiased, the mean $\tilde{\theta}$ is equal to the DOA of the desired source. The concentration parameter κ reflects the uncertainty in the DOA estimates, where larger concentration indicates larger DOA estimation error variance, while smaller concentration indicates smaller DOA estimation error variance. Factors which commonly affect the concentration include the

array geometry, the number of microphones, the coherent signal-to-diffuse signal ratio, as well as the DOA estimator. The concentration κ is an unknown model parameter and its computation is discussed in Section 5.3.2.

Assuming spatially isotropic background noise, where the sound may originate from all directions with equal probability, the likelihood $f\left(\hat{\theta}_{tk} \mid \mathcal{H}_v\right)$ is modelled by a uniform distribution on the circle, i.e., $f\left(\hat{\theta}_{tk} \mid \mathcal{H}_v\right) = (2\pi)^{-1}$. It remains to define the likelihood of the DOA estimates under the hypothesis $\mathcal{H}_i$. If the number and the DOAs of the interferers were known, a multimodal distribution on the circle would accurately model $f\left(\hat{\theta}_{tk} \mid \mathcal{H}_i\right)$. In practice, this information is unavailable and difficult to estimate. Instead, we propose to model $f\left(\hat{\theta}_{tk} \mid \mathcal{H}_i\right)$ as approximately uniform in regions sufficiently far from the desired source, and with a notch centred at the DOA of the source. We construct such a distribution by considering the following function $g(\theta, \tilde{\theta}, \kappa)$

$$g(\theta, \tilde{\theta}, \kappa) = -\mathrm{e}^{\kappa \cos\left(\theta - \tilde{\theta}\right)} + \mathrm{e}^{\kappa}, \tag{17}$$

which attains the minimum value at $\theta = \tilde{\theta}$ and approaches a uniform distribution as θ deviates from $\tilde{\theta}$. To obtain a probability density, $g(\theta, \tilde{\theta}, \kappa)$ is normalised by $c_{\mathcal{A}}$ such that

$$\int c_{\mathcal{A}}\, g(\theta, \tilde{\theta}, \kappa)\, \mathrm{d}\theta = c_{\mathcal{A}} \int -\mathrm{e}^{\kappa \cos(\theta - \tilde{\theta})} + \mathrm{e}^{\kappa}\, \mathrm{d}\theta = 1. \tag{18}$$

The integral of the first term is equal to the normalisation constant of the von Mises distribution, and $\int e^{\kappa}\, \mathrm{d}\theta = 2\pi\, \mathrm{e}^{\kappa}$. Therefore, the constant $c_{\mathcal{A}}$ is given by

$$\begin{aligned} c_{\mathcal{A}}(\kappa) &= \left[-2\pi (I_0(\kappa) - \mathrm{e}^{\kappa})\right]^{-1} \text{ and} \\ f\left(\hat{\theta}_{tk} \mid \mathcal{H}_i; \tilde{\theta}, \kappa\right) &= c_{\mathcal{A}}(\kappa) \left(-\mathrm{e}^{\kappa \cos\left(\hat{\theta}_{tk} - \tilde{\theta}\right)} + \mathrm{e}^{\kappa}\right). \end{aligned} \tag{19}$$

The von Mises distribution and the proposed notched distribution are illustrated in Fig. 1 for different values of the concentration parameter κ.

5.2 Desired speech presence probability and optimal detection

Having defined the likelihoods, the a posteriori DSPP and the a posteriori desired speech absence probability (DSAP) are given by the Bayes theorem as

$$p\left(\mathcal{H}_s \mid \hat{\theta}_{tk}\right) = \frac{q_s f\left(\hat{\theta}_{tk} \mid \mathcal{H}_s; \tilde{\theta}, \kappa\right)}{q_s f\left(\hat{\theta}_{tk} \mid \mathcal{H}_s; \tilde{\theta}, \kappa\right) + q_i f\left(\hat{\theta}_{tk} \mid \mathcal{H}_i; \tilde{\theta}, \kappa\right) + q_v f\left(\hat{\theta}_{tk} \mid \mathcal{H}_v\right)}, \tag{20}$$

$$p\left(\mathcal{H}_u \mid \hat{\theta}_{tk}\right) = \frac{q_i f\left(\hat{\theta}_{tk} \mid \mathcal{H}_i; \tilde{\theta}, \kappa\right) + q_v f\left(\hat{\theta}_{tk} \mid \mathcal{H}_v\right)}{q_s f\left(\hat{\theta}_{tk} \mid \mathcal{H}_s; \tilde{\theta}, \kappa\right) + q_i f\left(\hat{\theta}_{tk} \mid \mathcal{H}_i; \tilde{\theta}, \kappa\right) + q_v f\left(\hat{\theta}_{tk} \mid \mathcal{H}_v\right)}, \tag{21}$$

where the a priori probabilities $q_s = p(\mathcal{H}_s)$, $q_i = p(\mathcal{H}_i)$ and $q_v = p(\mathcal{H}_v)$ satisfy $q_s + q_i + q_v = 1$.

In scenarios with stationary undesired signals, such as background noise, the a posteriori DSPP is directly used for recursive estimation of the noise PSD matrix [15, 20, 21]. However, if the undesired signal contains speech, recursive updates using the DSPP introduce leakage of undesired signal into the desired signal PSD matrix and vice versa. Therefore, in this work, we employ the estimated DSPP and DSAP to compute an optimal binary detector at each TF bin, which minimises the Bayes risk for a false positive and a false negative costs $C_{su}, C_{\mathrm{us}} > 0$, as follows [31]

$$\begin{aligned} &\text{decide } \mathcal{I}_{\mathcal{H}_s} = 1,\ \mathcal{I}_{\mathcal{H}_u} = 0 \text{ if } \frac{p\left(\mathcal{H}_s \mid \hat{\theta}_{tk}\right)}{p\left(\mathcal{H}_u \mid \hat{\theta}_{tk}\right)} > \frac{C_{su}}{C_{us}}, \\ &\text{decide } \mathcal{I}_{\mathcal{H}_u} = 1,\ \mathcal{I}_{\mathcal{H}_s} = 0 \text{ otherwise,} \end{aligned} \tag{22}$$

where $\mathcal{I}_{\mathcal{H}_a}$ is a binary indicator which equals one if the hypothesis in the subscript is true, and zero otherwise. Using the binary indicator, only the PSD matrix of the dominant signal is updated, as discussed in Section 6 in more detail.

