A Compact Noise Covariance Matrix Model for MVDR Beamforming

Alastair H. Moore, Sina Hafezi, Rebecca R. Vos, Patrick A. Naylor, *Fellow, IEEE*, and Mike Brookes, *Life Member, IEEE*

***Abstract*—Acoustic beamforming is routinely used to improve the SNR of the received signal in applications such as hearing aids, robot audition, augmented reality, teleconferencing, source localisation and source tracking. The beamformer can be made adaptive by using an estimate of the time-varying noise covariance matrix in the spectral domain to determine an optimised beam pattern in each frequency bin that is specific to the acoustic environment and that can respond to temporal changes in it. However, robust estimation of the noise covariance matrix remains a challenging task especially in non-stationary acoustic environments. This paper presents a compact model of the signal covariance matrix that is defined by a small number of parameters whose values can be reliably estimated. The model leads to a robust estimate of the noise covariance matrix which can, in turn, be used to construct a beamformer. The performance of beamformers designed using this approach is evaluated for a spherical microphone array under a range of conditions using both simulated and measured room impulse responses. The proposed approach demonstrates consistent gains in intelligibility and perceptual quality metrics compared to the static and adaptive beamformers used as baselines.**

***Index Terms*—Beamforming, speech enhancement, covariance matrix estimation, spatial filtering, spherical microphone arrays, adaptive beamforming, microphone array, MVDR, MPDR.**

I. Introduction

THE use of microphone arrays and acoustic beamforming has become routine in devices such as cellphones, hearing aids, virtual assistants, teleconferencing and robot audition [1]–[5]. These devices share a need to acquire speech from a target talker in the presence of interfering noise from other sound sources. In many situations, especially those in which the talker is far from the microphones, the signal-to-noise ratio (SNR) of the received microphone signals will be inadequate and in these cases the spatial discrimination provided by beamforming allows the SNR to be improved with little or no distortion of the target speech. Existing beamformers perform well in laboratory conditions but may perform less well in real-world situations containing multiple interfering sound sources whose locations and characteristics vary with time. For clarity, the mathematical symbols used in the remainder of this introduction will only be given brief definitions here and will be fully defined in Section II.

Acoustic beamformers are conveniently implemented in the short time Fourier transform (STFT) domain [6]. In each time-frequency (TF) cell, the complex-valued output of the beamformer is given by $\mathbf{w}^H\mathbf{y}$ where $\mathbf{w}$ is a weight vector that determines the beamformer properties and the elements of $\mathbf{y}$ are the microphone signals. When choosing $\mathbf{w}$, a common goal is to maximize the SNR of the beamformer output signal subject to the so-called distortionless response constraint that the gain of the target be constant across all time-frequency cells. To achieve this, it is convenient to define the steering vector, $\mathbf{d}$, whose elements are proportional to the acoustic transfer functions between the target source and each of the microphones. With this definition, the optimum weight vector is given by $\mathbf{w} = (\mathbf{d}^H\mathbf{R}^{-1}\mathbf{d})^{-1}\mathbf{R}^{-1}\mathbf{d}$ where $\mathbf{R}$ is the covariance matrix either of the received microphone signals or, alternatively, of the noise component within those signals. We refer to these two alternative choices for $\mathbf{R}$ as the signal covariance matrix (SCM) and the noise covariance matrix (NCM) respectively and to the corresponding beamformers as the minimum power distortionless response (MPDR) and minimum variance distortionless response (MVDR) beamformers[1] [7], [8]. It can be shown using the matrix inversion lemma [9] that the MPDR and MVDR beamformers are identical provided that the noise is uncorrelated with the target and that the steering vector, $\mathbf{d}$, is precisely correct [10]. The advantage of the MPDR beamformer is that the SCM is independent of the choice of target source and is normally easier to estimate than the NCM. Its disadvantage, however, is that if $\mathbf{d}$ is inaccurate, the MPDR beamformer performance degrades and target cancellation can occur [11], [12]. For this reason, if the NCM is known or can be robustly estimated, the MVDR beamformer is the preferred choice. A popular way of implementing the MPDR is via the generalized sidelobe canceller (GSC) [13] structure which converts the constrained optimization problem into an unconstrained one that can be implemented as a recursive algorithm.

In this work, we concentrate on the estimation of $\mathbf{R}$ and assume that the steering vector, $\mathbf{d}$, is either known a *priori* or else can be estimated [6], [14]–[18]. We note that it is sufficient to determine $\mathbf{R}$ to within a scalar multiple since multiplying it by

Manuscript received September 17, 2021; revised March 22, 2022; accepted May 22, 2022. Date of publication June 7, 2022; date of current version June 22, 2022. This work was supported by Engineering and Physical Sciences Research Council under Grant EP/S035842/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefan Bilbao. *(Corresponding author: Sina Hafezi.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, United Kingdom (e-mail: alastair.h.moore@imperial.ac.uk; s.hafezi14@imperial.ac.uk; r.vos@imperial.ac.uk; p.naylor@imperial.ac.uk; mike.brookes@imperial.ac.uk).

Digital Object Identifier 10.1109/TASLP.2022.3180671

[1]We note that beamformer nomenclature varies and that the term MVDR is used by some authors for either or both of the MPDR and MVDR beamformers.

an arbitrary scale factor leaves the weight vector, $\mathbf{w}$, unchanged. If the characteristics of the interference are known a *priori*, then the NCM can be taken to have a fixed, signal-independent form. Examples of this include the matched-filter beamformer [1], [19] in which the NCM is taken to be a multiple of the identity matrix and the superdirective beamformer [20] in which it is taken to be the covariance matrix corresponding to a spherically or cylindrically isotropic diffuse noise field.

If the interference is non-stationary, then a higher SNR improvement can be obtained by making $\mathbf{R}$ adaptive and signal-dependent in which case the SCM and/or NCM must be estimated from the microphone signals. For the MPDR beamformer, an estimate of the SCM may be obtained by smoothing the instantaneous outer product, $\mathbf{y}\mathbf{y}^H$, using a recursive average filter whose time-constant is short enough to follow temporal variations in the noise characteristics but long enough to obtain a good estimate [6]. For the MVDR beamformer, however, it is necessary to estimate the noise component in each microphone signal in order to determine the NCM. Typically this estimate is obtained by using a speech activity detection (SAD) algorithm to identify when the target talker is inactive or alternatively an algorithm to estimate the speech presence probability (SPP) in each time-frequency cell [6]. More recently, deep neural networks (DNNs) have been used to identify time-frequency cells that are dominated by noise [21], [22]. In [23] noise is estimated by steering a null at the target while [24] assumes a stationary noise component. Also in [25] a time-invariant homogeneous noise field is assumed. Other works such as [16]–[18] propose methods to jointly estimate the power spectral density (PSD) of the speech and reverberation together with the early relative transfer functions. The robust estimation of the noise component in a noisy speech signal, however, remains a challenging problem when the SNR is poor or when, as is often the case in practical applications, the interference arises from other talkers whose characteristics are similar to those of the target.

If the matrix $\mathbf{R}$ is ill-conditioned, the weight vector, $\mathbf{w}$, becomes sensitive to errors in the steering vector, $\mathbf{d}$. In this case, the robustness of the beamformer may be improved by adding a multiple of the identity matrix onto the SCM or the NCM before calculating $\mathbf{w}$. This diagonal loading procedure may be formulated as imposing a constraint on the condition number of $\mathbf{R}$, on the norm of $\mathbf{w}$ or on the white-noise gain of the array [26]. It is also equivalent, within a scale factor, to allowing for uncertainty in $\mathbf{d}$ when maximizing the SNR gain of the beamformer [27].

In this paper, instead of estimating the NCM directly, we first estimate a parametric model for the SCM and then modify this to obtain a model for the NCM. The compact model that we use for the SCM is defined by only a small number of parameters. These can be estimated reliably even in high levels of noise and can adapt rapidly in non-stationary environments. The underlying signal model is similar to that introduced in [28] and expresses the sound field at the microphone array as a sum of plane-waves (PWs) together with isotropic and spatially white noise components. An advantage of the proposed model is that estimating its parameters does not require the interferer directions to be determined explicitly although it is nevertheless straightforward to account for any interferers whose target vectors are known a *priori*. Once the model parameters for the SCM have been determined, an estimate of the NCM can be made by excluding any sources that lie close to the direction of the target talker. This provides the robustness advantages of the MVDR beamformer without the need to estimate the noise component in each microphone signal. The principle underlying this approach is similar to that of the GSC-based method in [29] where the blocking matrix is modified to suppress an angular region around the target direction. An earlier version of the proposed method was presented in [30]; the present paper includes a richer model of the sound field, an explicit estimation of the NCM from the SCM, more extensive evaluation and a new procedure for determining the model parameters with lower cost and improved performance.

The remainder of the paper is organized as follows: Section II formulates the problem, defines the parametric model for the SCM and explains the estimation of the NCM from the modelled SCM. Section III describes the procedure for estimating the parameters of the model under the assumption of single unknown plane wave direction. Using test data generated respectively with simulated and measured room impulse responses (RIRs), Sections IV and V compare the performance of beamformers designed with the proposed method against that of beamformers using other baseline techniques. Finally, conclusions are given in Section VI.

