# Spatial audio signal enhancement by a two-stage source - system estimation with frequency smoothing for improved perception

Moti Lugasi, *Student Member, IEEE,* Anjali Menon, *Member, IEEE,* Vladimir Tourbabin, *Member, IEEE,* and Boaz Rafaely, *Senior Member, IEEE,*

***Abstract*— In many applications, such as hearing aids and virtual reality, spatial audio is used to provide a more natural experience to the users. However, when captured in the real world, the audio signals may suffer from noise and interference. The challenge in this case is to attenuate the undesired signals, while preserving the desired signals with their spatial information. In this paper, an approach for spatial signal enhancement is presented. This approach is based on two phases of estimation. The first phase is source signal estimation using a beamformer. Then, in the second phase, the acoustic transfer function (ATF) between the source and the array is estimated leading to an enhanced estimation of the desired signal at the microphones. This approach has been previously proposed but was not investigated in depth. In this paper, a model for the estimated desired signals is developed. In contrast to other methods of spatial enhancement, no trade-off between noise reduction and signal distortion is found in this model in the circumstance of a single desired source and single interfering source in a reverberant room. To overcome the limited accuracy of ATF estimation for short duration signals, frequency smoothing is applied. Listening tests verify the performance of the proposed approach.**

## I. Introduction

The important information that 3D sound carries and the natural way in which human beings process these signals motivate the incorporation of 3D sound in many applications, such as hearing aids [1], [2], virtual reality [3] and communication [4], [5]. In most of the aforementioned applications the sound field is captured in the real world by using a microphone array [6]. One problem encountered in this case is that the captured sound field may be composed of undesired components (e.g. noise sources and interference) in addition to the desired components. Hence, methods for the reduction of these undesired components are required.

There are a number of approaches that attempt to solve this problem. Most recent methods that are based on deep neural networks have been found to be useful for enhancing single channel signals, while preserving the monaural information [7], [8]. However, despite their effectiveness in noise reduction, these methods do not preserve spatial cues of the processed signals, and are therefore not suitable for spatial audio enhancement.

Binaural beamformers are a common solution for noise reduction and spatial cue preservation in the case of hearing aids. In [2] the authors suggest using a binaural beamformer, which, in certain acoustic cases, can be decomposed into a spatial filter (e.g. beamforming) and a single-channel postfilter. This method manages to preserve the spatial cues of the desired sound field, but may change the spatial cues of the residual noise; this may be problematic for acoustic awareness and environmental orientation. Hence, different extensions of this method have been developed to preserve the spatial cues of the residual noise in the cases of diffuse noise [9], [10] and directional interference [11], [12]. Another approach suggests applying common time-frequency masking to the binaural signal in order to attenuate the noise [13], [14]. However, in all of the aforementioned methods there is a trade-off between noise reduction and distortion of the desired components of the sound field. Moreover, these methods require the signals at the ears of the listeners, which restricts the choice of the microphone arrays that can be used and the range of operations that can be applied to those signals. One such operation enables to track the head rotation of the listener when the captured signals are played off-line [15].

In contrast, spherical microphone arrays (e.g. [16]) do not exhibit the last-mentioned limitations. Hence, they are commonly used for binaural reproduction [17] and for sound field reproduction [18], typically using Ambisonics signals [19]. Spherical arrays are also used for real-world recordings, motivating the development of methods that attempt to reduce the noise in the Ambisonics signals. In [20], [21] the authors proposed methods to attenuate some directions of the sound field and to amplify others, so that directional interference signals are attenuated. However, as a result, the desired sound field may be distorted as well. In [22], [23] the authors proposed two methods. The first method estimates the desired source signals, and then recovers the desired sound field by matching each source estimation to its corresponding steering vector. Hence, this method may not be suitable for reverberant sound fields. The second method proposes the application of a Wiener mask to each component of the sound field's plane wave amplitude density function (PWD), but this may change the direction of arrival (DOA) of the residual noise under certain acoustic conditions [24], [25]. In addition, this method distorts the desired sound field [23]. Another drawback of methods based on Ambisonics enhancement is that the Ambisonics signals typically require a specially designed microphone array, which may not be practical for all applications [19, Ch. 1]. It is noteworthy that in order to use Ambisonics signals for binaural reproduction, the head related transfer function (HRTF) of the user (the listener) is required [26]. For practical reasons, this HRTF is measured using a dummy

head, which may not generalize well for some users [27].

While in many cases the methods presented above provide a good solution for spatial audio signal enhancement, they all have clear limitations. Methods tailored for binaural and spherical arrays impose a severe constraint on array configurations, while many methods require information on the signal's statistics, which may not always be available. In summary, a method that can be applied to any microphone array and that does not require extensive prior information may be of great interest. One example of such a method was presented in [28]; in this method the desired source signal is estimated using a maximum directivity beamformer, followed by acoustic transfer function (ATF) estimation. However, this method was only proposed for spherical arrays, and lacked comprehensive theoretical and experimental performance analysis.