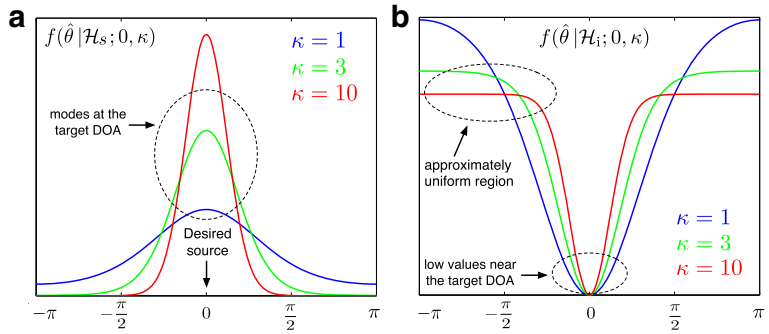

Fig. 1 Illustration of the DOA-based likelihoods under the different hypotheses **a** Von Mises distribution. **b** Notched distribution

5.3 Estimation of the likelihood model parameters

5.3.1 Estimation of the a priori probabilities q_s, q_i, and q_v

The Gaussian signal model has been successfully used to compute the DSPP in scenarios where the undesired signal consists of stationary noise [15, 20, 21, 32]. Denoting the speech presence hypothesis (desired or undesired speech) by $\mathcal{H}_{si} = \mathcal{H}_s \cup \mathcal{H}_i$, we can define the Gaussian likelihoods $f(\boldsymbol{y} \mid \mathcal{H}_v)$ and $f(\boldsymbol{y} \mid \mathcal{H}_{si})$ according to (14), with the appropriate PSD matrices $\boldsymbol{\Phi_v}$ and $\boldsymbol{\Phi_y} - \boldsymbol{\Phi_v}$, and compute an a posteriori speech presence probability (SPP) $p\left(\mathcal{H}_{si} \mid \boldsymbol{y}(t,k)\right)$ using the Bayes rule. The a posteriori SPP from the Gaussian model can then be used as an a priori SPP in our proposed DOA-based model, such that

$$\begin{aligned} q_v(t,k) &= 1 - p\left(\mathcal{H}_{si} \mid \boldsymbol{y}(t,k)\right) \quad \text{and} \\ q_s(t,k) &= q_i(t,k) = 0.5\left(1 - q_v(t,k)\right). \end{aligned} \tag{23}$$

In this manner, the a priori SPP in the proposed model exploits the spatio-temporal properties of the signal vector $\boldsymbol{y}(t,k)$ and knowledge of the noise PSD matrix to aid the discrimination between noise and speech-dominated TF bins, prior to the estimation of the narrowband DOA at the current TF bin.

5.3.2 Estimation of the concentration parameter κ

It was mentioned in Section 5.1 that the concentration parameter κ related to the mode and the notch of the DOA-related likelihoods often depends on the coherent-to-diffuse ratio (CDR), the array geometry, and the DOA estimator. For a given array geometry and a given DOA estimator, a single concentration parameter can be estimated for instance by collecting the DOA estimates from all TF bins during a training period when only the desired speech source and background noise are present, and finding the maximum likelihood (ML) estimate. However, this way of obtaining a single concentration parameter does not take into account the fact that many of the TF bins used for training are noise-dominated and do not contain significant speech energy. Instead, of providing an average concentration parameter, we seek to quantify the uncertainty of the DOA estimate at each TF bin. By quantifying the certainty of each DOA estimate, we provide additional information to the proposed signal detector for determining the dominant source at each TF bin. Therefore, the concentration parameter κ of the von Mises distribution needs to estimated for each TF bin as well.

In this work, we propose to use the short-time CDR $\widehat{\Gamma}_{tk}$ estimated from the microphone signals at each TF bin to control the concentration parameter. The motivation to use the CDR, stems from the fact that if the CDR is high at a given TF bin, it is more likely that the estimated DOA accurately indicates the DOA of the coherent sound. In such TF bins, $f(\hat{\theta}_{tk} \mid \mathcal{H}_s)$ and $f(\hat{\theta}_{tk} \mid \mathcal{H}_i)$ should have a high concentration κ, resulting in narrow mode or notch. If the CDR is low, the concentration should be lower, to reflect larger uncertainty in the DOA estimates. To estimate the CDR, we use the estimator proposed in [33], which is based on the short-term complex coherence between two microphone signals. The underlying assumption for this CDR estimator is that the sound field at each TF bin can be modelled as a superposition of a direct sound component with a given DOA (originating from a directional source, such as a speaker), and a diffuse sound component corresponding to late reverberation and diffuse background noise. If we express the CDR in dB, and consider that the range $[-\infty, \infty]$ of the CDR needs to be mapped to the non-negative concentration parameter, we propose the following sigmoid-like function for the mapping

$$\kappa_{tk} = f\left(\widehat{\Gamma}_{tk}\right) = l_{\max} \frac{10^{-c\rho}}{10^{-c\rho} + 10^{-\rho\,\widehat{\Gamma}_{tk}/10}}, \tag{24}$$

where $l_{\max}$ determines the maximum value of the function, $c \in \mathbb{R}$ controls the offset along the $\widehat{\Gamma}_{tk}$ axis, and $\rho \in \mathbb{R}^+$ controls the steepness of transition region of the sigmoid-like function. The minimum value of the function $f\left(\widehat{\Gamma}_{tk}\right) = 0$, attained in the limit $\widehat{\Gamma}_{tk} \to -\infty$ indicates that the distribution of the DOA is a uniform distribution on the circle in the absence of a coherent signal.

The remaining question is how to determine the parameters of the sigmoid function, so that the concentration parameter κ accurately describes the distribution of the DOA estimates for each value of the CDR. To do this, we perform a training phase in a controlled simulated environments as follows:

1. Simulate a short signal segment by convolving white Gaussian noise signal with an anechoic room impulse response, and add an ideally diffuse noise signal simulated according to [34], with a specified signal-to-noise ratio (SNR). Note that although the CDR also depends on the reverberation from directional sources, the spatial properties of late reverberation closely resemble those of a diffuse sound field.
2. Repeat the simulation for different SNRs (we used the range $[-30, 30]$ dB, with steps of 5 dB), and for each simulation store the CDR estimates and the DOA estimates for each TF bin.
3. Make a histogram of the CDR estimates stored from all simulations and associate to each histogram bin the corresponding DOA estimates.
4. If the set of DOA estimates associated with the n-th histogram bin is $\Theta_n = \left\{\theta_1, \ldots, \theta_{L_n}\right\}$, a maximum

likelihood estimate of the concentration parameter for this histogram bin is obtained by first computing

$$r = \sqrt{\left(\frac{1}{L_n}\sum_{i=1}^{L_n}\cos\theta_i\right)^2 + \left(\frac{1}{L_n}\sum_{i=1}^{L_n}\sin\theta_i\right)^2}, \tag{25}$$

and using the following approximation (see ([29], Section 5.3.1) for details)

$$\kappa_{n,\mathrm{ML}} = \begin{cases} 2\,r + r^3 + \frac{5}{6}\,r^5 & \text{if } r < 0.53, \\ -0.4 + 1.39\,r + \frac{0.43}{1-r} & \text{if } 0.53 \le r < 0.85 \\ \frac{1}{2(1-r)} & \text{if } r \ge 0.85. \end{cases} \tag{26}$$

5. For each histogram bin n, store the CDR value of the bin centre and the corresponding ML estimate of the concentration parameter as a pair $(\Gamma_n, \kappa_{n,\mathrm{ML}})$.