II. Formulation and Proposed Model

Assuming that all sources are in the far field, the time domain signal received at the m^{th} microphone, $y_m(t)$, can be decomposed into a sum of plane waves encompassing both direct-path propagation from sources as well as reflected components, together with a noise component

$$y_m(t) = \sum_{j=1}^{J} x_{m,j}(t) + v_m(t) \tag{1}$$

where t is the time index, J is the number of plane wave components, $x_{m,j}(t)$ is the signal due to the j^{th} plane wave component and $v_m(t)$ is sensor noise. In typical room acoustic scenarios, J will often be very large since it encompasses not only the direct path of the sources but also numerous reflections. Note that due to the coherence between (near-) simultaneously arriving reflections, the J plane waves do not necessarily have a direct, one-to-one, correspondence with reflection paths.

The microphone signals can equivalently be expressed in the STFT domain as

$$Y_m(\nu, \ell) = \sum_{j=1}^{J} X_{m,j}(\nu, \ell) + V_m(\nu, \ell) \tag{2}$$

where capitalized letters denote the STFT of the quantities denoted by the corresponding lowercase letters in (1), and ν and ℓ are the frequency and frame indices respectively. Since all frequency bins are processed independently, the dependence on ν will be omitted in the remainder of this paper for clarity.

Stacking the signals for all M microphones into a vector gives

$$\mathbf{y}(\ell) = \sum_{j=1}^{J} \mathbf{x}_j(\ell) + \mathbf{v}(\ell) \tag{3}$$

where $\mathbf{y}(\ell) = \begin{bmatrix} Y_1(\ell) & \dots & Y_M(\ell) \end{bmatrix}^T$ and $(\cdot)^T$ denotes the transpose. Other signals are stacked similarly.

In the design of a beamformer, the weight vector, $\mathbf{w}(\ell)$ can be calculated according to the MVDR formula [8], in which the dependence on ℓ has been omitted for clarity,

$$\mathbf{w} = \frac{\mathbf{R}_{\varepsilon}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{\varepsilon}^{-1}\mathbf{d}} \tag{4}$$

$$\text{with} \quad \mathbf{R}_{\varepsilon} = \hat{\mathbf{R}}_{\mathcal{N}} + \varepsilon\mathbf{I} \tag{5}$$

$$\text{and} \quad \varepsilon = \max\left(\frac{\lambda_{\max} - \kappa_0\lambda_{\min}}{\kappa_0 - 1}, 0\right), \tag{6}$$

where $\mathbf{d}$ is the steering vector, $\hat{\mathbf{R}}_{\mathcal{N}}$ is an estimate of the NCM, $(\cdot)^H$ and $(\cdot)^{-1}$ respectively denote the conjugate transpose and the inverse, $\mathbf{I}$ is the identity matrix, $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\hat{\mathbf{R}}_{\mathcal{N}}$, and κ_0 is the maximum permitted condition number of $\mathbf{R}_{\varepsilon}$ and, indirectly, constrains the norm of $\mathbf{w}$ and the sensitivity of the beamformer to errors in $\mathbf{d}$ [26], [27].

The proposed technique first introduces the parametric model of the SCM from which the NCM is then estimated.

A. Parametric Model of SCM

The proposed sound field model used to generate the modelled SCM makes three simplifying assumptions:

1) the array is sufficiently compact that, for a plane wave, the relative transfer function (RTF) to each microphone with respect to the reference microphone can be represented by a multiplicative constant in the STFT domain [31],
2) the signal in each time-frequency cell is dominated by a small number, K, of the J plane wave components enumerated in (2) and (3). This is a generalization of the W-disjoint orthogonality assumption in [32],
3) The combined effect of the remaining $J - K$ plane wave components may be approximated as a diffuse isotropic sound field, $\boldsymbol{\gamma}$.

With these assumptions, (3) is represented in terms of the model as

$$\dot{\mathbf{y}}(\ell) = \sum_{k=1}^{K} \mathbf{a}(\Omega_k(\ell))\dot{S}_k(\ell) + \boldsymbol{\gamma}(\ell) + \mathbf{v}(\ell) \tag{7}$$

where $\dot{\mathbf{y}}(\ell)$ is the modelled signal, $\boldsymbol{\gamma}(\ell)$ is the diffuse noise signal, and $\Omega_k(\ell)$ is the direction of arrival (DOA) of the k^{th} plane wave component(s) in the ℓ^{th} frame (which may be different in each frequency bin), $\dot{S}_k(\ell)$ models the complex-valued amplitude of the k^{th} plane-wave as observed at the arbitrarily selected, reference microphone and $\mathbf{a}(\Omega)$ is the plane-wave array manifold expressed as the RTF from a distant source to each microphone with respect to the reference microphone.

The signal covariance matrix (SCM) is the STFT-domain covariance of the microphone signals

$$\mathbf{R}_{\mathbf{y}}(\ell) = \mathbb{E}\{\mathbf{y}(\ell)\mathbf{y}^H(\ell)\} \tag{8}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator. The powers and covariance matrices of the quantities in (7) are similarly defined as

$$\begin{aligned} \sigma_k(\ell) &= \mathbb{E}\{|\dot{S}_k^2(\ell)|\} \\ \overline{\mathbf{R}}_{\mathbf{a}}(\Omega) &= \mathbf{a}(\Omega)\mathbf{a}^H(\Omega) \\ \mathbf{R}_{\boldsymbol{\gamma}}(\ell) &= \mathbb{E}\{\boldsymbol{\gamma}(\ell)\boldsymbol{\gamma}^H(\ell)\} \triangleq \sigma_{K+1}(\ell)\overline{\mathbf{R}}_{\boldsymbol{\gamma}} \\ \mathbf{R}_{\mathbf{v}}(\ell) &= \mathbb{E}\{\mathbf{v}(\ell)\mathbf{v}^H(\ell)\} \triangleq \sigma_{K+2}(\ell)\overline{\mathbf{R}}_{\mathbf{v}} \end{aligned} \tag{9}$$

where a normalized covariance matrix is written with an overbar to indicate that it is scaled so that the diagonal element corresponding to the reference microphone equals unity. In this work we will assume that the pre-calculated normalized covariance matrices $\overline{\mathbf{R}}_{\boldsymbol{\gamma}}$ and $\overline{\mathbf{R}}_{\mathbf{v}}$ are based respectively on a spherically isotropic noise field and on uncorrelated spatially white Gaussian sensor noise.

It can now be seen that the covariance matrix of each term in (7) can be expressed as the product of a fixed matrix and a scalar parameter. Assuming that these terms are uncorrelated, this leads to a compact model of $\mathbf{R}_{\mathbf{y}}(\ell)$ written as

$$\mathbf{R}_{\dot{\mathbf{y}}}(\ell) = \sum_{k=1}^{K} \sigma_k(\ell)\overline{\mathbf{R}}_{\mathbf{a}}(\Omega_k(\ell)) + \sigma_{K+1}(\ell)\overline{\mathbf{R}}_{\boldsymbol{\gamma}} + \sigma_{K+2}(\ell)\overline{\mathbf{R}}_{\mathbf{v}} \tag{10}$$

and defined by the $2K + 2$ parameters $\{\Omega_k(\ell), \sigma_k(\ell)\}_{1\le k\le K}$, $\sigma_{K+1}(\ell)$ and $\sigma_{K+2}(\ell)$. These denote respectively the DOA and power of each of the K plane-wave components, the power of the diffuse noise component and the power of the sensor noise component all at the reference microphone.

B. NCM Estimation From the Modelled SCM

Where the estimated SCM model identifies a dominant source in exactly the same direction as the target, which is a *priori* known, the MVDR and MPDR beamformers are equivalent, as discussed in Sec I. However, coherent reflections and/or estimation errors can lead to the direction of the dominant source being slightly offset from the true source direction. Therefore to avoid signal cancellation, an estimate of the NCM may now be obtained by excluding the components likely to be associated with the target from the modelled SCM and calculated as

$$\hat{\mathbf{R}}_{\text{CM}}(\ell) = \sum_{k\in\mathcal{K}} \sigma_k(\ell)\overline{\mathbf{R}}_{\mathbf{a}}(\Omega_k(\ell)) + \sigma_{K+1}(\ell)\overline{\mathbf{R}}_{\boldsymbol{\gamma}} + \sigma_{K+2}(\ell)\overline{\mathbf{R}}_{\mathbf{v}} \tag{11}$$

where $\mathcal{K}$ denotes the set of plane-waves for which the angle between Ω_k and steering direction exceeds $\Delta\Omega$. In other words, using the modelled SCM from (10), the plane-wave power associated with directions that lie within $\Delta\Omega$ of the target direction are set to zero to obtain an estimate of the NCM, $\hat{\mathbf{R}}_{\text{CM}}$. The estimated NCM is then used to calculate the beamformer weights from (4)–(6). The choice of value for $\Delta\Omega$ is discussed in Sections IV-C and V-C.

The general approach outlined above allows for the incorporation of a *priori* known interferer directions. The usefulness of such information depends on the accuracy with which it is known or can be estimated. Since in many situations interferer directions are unknown, or vary with time, in the remainder of this paper we restrict our analysis to cases $K = 1$ and $K = 2$ and where the target is the only source whose direction is known.