Motivated by [28], in this paper a general framework for spatial audio capture enhancement is presented. The desired source signal is first estimated using a spatial filter, and then the ATF is estimated using the method from [29] to reproduce the desired signal at the microphones. This framework can be applied under various acoustic conditions and with various microphone arrays, while the only required prior information is for the beamforming. Recently, this approach was shown, objectively and in a listening test, to be very effective for spatial enhancement in the case of a wearable array (the Facebook augmented reality glasses [30], [31]).

The contribution of this paper is threefold. First, models for the observed signal, the source signal estimate, and the ATF estimate are provided, facilitating analysis of the errors and artifacts of the processed signals. These models predict that the proposed approach will provide distortion-free enhancement of the processed signals in the case of a high output signal-to-noise ratio (SNR) at the source estimation stage. In addition, these models predict that, in contrast to other approaches [11], [12], the spatial cues of the residual noise may not be preserved using the proposed approach. Second, frequency smoothing is used in the ATF estimation stage to improve the perception experience. Third, objective analysis and listening tests show the superiority of the proposed approach over other state-of-the-art methods for spatial enhancement.

## II. Signal and system model

Assume that a desired source, with signal $s$, and $L$ interfering sources, with signals $u_1(f), u_2(f), ..., u_L(f)$, are located in a reverberant room, where $f$ denotes the frequency in Hz. In addition to these sources, assume an undesired noise, which cannot be described as a point source, also exists in this room, and the entire sound field is captured by using a microphone array with an arbitrary configuration, having $I$ microphones. The model of the observed signal, denoted $\mathbf{x}$, with size $I \times 1$, is given by:

$$\mathbf{x}(f) = s(f)\mathbf{h}(f) + \sum_{l=1}^{L} u_l(f)\mathbf{r}_l(f) + \tilde{\mathbf{n}}(f), \tag{1}$$

where $\mathbf{h}(f) = [h_1(f), h_2(f), ..., h_I(f)]^T$ and $\mathbf{r}_l(f) = [r_{l1}(f), r_{l2}(f), ..., r_{lI}(f)]^T$ for $l = 1, 2, ..., L$ are the ATFs of the desired and the $L$ interference sources at frequency $f$, respectively; the signal $\tilde{\mathbf{n}}(f)$ denotes the ambient noise, and $(\cdot)^T$ represents the transpose operator. The terms $h_i(f)$ and $r_{li}(f)$ are denoted the ATFs of the desired source and the $l^{\text{th}}$ interfering source for $1 \leq i \leq I$, respectively. For brevity the frequency index, $f$, is omitted from now on.

Equation (1) can be rewritten using matrix notation as follows:

$$\mathbf{x} = \mathbf{h}s + \mathbf{R}\mathbf{u} + \tilde{\mathbf{n}} = \mathbf{d} + \mathbf{n}, \tag{2}$$

where $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_L]$, $\mathbf{u} = [u_1, u_2, ..., u_L]^T$, $\mathbf{d} = \mathbf{h}s$ and $\mathbf{n} = \mathbf{R}\mathbf{u} + \tilde{\mathbf{n}}$; $\mathbf{d}$ and $\mathbf{n}$ are the desired and the undesired microphone signals, respectively. Throughout this work it is assumed that $\mathbf{d}$ and $\mathbf{n}$ are independent zero-mean random processes and the correlation matrix of vector $\mathbf{n}$ is given by:

$$\mathbf{P} = E\{\mathbf{n}\mathbf{n}^H\}, \tag{3}$$

where $E\{\cdot\}$ is the expectation operator, $(\cdot)^H$ represents the conjugate transpose operator.

## III. Distortionless enhancement

The aim of this work is to reduce the noise level in a spatial audio signal, while maintaining the desired microphone signal unchanged. In other words, the desired microphone signal $\mathbf{d}$ in (2) should be undistorted, while the variance of the undesired microphone signals at each microphone is required to be attenuated. This can be achieved by applying $I$ beamformers (as the number of microphones in the array), where each one of these beamformers provides a distortionless response to the desired signal, while minimizing the noise variance at its corresponding microphone. These beamformers will be defined first.

Let

$$\mathbf{w}_i = [w_{i,1}, w_{i,2}, ..., w_{i,I}]^T \tag{4}$$

denote the $i^{\text{th}}$ beamformer weights at frequency $f$, which aim to produce the undistorted desired signal of the $i^{\text{th}}$ microphone at the beamformer's output:

$$y_i = \mathbf{w}_i^H \mathbf{x}, \tag{5}$$

where $y_i$ denotes the output of the $i^{\text{th}}$ beamformer. Now, the contribution of the noise component to the variance of the beamformer's output, $E\{|y_i|^2\}$, is computed from (2), (3) and (5) as $\mathbf{w}_i^H \mathbf{P} \mathbf{w}_i$. Minimizing this noise term and constraining a distortionless response leads to $I$ optimization problems as follows:

$$\begin{aligned} &\underset{\mathbf{w}_i}{\text{minimize}} \;\; \mathbf{w}_i^H \mathbf{P} \mathbf{w}_i \\ &\text{subject to} \;\; \mathbf{w}_i^H \mathbf{h} = h_i, \;\; \text{for} \;\; i = 1, 2, ..., I. \end{aligned} \tag{6}$$