Following this data-driven procedure, we have experimentally found a correspondence between the CDR estimates and the concentration parameter κ. Given the pairs $\left(\Gamma_n, \kappa_{n,\mathrm{ML}}\right)$, we can now determine the parameters of the sigmoid-like mapping function. First, note that although in theory the logarithmic range of the CDR is $[-\infty, \infty]$, in practice, the CDR estimators saturate and are limited to a relatively small range of values around 0 dB. For our particular estimator, we observed that the range of estimates was $[-10, 20]$ dB, which allows us to determine the maximum value $l_{\max}$ of the concentration parameter by observing the values of $\kappa_{n,\mathrm{ML}}$ for the histogram bins where $\Gamma_n \approx 20$ dB. To find the parameter c that determines the offset along the $\widehat{\Gamma}$ axis, we note that for any value of ρ, the value of $\widehat{\Gamma}$ for which the resulting κ is exactly in the midpoint of its range $[0, l_{\max}]$, satisfies $\widehat{\Gamma} = 10\,c$. Therefore, by looking for the pair $(\Gamma_n, \kappa_{n,\mathrm{ML}})$ in our training results where $\kappa_{n,\mathrm{ML}}$ is as close as possible to $l_{\max}/2$, we can use the corresponding Γ_n to compute the parameter c. Having fixed c and $l_{\max}$ and noting that due to the aforementioned saturation of the CDR estimator, the concentration parameter is approximately 0 for $\Gamma_n \approx -10$, there is only a small range of values for ρ which satisfy the constraints on the maxima and the minima of the sigmoid-like function (i.e., $f(-10) \approx 0$ and $f(20) \approx l_{\max}$). This range was $\rho \in [0.2, 2]$ in our case, and the best fit for ρ can be easily found by visual inspection of the curves obtained by substituting several values for ρ from this range. The above described procedure for our data resulted in $l_{\max} = 8$, $c = 1.5$, and $\rho = 1.2$, which we kept constant for all the experiments.

6 Application to informed spatial filtering

We use the detector proposed in Section 5 to obtain the PSD matrix estimates $\widehat{\Phi}_{\boldsymbol{u}}$ and $\widehat{\Phi}_{\boldsymbol{s}}$, and the RTF vector estimate $\hat{\boldsymbol{g}}_1$. The RTF vector can be obtained using the generalized eigenvalue decomposition (GEVD) of the matrix pencil $\left(\widehat{\Phi}_{\boldsymbol{s}+\boldsymbol{v}}(t,k), \widehat{\Phi}_{\boldsymbol{v}}(t,k)\right)$ [35], such that if $\boldsymbol{u}(t,k)$ denotes the generalised eigenvector corresponding to the maximum generalised eigenvalue, $\hat{\boldsymbol{g}}_1(t,k)$ is given by

$$\hat{\boldsymbol{g}}_1(t,k) = \frac{\widehat{\Phi}_{\boldsymbol{v}}(t,k)\,\boldsymbol{u}(t,k)}{\boldsymbol{e}_1^{\mathrm{T}}\,\widehat{\Phi}_{\boldsymbol{v}}(t,k)\,\boldsymbol{u}(t,k)}, \text{with } \boldsymbol{e}_1 = [1, 0 \ldots, 0]\,, \tag{27}$$

where the denominator normalises the first entry to one. The PSD matrices are computed using the standard recursive updates

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{s}+\boldsymbol{v}}(t,k) = \alpha_s(t,k)\,\widehat{\boldsymbol{\Phi}}_{\boldsymbol{s}+\boldsymbol{v}}(t-1,k) + (1-\alpha_s(t,k))\;\boldsymbol{y}(t,k)\,\boldsymbol{y}^{\mathrm{H}}(t,k) \tag{28a}$$

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(t,k) = \alpha_u(t,k)\,\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(t-1,k) + (1-\alpha_u(t,k))\;\boldsymbol{y}(t,k)\,\boldsymbol{y}^{\mathrm{H}}(t,k), \tag{28b}$$

where the averaging parameters are computed using the output of the detector in (22) as follows

$$\begin{aligned} \alpha_s(t,k) &= 1 + \mathcal{I}_{\mathcal{H}_s}(t,k)\,(\tilde{\alpha}_s - 1)\,, \\ \alpha_u(t,k) &= 1 + \mathcal{I}_{\mathcal{H}_u}(t,k)\,(\tilde{\alpha}_u - 1)\,, \end{aligned} \tag{29}$$

where the values $\tilde{\alpha}_s, \tilde{\alpha}_u \in [0,1)$ are pre-defined constants determining the effective range of the averaging parameters, i.e., $\alpha_s \in [\,\tilde{\alpha}_s, 1]$ and $\alpha_u \in [\,\tilde{\alpha}_u, 1]$. The noise PSD matrix is computed using similar recursion as (28), however, the parameter α_v is computed using the Gaussian model-based SPP [15, 21] as follows

$$\alpha_v(t,k) = \tilde{\alpha}_v + p(\mathcal{H}_{si} \,|\, \boldsymbol{y}(t,k))\,(1-\tilde{\alpha}_v)\,. \tag{30}$$

where $\tilde{\alpha}_v \in [0,1)$ is a pre-defined constant. In contrast to (29), the noise averaging parameter (30) leads to a soft recursive update, as often done when the undesired signal is stationary [15, 20, 21].