III. Model Parameter Estimation

To determine the parameters of the SCM model in (10), an estimate, $\hat{\mathbf{R}}_{\mathbf{y}}(\ell)$, of the SCM in (8), is first obtained by applying a recursive estimator to the instantaneous covariance,

$$\hat{\mathbf{R}}_{\mathbf{y}}(\ell) = \alpha \hat{\mathbf{R}}_{\mathbf{y}}(\ell - 1) + (1 - \alpha)\mathbf{y}(\ell)\mathbf{y}^H(\ell), \tag{12}$$

where α determines a smoothing time constant whose choice is discussed in Section IV. This estimator is a lowpass filter with a gain of unity at zero frequency, as in [33]. The number of model parameters in (10) may be reduced by making the assumption that the DOAs, $\{\Omega_k(\ell)\}$, of all but the first of the dominant plane waves are known a *priori*. In this paper, we consider the cases $K = 1$ and $K = 2$ where, in the latter case, the second source corresponds to the target whose direction is assumed known. For each TF cell, the remaining $K + 3$ parameters are then chosen to minimize the Frobenius norm of the difference between the modelled and estimated covariance matrices as the solution to

$$\underset{\Omega_1, \{\sigma_k\}_{1 \le k \le K+2}}{\arg\min} \left\{ \left\| \mathbf{R}_{\dot{\mathbf{y}}} - \hat{\mathbf{R}}_{\mathbf{y}} \right\|_F^2 \right\} \tag{13}$$

where $\| \cdot \|_F$ denotes the Frobenius norm and the frame index, (ℓ), has been omitted for clarity.

In order to express (13) as a quadratic programming problem in standard form, an invertible, linear, norm-preserving transformation is applied that converts a complex-valued Hermitian covariance matrix, $\mathbf{R}$, into a real vector, $\mathbf{r}$, whose M^2 elements are defined by

$$\mathbf{r}_{i+(j-1)M} = \begin{cases} \sqrt{2}\Re\left(\mathbf{R}_{i,j}\right) & \text{for } 1 \le i < j \le M \\ \mathbf{R}_{i,j} & \text{for } 1 \le i = j \le M \\ \sqrt{2}\Im\left(\mathbf{R}_{i,j}\right) & \text{for } 1 \le j < i \le M \end{cases} \tag{14}$$

in which $\Re()$ and $\Im()$ take the real and imaginary parts of their arguments and $\mathbf{R}_{i,j}$ denotes the element of $\mathbf{R}$ at row i and column j. It can be verified that the transformation is linear and that

$$\|\mathbf{R}\|_F^2 = \mathbf{r}^T \mathbf{r}. \tag{15}$$

In the remainder of this section, the transformed version of a covariance matrix is denoted by a lower case $\mathbf{r}$ with the same subscripts and diacritics as the original matrix.

Substituting (10) into (13) and applying this transformation results in the following optimization problem

$$\begin{aligned} &\min_{\Omega_1} \left\{ \min_{\boldsymbol{\sigma}} \left(\left\| \mathbf{C}\boldsymbol{\sigma} - \hat{\mathbf{r}}_{\mathbf{y}} \right\|^2 \right) \right\} \\ &\quad \text{subject to} \quad \boldsymbol{\sigma}_k \ge 0 \quad \text{for } 1 \le k \le K + 2 \end{aligned} \tag{16}$$

where $\boldsymbol{\sigma}_k$ denotes the k^{th} element of $\boldsymbol{\sigma}$ and

$$\mathbf{C} = \left[\bar{\mathbf{r}}_{\mathbf{a}}(\Omega_1) \cdots \bar{\mathbf{r}}_{\mathbf{a}}(\Omega_K) \; \bar{\mathbf{r}}_{\boldsymbol{\gamma}} \; \bar{\mathbf{r}}_{\mathbf{v}} \right] \tag{17}$$

$$\boldsymbol{\sigma} = \left[\sigma_1 \cdots \sigma_K \; \sigma_{K+1} \; \sigma_{K+2} \right]^T. \tag{18}$$

The outer minimization in (16) is performed using an exhaustive search over all possible values of the discretized plane wave direction, Ω_1. For each possible value of Ω_1, the inner minimization in (16) is a quadratic programming problem whose solution is the unique pair of vectors, $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ that satisfy the Karush-Kuhn-Tucker (KKT) conditions [34]:

$$\boldsymbol{\mu} = \mathbf{C}^T \left(\mathbf{C}\boldsymbol{\sigma} - \hat{\mathbf{r}}_{\mathbf{y}} \right) \tag{19}$$

$$\text{with} \quad \boldsymbol{\sigma}_k \ge 0, \; \boldsymbol{\mu}_k \ge 0 \text{ and } \boldsymbol{\sigma}_k \boldsymbol{\mu}_k = 0 \quad \forall k. \tag{20}$$

This quadratic programming problem may be solved using the algorithm from [35]. The algorithm would normally be initialized with $\boldsymbol{\sigma} = \mathbf{0}$ but, because the solution must be found for many different values of Ω_1, it is more efficient to initialize with $\boldsymbol{\sigma} = \overline{\boldsymbol{\sigma}}$ where $\overline{\boldsymbol{\sigma}}$ is the solution to (16) but with the additional constraint $\overline{\sigma_1} = 0$ (i.e. without any directional component). It is found that, in the majority of TF cells, all the remaining component powers in $\overline{\boldsymbol{\sigma}}$ are strictly positive which implies that the corresponding elements of $\overline{\boldsymbol{\mu}}$ are necessarily zero because of the condition $\boldsymbol{\sigma}_k \boldsymbol{\mu}_k = 0$ in (20). To determine $\overline{\boldsymbol{\sigma}}$, it is therefore most efficient to apply the algorithm from [35] to the dual of the quadratic programming problem in which the roles of $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ are interchanged, (19) is rewritten as

$$\overline{\boldsymbol{\sigma}} = \left(\mathbf{C}^T \mathbf{C} \right)^{-1} \left(\overline{\boldsymbol{\mu}} + \mathbf{C}^T \hat{\mathbf{r}}_{\mathbf{y}} \right), \tag{21}$$

and the optimization is initialized with $\overline{\boldsymbol{\mu}}_k = 0$ for $k \neq 1$.

IV. Experiment Using Simulated RIRs

In this section an evaluation using simulated reverberant RIRs is conducted to compare the performance of the proposed compact model method with baseline methods under varying reverberation. An evaluation using the measured RIRs from a reverberant room under varying angular spacing and noise level is presented in Section V. The MATLAB code and some audio examples of the results presented in this paper are available at [36] and [37], respectively.

A. Scenario Setup

Recorded anechoic speech signals were convolved with simulated reverberant RIRs for a 32-element rigid spherical microphone array (SMA) with a radius of 4.2 cm (corresponding to the em32 Eigenmike [38]) using the spherical microphone arrays impulse response generator (SMIRgen) [39], [40] which is based on the image method [41]. Spatially white Gaussian sensor noise was also added. The reference microphone is chosen towards the top with azimuth of 180° and inclination of 21° where the array orientation is aligned with the X-axis, to minimise the dependence on azimuth of arrival of the sources.

As illustrated in Fig. 1, the array was placed at (4.65, 3.25, 1.5)m in a simulated shoebox room with dimension of

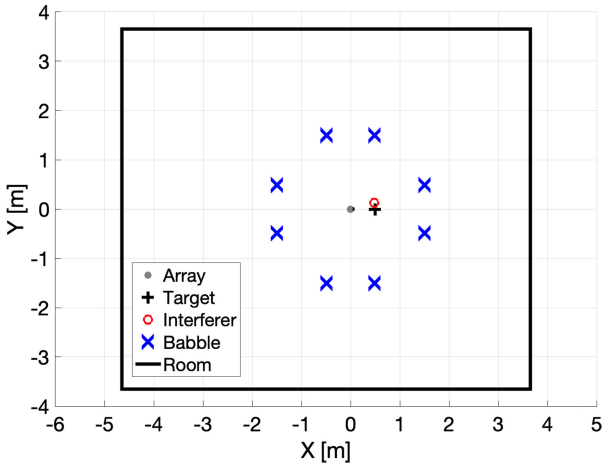


Fig. 1. The top view of the simulation setup. The origin is set to the array location.

$8.3 \times 6.5 \times 2.9$m. The reverberation time, T_{60}, was varied between 0.2 s and 1.0 s and the associated speech clarity, C_{50}, varied between 27.05 dB and 2.0 dB. The scenario involves two main sources (one target and one interferer) as well as eight sources which generate babble noise. All sources are stationary and lie on the array's horizontal plane. The target azimuth is 0° while the interferer azimuth is $\Delta\varphi = 15^\circ$. The direction of the interferer is unknown to the beamformer algorithms but the target direction is known. The eight babble sources have close-to-uniform circular distribution with azimuths of $\{18^\circ, 72^\circ, 108^\circ, 162^\circ, 198^\circ, 252^\circ, 288^\circ, 342^\circ\}$. The main sources (target and interferer) and the babble sources are respectively 0.5 m and 1.57 m away from the array.

The anechoic speech source signals were selected from Open Speech Repository [42] containing a selection of two female and six male talkers each reading different phrases in English. Fifty unique pairs of speech signals (with no repeated talker or phrase in a pair) were selected and each pair was used in two trials with swapped target and interferer positions giving a total of 100 trials per setting. For each trial, eight other speech signals were randomly selected as babble noise with a unique phrase per source in each trial. The signals were sampled at 8 kHz and truncated to a duration of 6s for processing.

The relative amplitudes of the sources at the reference microphone are controlled using three parameters defined as follows:

- signal-to-interference ratio (SIR) is the power ratio of the target source to the interferer source.
- signal-to-babble-noise ratio (SBNR) is the power ratio of the target source to the overall babble noise signal with the equal power at all babble sources.
- signal-to-sensor-noise ratio (SSNR) is the power ratio of the target source to the individual sensor noise.

In this experiment, SIR $= -5$ dB, SBNR $= 15$ dB, SSNR $=$ 40 dB and $\Delta\varphi = 15^\circ$ corresponding to a realistic challenging situation where an adjacent interferer masks the target. The effect of varying these parameters is investigated in Section V.

B. Methods

The proposed compact model (CM) method is compared with four baseline MVDR-based techniques including static and adaptive beamformers. They all share the same MVDR formulation and use (4)–(6) for the weight vector calculation but differ in the choice of covariance matrix. Table I provides a summary of the methods. An additional 'passthrough' case is also included, which is the unprocessed signal at the reference microphone. The description and calculation of the covariance matrix in the four baseline methods is as follows:

1) Isotropic (Iso): A stationary spherically diffuse sound field is assumed. This is equivalent to assuming the presence of interference equally from all directions. The spherically isotropic diffuse covariance matrix can be obtained as

$$\overline{\mathbf{R}}_{\boldsymbol{\gamma}} = \int_{\Omega} \mathbf{a}(\Omega)\mathbf{a}^H(\Omega)d\Omega, \tag{22}$$

where $\int_\Omega d\Omega = \int_0^{2\pi}\int_0^{\pi}\sin(\vartheta)d\vartheta d\varphi$ denotes integration along azimuth $\varphi \in [0, 2\pi)$ and inclination $\vartheta \in [0, \pi]$.

Equation (22) is approximated using a quadrature-weighted grid of discrete points $\mathcal{I}$ with 10° resolution in both azimuth and inclination giving

$$\hat{\mathbf{R}}_{\text{Iso}} = \sum_{i\in\mathcal{I}} w_i \mathbf{a}(\Omega_i)\mathbf{a}^H(\Omega_i), \tag{23}$$

where w_i is the quadrature weight for each sample point given by [43]

$$w_i = \frac{2\sin\vartheta_i}{N_\varphi N_\vartheta} \sum_{m=0}^{0.5N_\vartheta - 1} \frac{\sin\left((2m+1)\,\vartheta_i\right)}{2m+1} \tag{24}$$

in which ϑ_i is the inclination of sample point i and the number of sample points in azimuth and inclination are N_φ and N_ϑ respectively.

2) MPDR: An estimate of the SCM in (8) is obtained using (12) giving

$$\hat{\mathbf{R}}_{\text{MPDR}}(\ell) = \hat{\mathbf{R}}_{\mathbf{y}}(\ell). \tag{25}$$

3) Oracle-VAD: Using a voice activity detector (VAD) [44] on the oracle target signal in isolation at the array's reference channel, an estimate of the SCM as in (25) is used for the duration when the target is inactive giving

$$\hat{\mathbf{R}}_{\text{Oracle-VAD}}(\ell) = \begin{cases} \hat{\mathbf{R}}_{\mathbf{y}}(\ell) & \text{if VAD}_{\text{Target}}(\ell) = 0 \\ \hat{\mathbf{R}}_{\mathbf{y}}(\ell-1) & \text{otherwise} \end{cases} \tag{26}$$

where $\text{VAD}_{\text{Target}}(\ell)$ is the VAD state of the target at frame ℓ.

4) Oracle-NCM: The direct-path target-only signal, $\mathbf{y}_{\text{Target}}$ is used to estimate the true NCM as

$$\hat{\mathbf{R}}_{\text{Oracle-NCM}}(\ell) = \alpha\hat{\mathbf{R}}_{\text{Oracle-NCM}}(\ell-1) + (1-\alpha)(\mathbf{y}(\ell) - \mathbf{y}_{\text{Target}}(\ell))(\mathbf{y}(\ell) - \mathbf{y}_{\text{Target}}(\ell))^H. \tag{27}$$

C. Parameter Settings

The time-domain signals are transformed to the STFT domain using 16 ms frames with 50% overlap. The steering vectors used for all methods are anechoic simulated impulse responses for the rigid spherical array, as described in Section IV-A, and are

TABLE I
METHODS SUMMARY

	Iso	CM	MPDR	Oracle-VAD	Oracle-NCM
Signal Dependent / Independent	Independent	Dependent	Dependent	Dependent	Dependent
Static / Adaptive	Static	Adaptive	Adaptive	Adaptive	Adaptive
Real-time / Offline	N/A	Real-time	Real-time	Real-time	Real-time
Covariance Matrix	Spherically Diffuse	Modelled NCM	SCM	VAD-SCM	True NCM

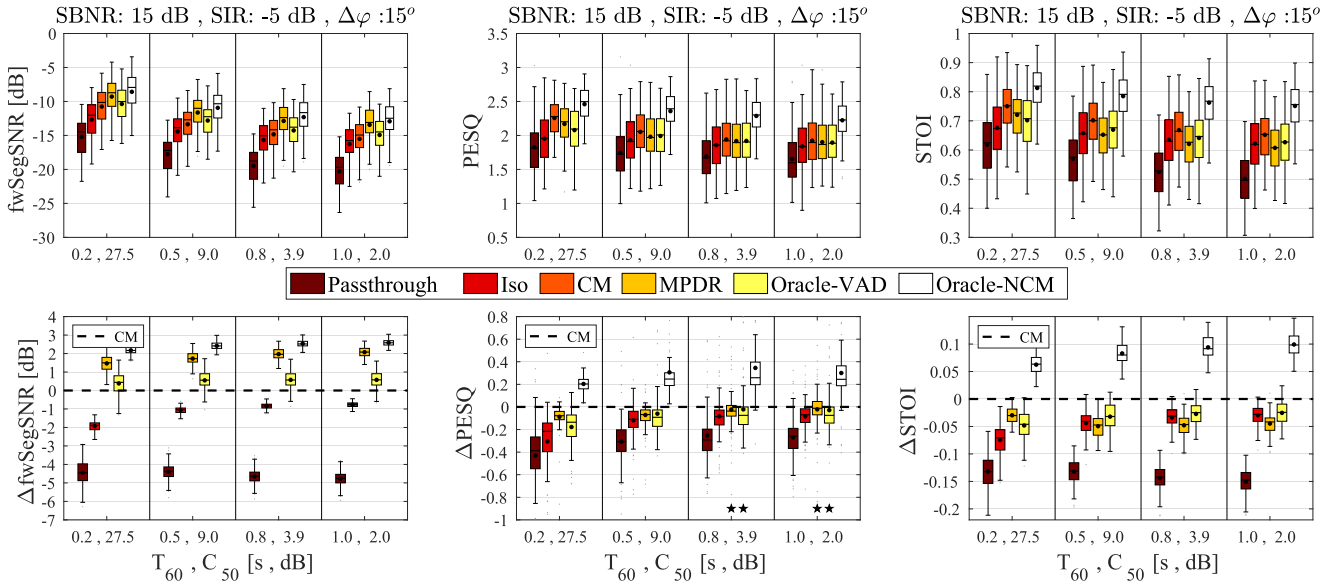


Fig. 2. Distribution of the evaluation metrics as absolute values (top row) and relative to CM (bottom row) for fixed SIR $= -5$ dB, SBNR $= 15$ dB, SSNR $= 40$ dB, $\Delta\varphi = 15^\circ$ and varying reverberation using simulated RIRs.

obtained using SMIRgen [39], [40]. The maximum condition number $\kappa_0 = 100$ in (6) is empirically chosen and used in all methods. The smoothing factor α used for the real-time covariance matrix estimation in the proposed, MPDR, Oracle-VAD and Oracle-NCM methods is calculated as

$$\alpha = \exp\left(\frac{-\Delta t}{\tau}\right), \tag{28}$$

where $\Delta t = 8$ ms is the frame hop and $\tau = 50$ ms is the time constant of the exponential moving average. The initial covariance matrix at $\ell = 0$ for the smoothing operation is the mean SCM in (8) averaged over the first 100 ms assuming no activity of the target in this initialisation interval.

In these experiments, the proposed method used $K = 1$ in (10) and a spatial search domain of discrete grid points with 5° resolution along azimuth and inclination covering the full azimuth circle and $90^\circ \pm 10^\circ$ in inclination. This implicitly assumes that the dominant sources lie within $\pm 10^\circ$ of the horizontal plane of the array. The mis-steering angular range $\Delta\Omega$ used to exclude the target-nearby estimated directions in (11) was empirically set to 10°.

D. Evaluation Metrics

The beamformer output was evaluated using three intrusive metrics; fwSegSNR [45], PESQ [46], [47], and STOI [48]. The PESQ and STOI metrics aim to measure speech quality and speech intelligibility respectively. Each metric compares the output of the beamformer to the target-only direct-path signal, at the reference microphone.

E. Results and Discussion

Fig. 2 shows the distribution of the evaluation metrics both as absolute values (upper row) and relative to CM (lower row) for four different values of T_{60} and C_{50} using simulated RIRs. The boxes show the upper and lower quartiles with the whiskers extended to 1.5 times the interquartile range with any values outside this range marked as grey dots. The mean and median are respectively indicated as black dots and horizontal solid black lines in the boxes. Using a paired t-test, all the relative differences plotted in the lower row are significant at the 5% level except where indicated with a star at the bottom of the plot. The results for each metric are discussed in turn below. In all cases lowest and highest metric scores were obtained with Passthrough and Oracle-NCM respectively, which confirms that the beamformers always improve the passthrough signal.