Given the estimates $\widehat{\Phi}_{\boldsymbol{u}}$ and $\hat{\boldsymbol{g}}_1$, an informed MVDR filter to extract the desired source in (3) is computed as

$$\boldsymbol{w}_{\mathrm{mvdr}}(t,k) = \frac{\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}^{-1}(t,k)\,\hat{\boldsymbol{g}}_1(t,k)}{\hat{\boldsymbol{g}}_1^{\mathrm{H}}(t,k)\,\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(t,k)^{-1}\hat{\boldsymbol{g}}_1(t,k)}. \tag{31}$$

Note that, as the desired source PSD matrix $\boldsymbol{\Phi}_{\boldsymbol{s}}$ is of rank one, the MVDR filter can be expressed only in terms of $\boldsymbol{\Phi}_{\boldsymbol{u}}$ and $\boldsymbol{\Phi}_{\boldsymbol{s}}$ [4]. However, when the undesired signal is non-stationary, $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{s}}$ contains errors which can be detrimental to the signal quality when used in such filter formulation. Therefore, we first estimate the desired signal RTF vector $\hat{\boldsymbol{g}}_1(t,k)$, and compute an MVDR filter using (31). The complete source extraction framework is summarized in Fig. 2. In addition, the DSPP $p\left(\mathcal{H}_s \,|\, \hat{\theta}_{tk}\right)$ can be applied as a multiplicative factor to the output of the MVDR filter, i.e.,

$$\widehat{S}_m(t,k) = p\left(\mathcal{H}_s \,|\, \hat{\theta}_{tk}\right) \cdot \boldsymbol{w}_{\mathrm{mvdr}}^{\mathrm{H}}(t,k)\,\boldsymbol{y}(t,k), \tag{32}$$

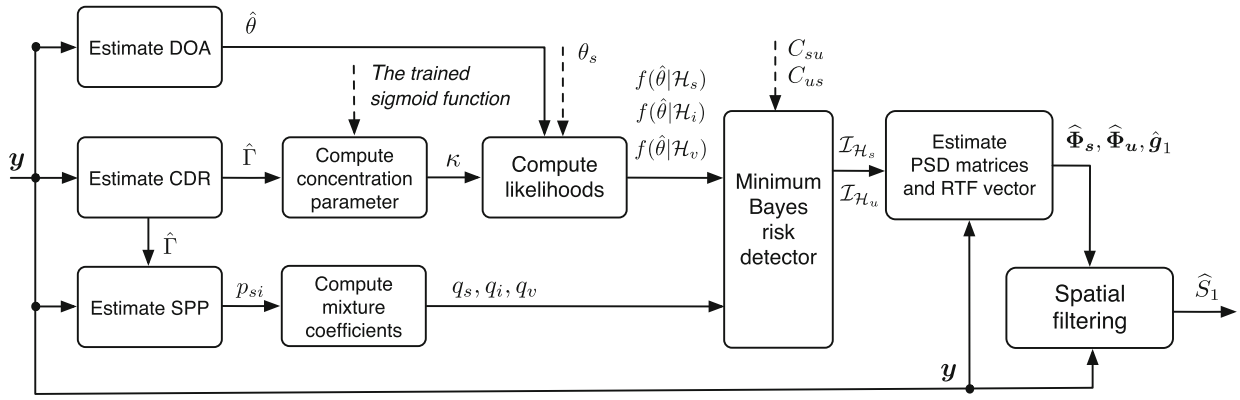

Fig. 2 Processing blocks of the DOA-informed spatial filtering framework

which has a similar role as the single-channel DOA-based gain in [23] and the DOA-based TF mask common for source separation [36]. Applying the DSPP as a multiplicative factor provides additional undesired signal reduction, however, when inaccurately estimated, it causes audible distortion to the desired signal. This is further evaluated in the experiments in Section 7.2.

7 Performance evaluation

To evaluate the proposed system, measurements were performed in a room with $T_{60} = 0.16$ s using the setup illustrated in Fig. 3. We simultaneously recorded three sources captured at three arrays, and the objective is to extract each source using the nearest array to that source, while reducing the remaining two sources. Each array (uniform circular with a diameter of 3 cm) consisted of three omnidirectional DPA microphones (model DPA d:screet SMK-SC4060). The distance between each source and the nearest array is 0.7 m, and each source is extracted using the nearest array only (hence the framework is implemented with only three microphones). To generate background noise, ten loudspeakers were placed facing the walls and babble speech signals were convolved with the measured room impulse responses (RIRs) of the ten loudspeakers. Finally, clean speech signals were convolved with the measured RIRs for the three sources in Fig. 3, and added with the babble noise signal, appropriately scaled to provide a desired SNR (exact values are given in the experiment descriptions). In addition, measured sensor noise was added with an SNR of 35 dB for all experiments. The SNRs were computed segmentally over 30 ms signal segments, as the power ratio of the desired speech signal and the background (or sensor) noise captured at the reference microphone. The final SNRs indicated in the experiments are obtained by averaging across all segments with SNR in the range [-20,40].

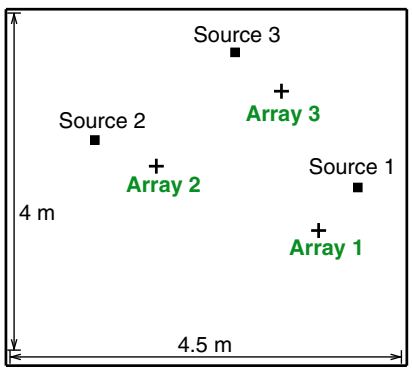

Fig. 3 Measurement setup

To evaluate the performance for different reverberation, simulated data was used. RIRs were computed using the simulator in [37]. Diffuse noise was simulated as described in [34] and the microphone signals were obtained by adding the speech signals convolved with the RIRs, the diffuse noise signal, and spatially and temporally uncorrelated noise signal. The processing was done at a sampling rate of 16 kHz, with an STFT frame length of 64 ms with 50% overlap, windowed by a Hamming window. Unless stated otherwise, the DOA estimator with instantaneous phase differences, described in Section 3.1 was employed.