1) fwSegSNR Evaluation: Using the fwSegSNR metric, the order of performance from high to low is consistently Oracle-NCM, MPDR, Oracle-VAD, CM, Iso and Passthrough, as shown in the leftmost column of Fig. 2. CM significantly outperforms Iso by up to 2 dB and with an average improvement of 1 dB due to utilising an extra component for the presence of interferer in

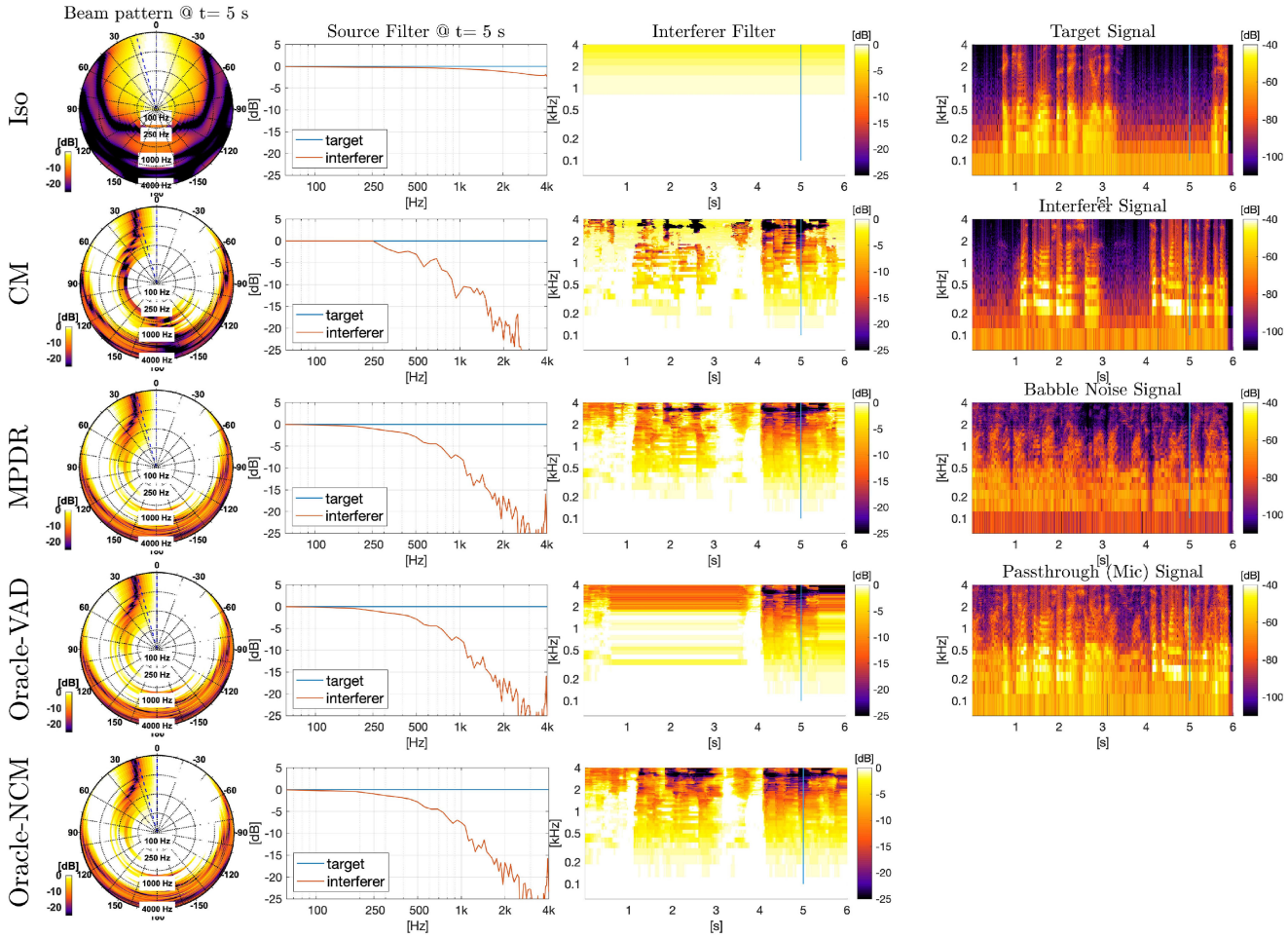


Fig. 3. The spatial, spectral and temporal visualisations of the filter gain (normalised by the transfer function of the reference microphone) for each beamformer in a trial with $T_{60} = 0.2$ s, $C_{50} = 27.5$ dB, SIR $= -5$ dB, SBNR $= 15$ dB and $\Delta\varphi = 15^\circ$ using simulated RIR. The fourth column shows the spectrogram of the oracle signals for the anechoic target, anechoic interferer, reverberant babbles as well as the passthrough at the reference microphone. For each method (per row), the first three columns respectively show the instantaneous spatio-spectral beam pattern on the horizontal plane at time $t = 5$ s, the instantaneous filter response in the direction of target and interferer, and the spectrogram of the filter in the interferer's direction. The solid blue line on the plots in the third and fourth columns shows the time used for the instantaneous plots in the first and second columns and the dashed blue lines on the first column plots indicate the azimuth of the target and interferer.

the model. It is observed that MPDR and Oracle-VAD achieve more noise reduction than CM. However, this comes at the cost of signal quality and intelligibility degradation, as is seen in the subsequent metrics.

2) PESQ Evaluation: As seen in the centre column of Fig. 2, CM outperforms Iso by up to 0.6 and with an average improvement of 0.2 in varying amount of reverberation due to interferer suppression. In high reverberation, MPDR and CM show similar performance whereas, in low and moderate reverberation, CM slightly leads by an average of 0.05 ($T_{60} \leq 0.5$ s, $C_{50} \leq 9.0$ dB).

3) STOI Evaluation: The absolute STOI plot in the top right plot of Fig. 2 shows that STOI decreases with increasing T_{60}. The Δ STOI plot shows that CM outperforms Iso, MPDR, and Oracle-VAD in almost all cases by up to 0.15 in STOI with an average improvement of 0.05.

Compared to Iso, CM shows a greater performance advantage in low reverberation ($T_{60} = 0.2$ s, $C_{50} = 27.5$ dB) than in moderate and high reverberation. This is expected since the suppression of reverberation is handled by the isotropic component, which is shared in both methods. Hence as the reverberation increases, the isotropic component (present in both Iso and CM) plays a greater role in the beamforming. However the distinguishability threshold of Iso and CM also depends on other parameters such as SIR, source angular spacing and the width of the main beam in Iso that is governed by the number of microphones in the array.

Compared to MPDR, CM gives a consistent STOI improvement of about 0.05 at all reverberation levels despite having a worse fwSegSNR. Compared to Oracle-VAD, CM still provides better STOI due to its target-robust adaptive suppression of interferer when the target is active, especially as the reverberation decreases where the presence of interferer is more prominent than the diffuse noise.

F. Beam Patterns Analysis

Fig. 3 shows the spatial, spectral and temporal representations of the filter gain for each beamformer as well as microphone and

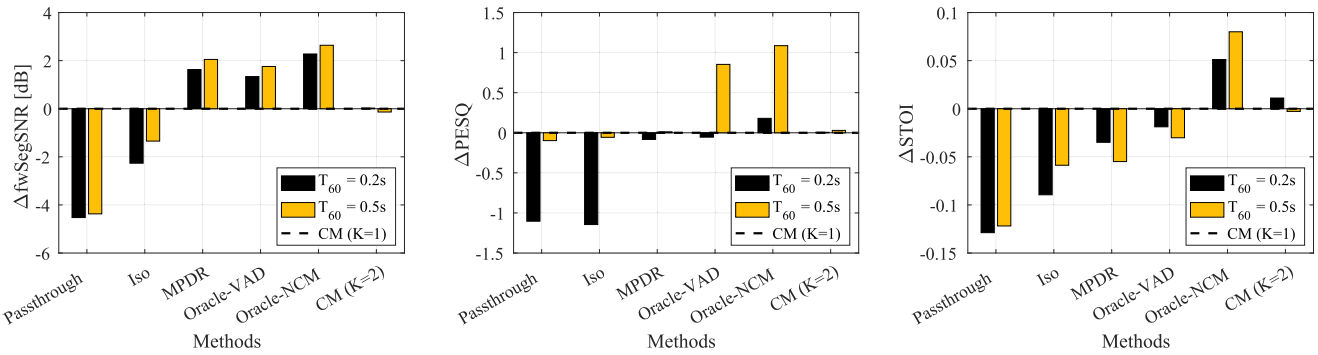


Fig. 4. The evaluation metrics of the methods relative to CM ($K = 1$) in a trial with SIR $= -5$ dB, SBNR $= 15$ dB and $\Delta\varphi = 15^\circ$ for varying $T_{60} = [0.2\,\text{s},\ 0.5\,\text{s}]$ using simulated RIR.

oracle signals for a single trial. Columns one to three illustrate the beamforming filter gains for each beamformer per row. The first two columns show the instantaneous filters at time $t = 5$ s in the azimuth-frequency and the frequency domains, respectively. The third column shows the filter in the interferer direction in the STFT-domain. The fourth column shows the oracle direct-path target and interferer, reverberant babble noise as well as the passthrough microphone signal all at the reference microphone. As expected, Iso's beam pattern is independent of the interferer unlike that of the three adaptive beamformers. Hence, due to the small angular spacing between interferer and target, the attenuation in the interferer direction is relatively low (≤ 2 dB) whereas the adaptive beamformers have much higher (≥ 10 dB) interferer attenuation above 1 kHz.