7.1 Detector evaluation in terms of receiver operating characteristics (ROC)

We compare a minimum Bayes risk detector obtained using the proposed DOA model-based DSPP, to the one obtained using the Gaussian signal model, as in [22]. The false positive rate (FPR) and the false negative rate (FNR) are defined as

$$\begin{aligned} \text{FPR} &= \textstyle\sum_{t,k} [\mathcal{H}_s{=}1 \wedge \mathcal{H}_{\text{ideal}}{=}0] / \sum_{t,k} [\mathcal{H}_{\text{ideal}}{=}0], \\ \text{FNR} &= \textstyle\sum_{t,k} [\mathcal{H}_s{=}0 \wedge \mathcal{H}_{\text{ideal}}{=}1] / \sum_{t,k} [\mathcal{H}_{\text{ideal}}{=}1], \end{aligned} \tag{33}$$

where $\sum_{t,k} [\cdot\,]$ denotes summation of the value of the logical expression in the brackets. The ideal detector $\mathcal{H}_{\text{ideal}}$ is

obtained by comparing the spectra of the desired signal to the sum of all undesired signals, namely,

$$\mathcal{H}_{\text{ideal}}(t,k) = \begin{cases} 1, \text{if } |S_m(t,k)|^2 > \left|I_m(t,k) + V_m(t,k)\right|^2 \\ 0, \text{otherwise.} \end{cases} \tag{34}$$

The ROC curves are obtained by computing the FPR and FNR as $\frac{C_{su}}{C_{us}}$ varies from 0 to ∞. The FPR and FNR are computed for the three sources (French female, English female, and English male) during 20 s of multi-talk. The average FPR and FNR used for the ROC curves are obtained by averaging over the segments for each of the three sources (hence over 60 s of speech in total). In all experiments, the desired-to-interfering speech ratio (DSIR) was in the range [5, 8] dB. The DSIR for each source is computed at one of the microphones from the nearest array.

7.1.1 Experiment 1

In this experiment, we evaluate the detection accuracy for different noise and reverberation levels. To investigate the effect of background noise, the experiment was repeated for input SNR of 3, 8, 13, and 18 dB using the measurement data. The ROC curves for the different SNRs are shown in Fig. 4a, for the two detectors. The detection accuracy is not notably affected by the SNR, as shown by the overlapping ROC curves and only a minor increase in the error rate can be observed for decreasing SNR. It is worthwhile noting that although in non-stationary scenarios, both types of errors are critical for the extracted signal quality [28], false positives are more detrimental as they lead to errors in the RTF vector and distortion of the desired signal. In contrast, if the RTF vector is accurate, which can be achieved if the FPR is low, false negatives do not affect the performance.

To evaluate the detectors for different reverberation levels, the setup shown in Fig. 3 was simulated for T_{60} of 0.2 s, 0.35 s, 0.5 s, and 0.65 s and diffuse babble noise with an SNR of 22 dB. As shown in Fig. 4c, reverberation has a stronger effect on the ROC than the noise, as the curves shift more notably with increasing reverberation. However, the proposed detector clearly outperforms the signal-model based detector in all cases.

7.1.2 Experiment 2

In this experiment, we simulated scenarios with one desired source and one interferer, for different angular separations between the desired source and the interferer, namely, 160°, 95°, 50°, 25°, and 0°. In all cases, the desired source is located at 0.7 m from the array, whereas the interferer at 1.5 m from the array. The reverberation time was $T_{60} = 0.35$ s and diffuse babble noise with an SNR of 22 dB and uncorrelated sensor noise with an SNR of 35 dB were added. As expected, with decreasing angular separation, the detection accuracy deteriorates, as visible in Fig. 4b. Note that even when the desired and the undesired source have equal DOA, the detector provides good accuracy due to the fact that the desired signal is stronger than the interferer at its respective nearest array. Another reason is that the CDR in interferer-dominated TF bins is lower than the CDR in desired signal-dominated TF bins, hence allowing the CDR-controlled concentration κ to aid the detection even when the sources have equal DOA.

7.1.3 Experiment 3

The detector ROC curves obtained with the two DOA estimators discussed in Section 3, the one with instantaneous, and the one with time-averaged phase differences, are illustrated in Fig. 4d. Although time averaging of the phase differences generally provides less noisy DOA estimates (i.e., smoother across the TF spectrum), the detector performance is better when instantaneous DOA estimates are used.

7.2 Objective evaluation of extracted signals

To estimate the PSD matrices and the RTF vector, we computed a Bayes detector according to (22), where we used $C_{su} = 1$ and $C_{us} = 2$. These costs were chosen after investigating the objective performance measures and the results of informal listening tests in different acoustic conditions, where they proved to achieve the best performance from all (C_{su}, C_{us}) pairs across the ROC. The chosen costs resulted in an FPR of 0.1 and an FNR of 0.9 on average (across the different experiments), which corroborates the observation made in Section 7.1.1, that the FPR needs to be very low in order to ensure a good extracted signal quality. The averaging constants for the PSD matrices were $\tilde{\alpha}_v = 0.95$, $\tilde{\alpha}_s = \tilde{\alpha}_u = 0.92$ (corresponding to time constants of 0.62 and 0.38 s). The performance was evaluated in terms of segmental noise reduction (NR), segmental interference reduction (IR), speech distortion (SD) index ν_{sd}, PESQ score improvement Δ_{PESQ} [38], and improvement of the short-time objective intelligibility (STOI) score [39], Δ_{STOI}. Five spatial filtering frameworks are evaluated: (i) An oracle MVDR filter, where the PSD matrices are computed using recursive averaging with an ideal detector, denoted by $\mathcal{D}_{\text{id}}$, (ii) a DSB steered to the desired source DOA, (iii) an MPDR filter steered to the desired source DOA, iv) an informed MVDR filter obtained using the Gaussian signal model-based detector with a DOA-based a priori SPP, denoted by $\mathcal{D}_{\text{sm}}$ and v) an informed MVDR filter obtained using the proposed DOA-based detector, denoted by $\mathcal{D}_{\text{dm}}$.

7.2.1 Experiment 1

This experiment is performed using measured data for two SNR conditions, indicated in Table 1. Although $\mathcal{D}_{\text{sm}}$

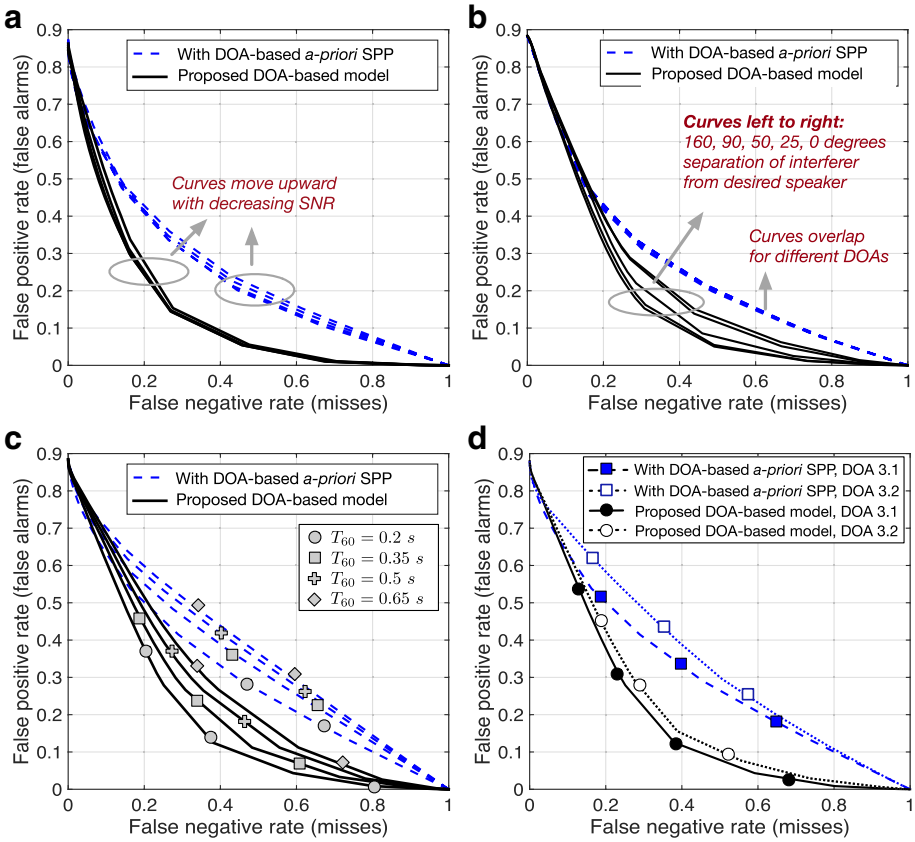

Fig. 4 ROC curves. Comparison of the Gaussian model-based detector with a DOA-based a-priori SPP, and the proposed DOA model-based detector, in different scenarios. **a** Vary SNR, measured data. **b** Vary interferer DOA, simulated data. **c** Vary reverberation time simulated data. **d** Compare DOA estimators, measured data

and $\mathcal{D}_{\text{dm}}$ perform similarly in terms of NR, $\mathcal{D}_{\text{dm}}$ offers by of up to 8 dB better IR than $\mathcal{D}_{\text{sm}}$ for SNR of 10 dB, and up to 6 dB better IR for SNR of 2 dB. The SD index is lower for the proposed $\mathcal{D}_{\text{dm}}$ in all cases. The better performance of the proposed system is due to the higher accuracy of the DOA-based detector compared to the Gaussian model-based one. The improvement in PESQ and STOI scores at the output of the proposed system with respect to the unprocessed signal is notably higher than the improvement offered by the other systems. The severe distortions of MPDR filter often result in lower PESQ and STOI scores than the unprocessed signal, as visible in Table 1.

7.2.2 Experiment 2

In this experiment, the proposed system $\mathcal{D}_{\text{dm}}$ and the output of the DSB are multiplied by the a posteriori DSPP. A system where DOA-based DSPP is applied at the output of a fixed spatial filter is proposed in [23], and the goal of the current experiment is to confirm that the benefit of the DSPP is even larger when it is used in combination with a data-dependent, informed spatial filter, rather than a fixed spatial filter. The experiment is repeated with the two DOA estimators discussed in Section 3.

The results in Table 2 are shown for average input SNR of 10 dB and confirm that the informed MVDR filter outperforms the DSB when the DSPP-based mask is applied after spatial filtering. We also note that multiplying the DSPP is critical as it introduces SD, even though, the NR and IR are significantly improved. The choice whether to multiply the MVDR output by the DSPP, depends on the accuracy of the DSPP and the importance of undistorted speech for a given application. Finally, note that the system with the DOA estimator based on instantaneous phase differences slightly outperforms the one with cross PSD phase differences, which is consistent with the detection performance evaluation in Fig. 4d. Time-averaged phase differences result in time-smoothing of the DSPP, which can distort the speech onsets and degrade the overall performance.

7.2.3 Experiment 3

In this experiment, we investigate the performance for varying angular separation between the desired and

Table 1 Results for `Source1` (top), `Source2` (middle), and `Source3` (bottom)

	Average input SNR 10 dB					Average input SNR 2 dB				
	DSB	MPDR	$\mathcal{D}_{\text{ld}}$	$\mathcal{D}_{\text{sm}}$	$\mathcal{D}_{\text{dm}}$	DSB	MPDR	$\mathcal{D}_{\text{ld}}$	$\mathcal{D}_{\text{sm}}$	$\mathcal{D}_{\text{dm}}$
NR	1.4	6.4	**7.5**	6.1	**7.2**	1.4	7.0	**8.9**	7.6	**9.2**
IR	1.9	8.1	**14.4**	5.5	**12.9**	1.9	8.0	**12.8**	6.3	**11.8**
ν_{sd}	**0.03**	0.25	**0.02**	0.11	0.06	**0.03**	0.28	**0.03**	0.08	0.07
Δ_{PESQ}	0.02	0.17	**0.75**	-0.02	**0.68**	0.02	0.12	**0.59**	0.17	**0.53**
Δ_{STOI}	0.01	0.07	**0.17**	-0.01	**0.15**	0.01	0.05	**0.18**	0.04	**0.16**
NR	0.9	4.3	**7.1**	**7.0**	6.0	0.9	2.8	**11.7**	**7.5**	6.1
IR	0.5	2.9	**15.7**	5.3	**13.9**	0.5	2.1	**13.9**	6.2	**11.0**
ν_{sd}	0.06	0.19	**0.03**	0.11	**0.03**	0.06	0.17	**0.03**	0.10	**0.03**
Δ_{PESQ}	0.03	-0.01	**0.85**	0.20	**0.76**	0.03	0.04	**0.73**	0.31	**0.51**
Δ_{STOI}	0.01	-0.03	**0.13**	0.01	**0.12**	0.01	-0.02	**0.16**	0.05	**0.11**
NR	1.4	**10.6**	**6.4**	5.6	6.5	1.4	**12.8**	**8.4**	7.7	8.1
IR	1.7	**15.9**	**13.8**	5.6	11.8	1.7	**15.0**	**11.2**	6.4	10.4
ν_{sd}	**0.03**	0.87	**0.02**	0.09	0.04	**0.04**	0.81	**0.02**	0.07	0.04
Δ_{PESQ}	0.02	-1.10	**0.61**	0.20	**0.53**	0.02	-0.63	**0.51**	0.31	**0.47**
Δ_{STOI}	0	-0.37	**0.07**	0.01	**0.07**	0.01	-0.25	**0.10**	0.06	**0.09**

The segmental DSIR at the reference microphone of each Source is 6.8, 5.7, and 8 dB. The result with the oracle detector and the second best result are indicated in bold

the interfering source. Signals were simulated with reverberation time $T_{60} = 0.2$ s and $T_{60} = 0.4$ s. The distances of the desired source and the interferer from the array were 0.7 and 1.5 m, respectively, and the results in Fig. 5 are averaged over three experiments with different locations of the constellation. The NR and SD of the different frameworks are rather unaffected by the angular separation, while the IR decreases with decreasing angular separation. As the angular separation decreases, $\mathcal{D}_{\text{dm}}$ and $\mathcal{D}_{\text{sm}}$ achieve similar IR, due to the fact that in both cases, the performance is limited by the spatial resolution of the array.