Although MPDR and Oracle-NCM look very similar, there are times where detailed differences can be observed in the third column plots, e.g. the last 0.5 s period, where strong attenuation of the interferer is continued to the end in Oracle-NCM, unlike MPDR. This is due to the presence of the target signal in the MPDR's SCM when the target is active. Oracle-VAD is similar to MPDR when there is no target signal as expected.

CM generally results in fewer STFT cells with significant interferer attenuation, compared to MPDR and Oracle-NCM. However, as the PESQ and STOI results in Fig 2 show, this is at the cost of degrading the target quality and intelligibility in MPDR. Note that the lack of attenuation at low frequencies is a consequence of the array dimensions, the condition-number limiting in (5)–(6) and the low angular spacing between interferer and target.

G. Multi-PW CM

In this subsection, the performance of CM with $K = 2$ is compared with the other methods, in particular with CM with $K = 1$. For CM with $K = 2$ the DOA of one of the PWs is assumed to be known and set to the target direction whereas the DOA of the second PW is estimated. Note that the power of both PWs are still to be estimated in CM. Although the PW component with the known DOA in the target direction is considered in the modelled SCM, it is excluded in the modelled NCM, $\hat{\mathbf{R}}_{\text{CM}}$, as in (11), since the target is within the target-exclusion zone $\Delta\Omega$ by definition.

TABLE II
RELATIVE LOCATION OF SOURCES WITH RESPECT TO THE ARRAY

	Target	Interferer A	Interferer B	Babbles
Azimuth [°]	0	27	45	18 to 342
Distance [m]	1.0	1.1	1.4	1.6
C_{50} [dB]	5.4	4.5	3.3	2.3 to 3.3

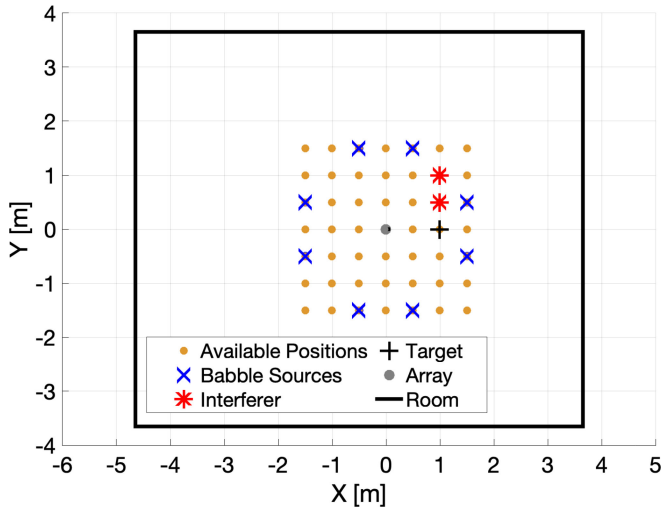


Fig. 5. The top view of the measured RIR setup. The origin is set to the array location.

Fig. 4 shows the evaluation metrics of the methods relative to CM ($K = 1$) in a trial with SIR $= -5$ dB, SBNR $= 15$ dB and $\Delta\varphi = 15^\circ$ for varying $T_{60} = [0.2\,\text{s},\ 0.5\,\text{s}]$ using simulated RIR. It can be seen that the additional PW results in a marginal improvement in STOI for low reverberation since the importance of PW component in CM becomes more significant as the reverberation decreases, as discussed in Section IV-E3.

V. EXPERIMENTS USING MEASURED RIRS

This section presents evaluations using RIRs recorded in a real reverberant room under varying angular spacing and noise level. Some audio examples of the results are available at [37].

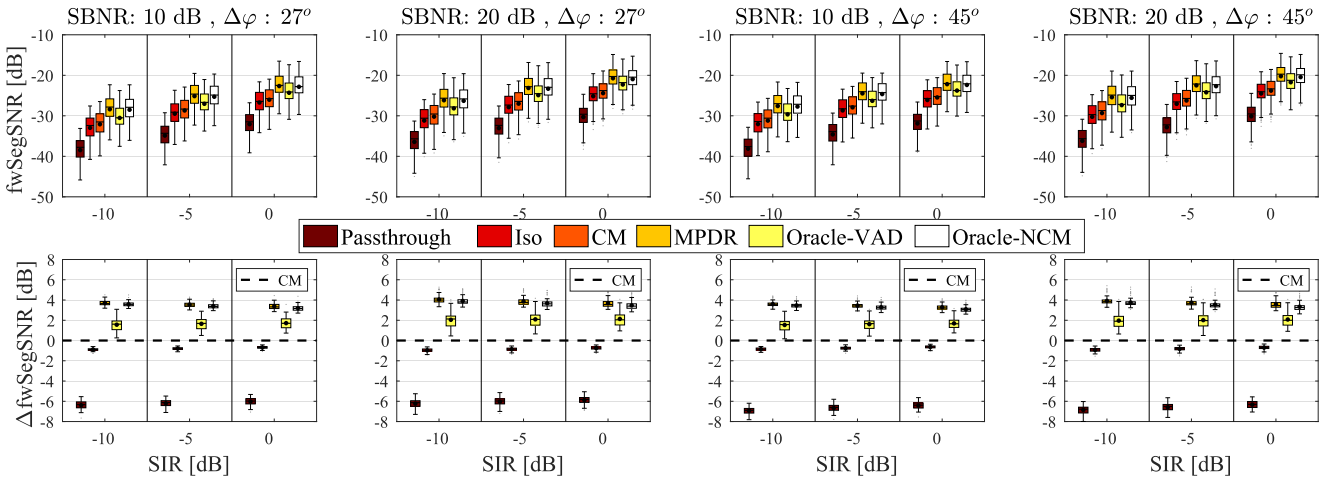


Fig. 6. Distribution of fwSegSNR (top row) and ΔfwSegSNR, relative to CM, (bottom row) for varying SIR, SBNR and $\Delta\varphi$ using measured RIRs.

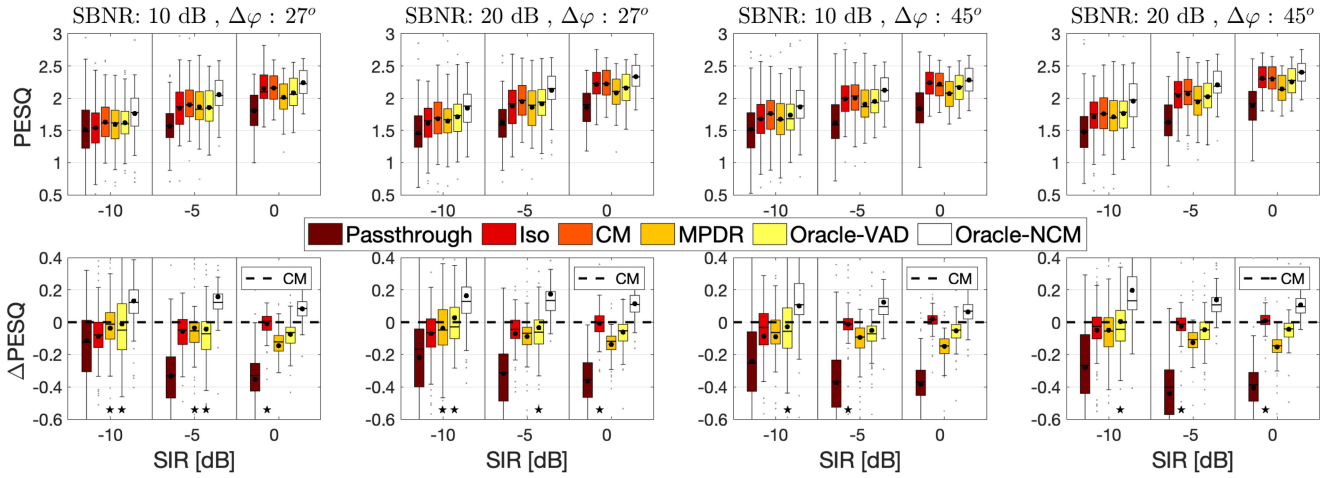


Fig. 7. Distribution of PESQ (top row) and ΔPESQ, relative to CM, (bottom row) for varying SIR, SBNR and $\Delta\varphi$ using measured RIRs.

A. *Experiment Setup*

The measured RIRs from METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset [49], [50] were convolved with the same anechoic speech source signals used in the previous section with additive spatially white Gaussian sensor noise.

The RIRs were recorded in an empty approximately rectangular classroom with a high reverberation time ($T_{60} \approx 1.12\,\mathrm{s}$) that is illustrated in Fig. 5. The room dimensions of $8.3 \times 6.5 \times 2.9\,\mathrm{m}$ and the array location of (4.65, 3.25, 1.5)m match the setup in the previous section. The available measurement points, indicated by dots, were symmetrically distributed on a grid with $0.5\,\mathrm{m}$ spacing surrounding the array and at the same height as the array. Table II lists the azimuth, source-to-array distance and associated C_{50} for the different sources. Note that although two positions for the interferer are considered, all trials involve a single-interferer scenario by including only one of the two possible positions giving two categories of $\Delta\varphi = 27°$ or $45°$. As in Section IV, the DOA of the target is known a *priori* while the DOA of the interferer is unknown. The same parameter settings and steering vectors as in Section IV-C are used for all methods including $K = 1$ for CM.

B. *Results and Discussion*

Figs. 6, 7 and 8 respectively show the distribution of fwSegSNR, PESQ and STOI using measured RIRs. For each metric, the four columns show different combinations of SBNR and $\Delta\varphi$ while, within each plot, the horizontal axis shows three different values of SIR. The discussion of the results based on each metric is as follows.

1) fwSegSNR Evaluation: As shown in Fig. 6, MPDR enhances SNR as much as Oracle-NCM since its covariance matrix partially contains the true NCM. Oracle-VAD has the third highest SNR improvement due to its utilisation of SCM but still less than Oracle-NCM and MPDR due to its lack of adaptive suppression during target activity. CM still leads in SNR enhancement by 1 dB over Iso due to its ability to suppress the interferer. The amount of SNR enhancement is not significantly high as strong attenuation of the interferer can be only achieved at high frequencies ($\geq$ 1kHz) whereas the the main interference

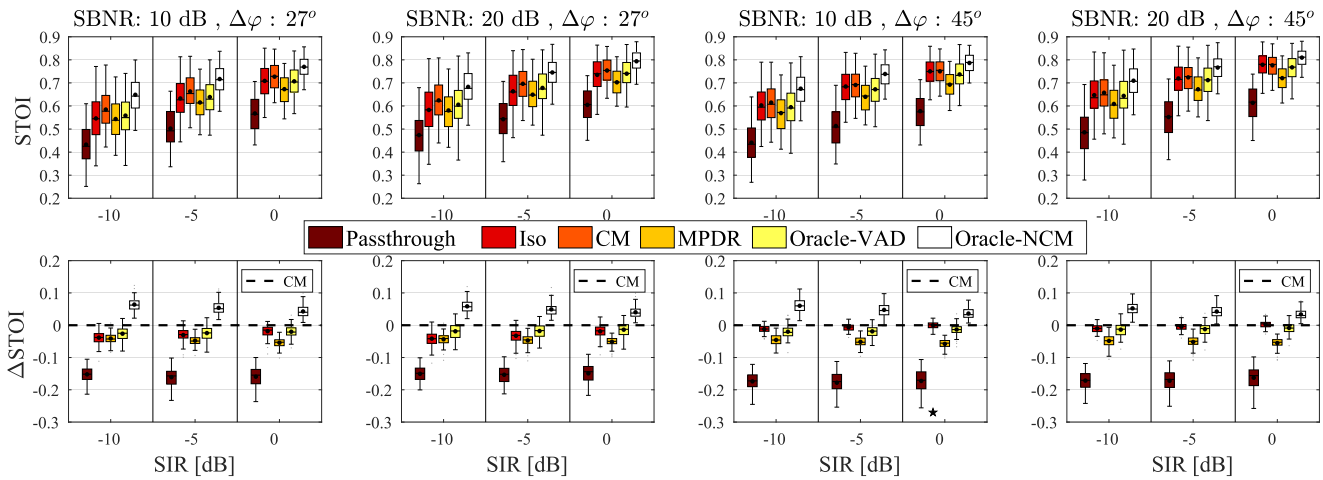


Fig. 8. Distribution of STOI (top row) and ΔSTOI, relative to CM, (bottom row) for varying SIR, SBNR and $\Delta\varphi$ using measured RIRs.

power is at lower frequencies, as shown in the rightmost column of Fig. 3.

2) PESQ Evaluation: The results in Fig. 7 (bottom row) show higher or similar PESQ scores for CM compared to Iso, MPDR and Oracle-VAD. This validates the enhancement of the target signal perceptual quality compared to other baselines due to CM's robustness to mis-steering, when compared to MPDR, and the use of an interferer component in the modelled covariance matrix, when compared to Iso as well as the robustness to target activity, when compared to Oracle-VAD.

3) STOI Evaluation: As seen in Fig. 8 (bottom row), CM consistently outperforms all except Oracle-NCM with a few cases being similar to Iso. Compared to Iso, as the SIR or $\Delta\varphi$ decreases, CM becomes more superior. The decrease in SIR increases the importance of interferer component in CM as an advantage over Iso whereas the decrease in $\Delta\varphi$ exposes the interferer more to the low or no-attenuation zone of the Iso's main lobe, as shown in Fig. 3.

Compared to MPDR, CM performs significantly better in terms of STOI by up to 0.1 and with an average improvement of 0.05. This is due to CM being more robust to mis-steering, to which MPDR is prone. These are likely to be more erroneous in measured RIRs than simulated RIRs. The superiority gap between CM and MPDR does not significantly vary over different SIR, SBNR and angular spacing as both methods adaptively include interferer and isotropic noises in their covariance matrix. CM also outperforms Oracle-VAD due to its adpative suppression of interferer during target activity, unlike Oracle-VAD.

C. Mis-Steering Evaluation

This subsection evaluates the effect of mis-steering for each method as well as the choice of $\Delta\Omega$ in CM. Fig. 9 shows the STOI for all methods including the variations of CM with varying choice of $\Delta\Omega = \{5°, 10°, 20°\}$ for a trial with SIR $= -5$ dB, SBNR $= 10$ dB and $\Delta\varphi = 45°$ using measured RIR. As expected, Oracle-NCM shows the highest robustness to mis-steering. The STOI drops sharper at the positive target DOA error due to the steering direction getting closer to the interferer placed at $+45°$

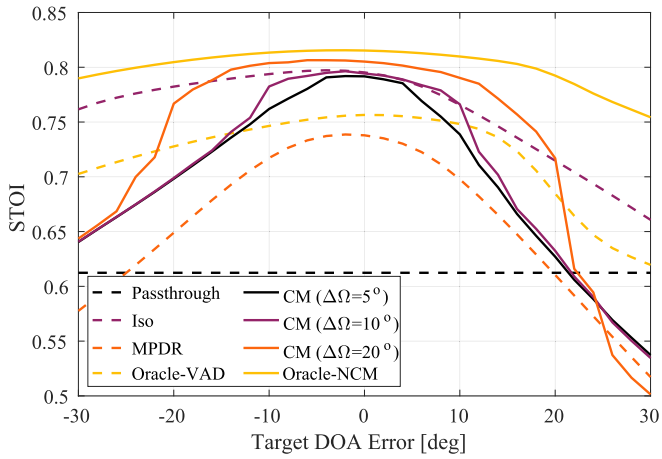


Fig. 9. The effect of mis-steering on STOI for a trial with SIR $= -5$ dB, SBNR $= 10$ dB and $\Delta\varphi = 45°$ using measured RIR.

from the target. The MPDR shows the highest sensitivity to mis-steering, as expected. It can clearly be seen that the increase in the size of target-exclusion zone $\Delta\Omega$ in CM improves the robustness to mis-steering as CM with $\Delta\Omega = 20°$ provides a wider safe zone (with relatively consistent STOI) than $\Delta\Omega = 10°$ and $15°$. Although the performance of CM improves with the increase in the size of target-exclusion zone, the choice of $\Delta\Omega$ needs to be less than the minimum angular separation of the interferer(s) to the target. Since a minimum of $\Delta\varphi = 15°$ was used in this paper for the interferer azimuth, the choice of $\Delta\Omega = 10°$ was chosen for our standard CM as stated previously in Section IV-C.

D. Computational Complexity

A comparison of the average run time for each method relative to MPDR is shown in Fig. 10. CM superiority is shown to be at the cost of approximately 4.5 times more computational cost than other conventional beamformers. Several optimizations could be considered including for example fixed, frequency-dependent, regularization in (6).

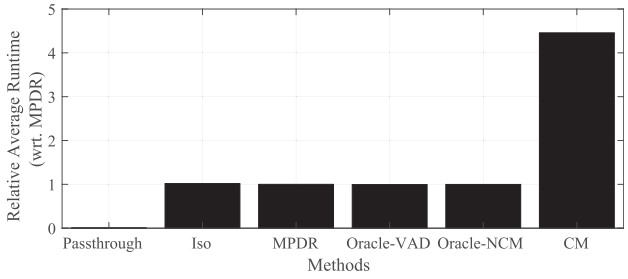


Fig. 10. Relative average run times for all methods with respect to MPDR.

VI. Conclusion

A method employing a compact model of the noise covariance matrix is proposed for adaptive beamforming. Using the signal covariance matrix as the reference, the model optimizes a small number of parameters for each TF cell including the direction of the dominant source(s) and power components associated with the plane-wave source(s), spherically isotropic noise and spatially white sensor noise. This parametric representation of the signal covariance matrix provides the ability to constrain the direction of interferer(s) over time, which increases the robustness to mis-steering or imperfect steering vectors, unlike the other adaptive baseline methods, which are based on the signal covariance matrix or a VAD-controlled estimate of the NCM. In addition, the proposed method uses a model with relatively few parameters to be estimated and pre-calculated basis elements to define the complex noise sound field. An evaluation based on simulated and measured RIRs for a 32-element spherical microphone array was conducted to compare the proposed method with other baseline methods, which employs static isotropic (Iso) sound field, SCM (MPDR and Oracle-VAD) and true NCM (Oracle-NCM) as well as passthrough signal (no processing), under varying interference and babble noise conditions, angular spacing and reverberation. The results show that all baseline methods (excluding the Oracle-NCM) are outperformed by the proposed method in terms of improving the target intelligibility and perceptual quality by up to 0.15 and 0.6 with average improvements of 0.05 and 0.1 in STOI and PESQ, respectively. Informal listening by the authors indicates that the subjective improvement is more perceptually significant than indicated by the metrics. The interested reader is invited to listen to the audio demonstrations available at [37]. Although the proposed method is shown to result in less improvement of fwSegSNR than MPDR and Oracle-VAD, CM is shown to have noticeable superiority and robustness in terms of enhancement of target intelligibility and perceptual quality due to its target-exclusion zone constraint over the interference direction as well as ability to suppress the interferer(s) during target activity. The superiority of CM over Iso increases as the relative level of interference noise increases, and the sources' angular spacing or reverberation reduces.