Table 2 Results when the spatial filter output is multiplied by the estimated DSPP

	DSB-inst	DSB-cPSD	$\mathcal{D}_{\text{dm}}$-inst	$\mathcal{D}_{\text{dm}}$-cPSD
NR	14.3	14.2	**19.3**	18.9
IR	15.7	15.6	**24.9**	24.3
ν_{sd}	0.36	0.36	**0.33**	0.34
Δ_{PESQ}	0.52	0.47	**0.88**	0.79
Δ_{STOI}	0.10	0.08	**0.13**	0.11
NR	9.1	8.2	**15.7**	15.4
IR	12.3	11.5	**25.0**	23.1
ν_{sd}	**0.12**	0.13	0.15	0.15
Δ_{PESQ}	0.67	0.60	**1.07**	1.00
Δ_{STOI}	0.08	0.07	**0.11**	0.10
NR	12.2	11.4	**16.7**	16.0
IR	14.0	13.2	**22.5**	21.6
ν_{sd}	0.27	0.26	0.24	**0.23**
Δ_{PESQ}	0.54	0.48	**0.87**	0.84
Δ_{STOI}	0.04	0.02	**0.05**	0.04

`Source1` (top), `Source2` (middle), and `Source3` (bottom). The segmental DSIR at the reference microphone of each source is 6.8, 5.7, and 8.0 dB, "-inst" indicates the DOA estimator with instantaneous phase differences, while "-cPSD" the one with cross-PSD phase differences. The best result is indicated in bold

7.2.4 Experiment 4

An important motivation for the current work was to mitigate the sensitivity to DOA mismatch typical for the MPDR beamformer, where DOA errors lead to severe distortion of the desired signal. Provided that the desired signal detector is accurate besides the presence of small DOA errors, the RTF vector and the undesired signal PSD matrix can be estimated and the desired signal can be extracted with a good quality using an informed MVDR filter. In Fig. 6, the DOA mismatch is varied such that the system is given a wrong information about the true source DOA, with an error from 1 to 19°, including an error-free case. The difference between the DOA of the desired source and the interferer is 100 degrees. In Fig. 6a, the SD index and the PESQ improvement are shown on the y-axis, whereas in Fig. 6b, the NR and the IR are shown. Notably, besides minor performance loss as the mismatch angle increases, the ISF is very robust to DOA errors, which is crucial for practical applications where the DOA might be given only approximately.

8 Conclusions

We addressed the problem of source extraction in the presence of background noise and speech interferers.

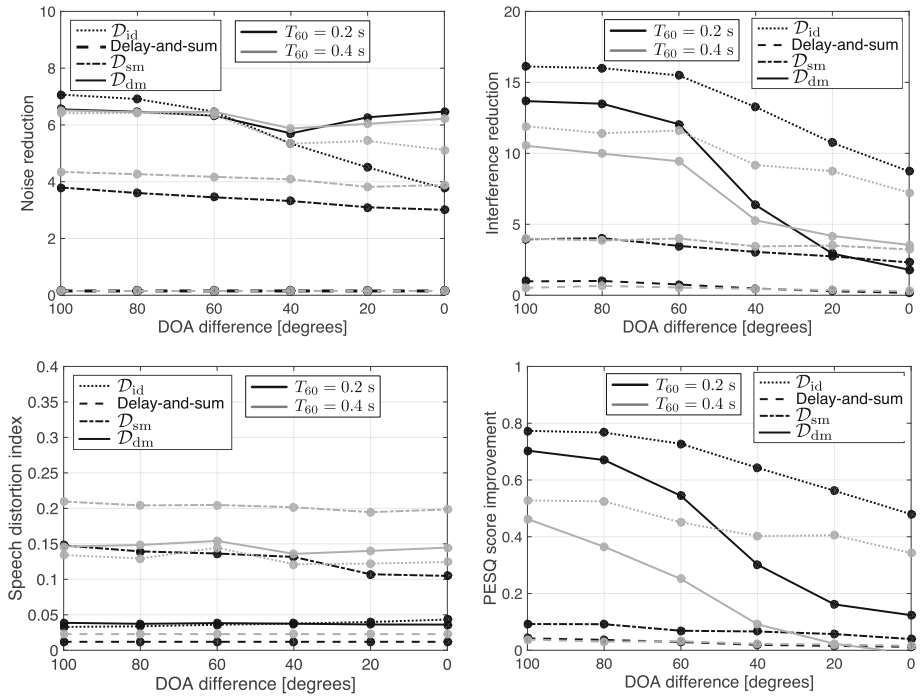

Fig. 5 Evaluation of the extracted signal quality as a function of the DOA difference between the desired and the undesired source. Input SIR = 9 dB and input SNR = 14 dB

The DOA of the desired source was assumed to be approximately known, while the number and locations of the interferers were unknown. Designing robust spatial filters is a challenging task in such scenarios, as the PSD matrix of the undesired speech signals needs to be estimated from the data. We proposed an informed spatial filtering framework, where the first step is to design appropriate desired signal detector. We discussed and experimentally showed that the commonly used Gaussian signal model-based detector is not suitable when the undesired signals contain speech. Therefore, we proposed a DOA model-based detector, where narrowband DOA estimates are used for discrimination of desired and undesired speakers, while the Gaussian signal model aids the detection of noisy TF bins. The performance of the detector was evaluated in terms of ROC curves, and by objective evaluation of the extracted signals when the detector is applied for PSD matrix estimation in an informed MVDR filtering framework.

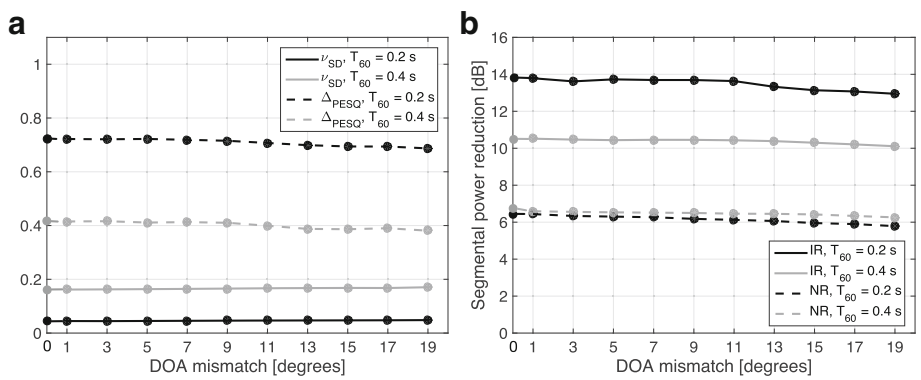

Fig. 6 Evaluation of extracted signal quality in the presence of DOA mismatch. The *y*-axis on the left plot illustrates the SD index and the PESQ improvement, while the *y*-axis on the right plot illustrates the NR and IR. Input SIR = 9 dB and input SNR = 14 dB **a** Speech distortion and PESQ improvement **b** Noise reduction and interference reduction

Acknowledgements
The authors would like to thank the Editorial board and the Reviewers for considering and revising this manuscript.

Funding
No funding was received or used to prepare this manuscript.

Authors' contributions
Both authors had significant contribution to the development of early ideas and design of the final algorithms. Throughout all stages, the authors discussed the importance and the quality of the algorithms and the structure of the manuscript. Both authors read and approved the submitted version of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 24 March 2017 Accepted: 10 August 2017
Published online: 22 August 2017

References

1. JL Flanagan, JD Johnston, R Zahn, GW Elko, Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. **5**(78), 1508–1518 (1985)
2. S Doclo, M Moonen, Superdirective beamforming robust against microphone mismatch. IEEE Signal Process. Lett. **15**(2), 617–631 (2007)
3. J Benesty, J Chen, Y Huang, *Microphone Array Signal Processing*. (Springer, Berlin, 2008)
4. M Souden, J Benesty, S Affes, On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio Speech Lang. Process. **18**(2), 260–276 (2010)
5. KL Bell, Y Ephraim, HL Van Trees, A Bayesian approach to robust adaptive beamforming. IEEE Trans. Signal Process. **48**(2), 386–398 (2000)
6. CJ Lam, AC Singer, Bayesian beamforming for DOA uncertainty: theory and implementation. IEEE Trans. Signal Process. **54**(11), 4435–4445 (2006)
7. Y Grenier, A microphone array for car environments. Speech Commun. **12**, 25–39 (1993)
8. MK Buckley, Spatial/spectral filtering with linearly constrained minimum variance beamformers. IEEE Trans. Acoust. Speech Signal Process. **35**(3), 249–266 (1987)
9. CA Anderson, PD Teal, MA Poletti, Spatially robust far-field beamforming using the von Mises(-Fisher) distribution. IEEE Trans. Acoust. Speech Signal Process. **23**(12), 2189–2197 (2015)
10. O Hoshuyama, A Sugiyama, A Hirano, A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Trans. Signal Process. **47**(10), 2677–2684 (1999)
11. BJ Yoon, I Tashev, A Acero, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Robust adaptive beamforming algorithm using instantaneous direction of arrival with enhanced noise suppression capability, (HI, USA, 2007), pp. 133–136
12. O Hoshuyama, B Begasse, A Sugiyama, A Hirano, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. A real time robust adaptive microphone array controlled by an SNR estimate (ICASSP, Seattle, 1998), pp. 3605–3608
13. M Taseska, EAP Habets, Informed spatial filtering with distributed arrays. IEEE Trans. Audio Speech Lang. Process. **22**(7), 1195–1207 (2014)
14. M Taseska, EAP Habets, Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1291–1304 (2016)
15. M Souden, J Chen, J Benesty, S Affes, An integrated solution for online multichannel noise tracking and reduction. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2159–2169 (2011)
16. T Higuchi, N Ito, T Yoshioka, T Nakatani, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, (Shanghai, 2016)
17. S Affès, Y Grenier, A signal subspace tracking algorithm for microphone array processing of speech. IEEE Trans. Speech Audio Process. **5**(5), 425–437 (1997)
18. S Gannot, D Burshtein, E Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
19. D Van Compernolle, Adaptive filter structures for enhancing cocktail party speech from multiple microphone recordings. Colloque sur le traitement du signal et des images, 513–516 (1989). http://documents.irevues.inist.fr/handle/2042/11518?show=full
20. I Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
21. M Taseska, EAP Habets, in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*. MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator, (2012)
22. DP Jarrett, EAP Habets, PA Naylor, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation, (Vancouver, 2013)
23. I Tashev, A Acero, in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Microphone array post-processor using instantaneous direction-of-arrival, (Paris, 2006)
24. M Taseska, EAP Habets, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Minimum Bayes risk signal detection for speech enhancement based on a narrowband DOA model, (Brisbane, 2015)
25. R Roy, T Kailath, ESPRIT - estimation of signal parameters via rotational invariance techniques. IEEE Trans. Acoust. Speech Signal Process. **37**, 984–995 (1989)
26. S Araki, H Sawada, R Mukai, S Makino, DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors. J. Signal Process. Syst. **63**, 265–275 (2011)
27. O Thiergart, W Huang, EAP Habets, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. A low complexity weighted least squares narrowband DOA estimator for arbitrary array geometries (ICASSP, Shanghai, 2016), pp. 340–344
28. H Cox, RM Zeskind, MM Owen, Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. **35**(10), 1365–1376 (1987)
29. KV Mardia, PE Jupp, *Directional Statistics*. (Wiley-Blackwell, New York, 1999)
30. Stegun IA, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. (M Abramowitz, ed.) (United States Department of Commerce, USA, 1972), p. 1046
31. S Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. (Prentice-Hall, Inc., NJ, USA, 1998)
32. M Souden, J Chen, J Benesty, S Affès, Gaussian model-based multichannel speech presence probability. IEEE Trans. Audio Speech Lang. Process. **18**(5), 1072–1077 (2010)
33. O Thiergart, G Del Galdo, EAP Habets, On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation. J. Acoust. Soc. Am. **132**(4), 2337–2346 (2012)
34. EAP Habets, I Cohen, S Gannot, Generating nonstationary multisensor signals under a spatial coherence constraint. J. Acoust. Soc. Am. **124**(5), 2911–2917 (2008)
35. A Krueger, E Warsitz, R Haeb-Umbach, Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation. IEEE Trans. Audio Speech Lang. Process. **19**(1), 206–219 (2011)
36. S Araki, H Sawada, R Mukai, S Makino, in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. A novel blind source separation method with observation vector clustering (IWAENC, Eindhoven, 2005)
37. EAP Habets, Room impulse response generator. Technical report, Technische Universiteit Eindhoven (2006)
38. ITU-T: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs
39. CH Taal, RC Hendriks, R Heusdens, J Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)