Acknowledgment

The authors would like to thank the associated editor and the reviewers for their constructive comments and suggestions, which have significantly benefited the paper.

References

[1] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds, Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.

[2] H. W. Löllmann *et al.*, "Microphone array signal processing for robot audition," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, 2017, pp. 51–55.

[3] T. J. Klasen, T. Van denM. BogaertMoonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[4] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 461–465.

[5] R. Haeb-Umbach *et al.*, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, Nov. 2019.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[7] H. L. van Trees, "Optimum waveform estimation," in *Optimum Array Processing*. Hoboken, NJ, USA: Wiley, 2002, pp. 428–709.

[8] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[9] W. W. Hager, "Updating the inverse of a matrix," *SIAM Rev.*, vol. 31, no. 2, pp. 221–239, Jun. 1989.

[10] H. L. van Trees, *Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2002.

[11] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoustical Soc. Amer.*, vol. 54, no. 3, pp. 771–785, Sep. 1973.

[12] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. IEEE Conv. Elect. Electron. Eng.*, 2010, pp. 416–420.

[13] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[14] X. Sun, Z. Wang, R. Xia, J. Li, and Y. Yan, "Effect of steering vector estimation on MVDR beamformer for noisy speech recognition," in *Proc. IEEE 23rd Int. Conf. Digit. Signal Process.*, 2018, pp. 1–5.

[15] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1507–1519, Oct. 2019.

[16] O. Schwartz, S. Gannot, and E. A. P. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, Sep. 2016.

[17] M. Tammen, S. Doclo, and I. Kodrasi, "Joint estimation of RETF vector and power spectral densities for speech enhancement based on alternating least squares," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[18] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, Jul. 2019.

[19] E. E. Jan and J. Flanagan, "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1996, pp. 917–920.

[20] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications, ser. Digital Signal Processing*, M. Brandstein and D. Ward, Eds, Berlin, Germany: Springer, 2001, pp. 19–38.

[21] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6717–6721.

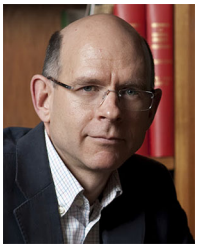
[22] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[23] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.
[24] H. H. Dam, H. Q. Dam, and S. Nordholm, "Noise statistics update adaptive beamformer with PSD estimation for speech extraction in noisy environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1633–1641, Nov. 2008.
[25] I. Kodrasi and S. Doclo, "Joint late reverberation and noise power spectral density estimation in a spatially homogeneous noise field," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.
[26] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.
[27] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.
[28] O. Thiergart and E. A. P. Habets, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 659–663.
[29] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 514–518, Nov. 1993.
[30] A. Moore, P. Naylor, and M. Brookes, "Improving robustness of adaptive beamforming for hearing devices," in *Proc. Int. Symp. Auditory Audiological Res.*, 2019, pp. 305–316.
[31] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[32] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[33] S. Braun *et al.*, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.
[34] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proc. Berkeley Symp. Math. Statist. Probability*, 1951, pp. 481–492.
[35] P. E. Gill and E. Wong, "Methods for convex and general quadratic programming," *Math. Program. Computat.*, vol. 7, no. 1, pp. 71–112, Mar. 2015.
[36] S. Hafezi, A. H. Moore, M. Brookes, R. R. Vos, and P. A. Naylor, "MATLAB code for compact noise covariance matrix model." [Online]. Available: https://github.com/ImperialCollegeLondon/sap-papers-2021-elospheres-compact-model
[37] Audio examples. [Online]. Available: https://imperialcollegelondon.github.io/sap-papers-2021-elospheres-compact-model/
[38] mh acoustics, "EM32 eigenmike microphone array release notes (v17.0)," *M. H. Acoust.*, NJ USA, Hardware, 2013. [Online]. Available: https://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf
[39] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.
[40] D. P. Jarrett, "Spherical microphone array impulse response (SMIR) generator." [Online]. Available: https://github.com/ehabets/SMIR-Generator
[41] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[42] Open speech repository. [Online]. Available: http://www.voiptroubleshooter.com/open_speech/
[43] J. R. Driscoll and D. M. Healy, "Computing Fourier transforms and convolutions on the 2-sphere," *Adv. Appl. Math.*, vol. 15, no. 2, pp. 202–250, Jun. 1994.
[44] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[45] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1978, pp. 586–590.
[46] 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," *Int. Telecommun. Union (ITU-T), Recommendation P.862*, Nov. 2003.
[47] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.
[48] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[49] M. B. Coteli, O. Olgun, and H. Hacihabiboglu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2215–2229, Nov. 2018.
[50] O. Olgun and H. Hacihabiboglu, "METU SPARG eigenmike em32 acoustic impulse response dataset v0.1.0 (Version 0.1.0)," 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2635758

Alastair H. Moore received the M.Eng. degree in electronic engineering with music technology systems and the Ph.D. degree from the University of York, York, U.K, in 2005 and 2010, respectively. He is currently a Research Fellow with Imperial College London, London, U.K., and spatial audio consultant with Square Set Sound. He spent three years as a Hardware Design Engineer for Imagination Technologies PLC designing digital radios and networked audio consumer electronics products. In 2012, he joined Imperial College, where he has contributed to a series of projects in the field of speech and audio processing applied to voice over IP, robot audition, and hearing aids. His research interests particularly include microphone array signal processing, modeling and characterization of room acoustics, dereverberation, and spatial audio perception. His current research interests include signal processing for moving and head-worn microphone arrays.

Sina Hafezi received the B.Eng. degree in electronic engineering in 2012, the M.Sc. degree in digital signal processing in 2013, from Queen Mary University of London, London, U.K., and the Ph.D. degree in acoustic source localisation using spherical microphone arrays in 2018 from Imperial College London, London, U.K. He is currently a postdoctoral Research Associate with Imperial College London, London. U.K. He was with the Centre for Digital Music as a Researcher and software engineer on autonomous multitrack mixing systems, which led to patent and spin-out company. He spent 2.5 years with Silixa as Senior Signal Processing Engineer developing algorithms and software for Distributed Acoustic Sensing systems. In 2021, he re-joined Imperial College London, where he has contributed to academic and industrial projects on hearing aids and spatial audio. His research interests include microphone array processing, spatial audio rendering, beamforming, source localisation and room acoustic modeling with applications for augmented, and virtual reality.

Rebecca R. Vos received the B.Sc. degree in physics from the University of Manchester, Manchester, U.K., in 2012, the M.Sc. degree in audio acoustics from the University of Salford, Salford, U.K., in 2013, and the Ph.D. degree in electronic engineering from the University of York, York, U.K, in 2018. She is currently a Postdoctoral Research Associate with Imperial College London, London, U.K. In 2019, she joined Imperial college with the ELO-SPHERES project. Her research interests include beamforming, microphone array processing, hearing aids, and singing perception.



Patrick A. Naylor (Fellow, IEEE) received the B.Eng. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., and the Ph.D. degree from Imperial College London, London, U.K. He is currently a Professor of speech and acoustic signal processing with Imperial College London. His research interests include speech, audio and acoustic signal processing. His current research addresses microphone array signal processing, speaker diarization, and multichannel speech enhancement for application to binaural hearing aids and robot audition. He has also worked on speech dereverberation including blind multichannel system identification and equalization, acoustic echo control, non-intrusive speech quality estimation, and speech production modeling with a focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several collaborative links with industry. He is currently a member of the Board of Governors of the IEEE Signal Processing Society and President of the European Association for Signal Processing. He was the formerly Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. He was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and is currently a Senior Area Editor of the IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING.

Mike Brookes (Life Member, IEEE) is currrently a Senior Research Investigator in Signal Processing with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. After graduation in mathematics from Cambridge University, Cambridge, U.K., in 1972, he was with the Massachusetts Institute of Technology, Cambridge, MA, USA and briefly, the University of Hawaii, Honolulu, HI, USA, before returning to the U.K. and joining Imperial College in 1977. Within the area of speech processing, he has concentrated on the modeling and analysis of speech signals, the extraction of features for speech and speaker recognition and on the enhancement of poor quality speech signals. He is the primary author of the VOICEBOX speech processing toolbox for MATLAB. Between 2007 and 2012, he was the Director of the Home Office sponsored Centre for Law Enforcement Audio Research (CLEAR) which investigated techniques for processing heavily corrupted speech signals. Between 2015 and 2019, he was Principal investigator of the E-LOBES project that addressed environment-aware enhancement algorithms for binaural hearing aids.