The solution is given by [19, p. 163], [32, pp. 428-709] as:

$$\mathbf{w}_{\text{oracle}}^i = h_i^* \cdot \frac{\mathbf{P}^{-1}\mathbf{h}}{\mathbf{h}^H \mathbf{P}^{-1} \mathbf{h}} = h_i^* \cdot \mathbf{w}_{\text{oracle}}, \;\; \text{for} \;\; i = 1, 2, ..., I. \tag{7}$$

Note that these $I$ oracle beamformers have a common term, denoted as $\mathbf{w}_{\text{oracle}}$, while the term $h_i^*$ is different for each beamformer $i$. This separation of terms provides a unique interpretation for the beamformers, that stems from the common

beamformer $\mathbf{w}_{\text{oracle}}$. This beamformer ($\mathbf{w}_{\text{oracle}}$) is the optimal solution of the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{w}}{\text{minimize}} \;\; \mathbf{w}^H\mathbf{P}\mathbf{w} \\ &\text{subject to} \;\; \mathbf{w}^H\mathbf{h} = 1, \end{aligned} \tag{8}$$

where $(\cdot)^*$ is the conjugate operator and $\mathbf{w} = [w_1, w_2, ..., w_I]^T$. The solution of (8) is a beamformer that aims to provide a distortionless estimate of the source signal $s$, while minimizing the noise component at the beamformer output. Once this single beamformer has been computed, all $I$ beamformers can now be obtained directly from (7).

Following the derivation above, the goal of producing a distortionless response for the desired microphone signals, while minimizing the contribution of the noise, can be achieved in two stages: (i) apply a minimum-variance distortionless response beamformer, with weights $\mathbf{w}_{\text{oracle}}$ from (7), to estimate the desired source signal $s$, denoted here $\hat{s}$; (ii) multiply $\hat{s}$ by $\mathbf{h}$ to obtain the noise-attenuated desired signal at the microphones, leading to:

$$\mathbf{y}_{\text{known } \mathbf{h}} = \hat{s} \cdot \mathbf{h}, \tag{9}$$

where $\mathbf{y}_{\text{known } \mathbf{h}} = [y_1, ..., y_I]^T$.

Note that, as detailed above, the optimal solution of (6) requires knowledge of the ATF $\mathbf{h}$, even at the source estimation stage. Unfortunately, in practice, the ATF corresponding to the desired source may not be known. For that reason the beamformer from (7) is assumed to be an oracle beamformer. In the following sections an approach for estimating the desired source signal and its corresponding ATF is presented with the aim of approximating the ideal solution in (9).

## IV. Desired microphone signal estimation process

In order to estimate the desired microphone signal from the observed signal, two phases are used. In the first phase, the desired source signal is estimated by using a spatial filter (beamformer). In the second phase, the ATF, $\mathbf{h}$, is estimated using the desired source signal estimate, $\hat{s}$, and the observed signal, $\mathbf{x}$. Then, these two estimates are used to recover the desired microphone signals, $\mathbf{d}$. In this section, these phases are presented.

### *A. Source signal estimation using beamforming*

By using spatial filtering (beamforming) and the observed signal, the source can be estimated as follows:

$$\hat{s} = \mathbf{w}^H\mathbf{x}, \tag{10}$$

where $\mathbf{w} = [w_1, w_2, ..., w_I]^T$ is the vector of the beamformer weights. Typically, the beamformer is steered towards the direct sound. Therefore, the steering vector in the desired source DOA should form a basic beamformer. Assuming that the required information is available, by using a beamformer with a distortionless response in the DOA of the desired source, the desired source signal is correctly obtained at the output of the beamformer. In addition, the reflections and the noise sources also appear at the output of the beamformer and add distortion and noise to the source signal estimate. By substituting (2) into (10), the following is obtained:

$$\begin{aligned} \hat{s} &= \mathbf{w}^H\mathbf{d} + \mathbf{w}^H\mathbf{n} \\ &= \mathbf{w}^H\mathbf{h}s + \mathbf{w}^H\mathbf{n}. \end{aligned} \tag{11}$$

It can be shown that the ideal beamformer from (7) maximizes the SNR at the output of the beamformer; this will later be shown to be an advantage. However, the ATF $\mathbf{h}$ is required in order to construct this beamformer. Hence, in many acoustic scenes where the ATF is not available, the ideal beamformer is not a practical solution, and other alternatives are needed.

Assuming that there are segments in time where the observation signal $\mathbf{x}$ does not contain the desired microphone signals $\mathbf{d}$, the matrix $\mathbf{P}$ can be estimated. In addition to the assumption that the steering vector in the desired source DOA is given, a realistic approximation of $\mathbf{w}_{\text{oracle}}$ from (7), which solves the optimization problem:

$$\begin{aligned} &\underset{\mathbf{w}}{\text{minimize}} \;\; \mathbf{w}^H\hat{\mathbf{P}}\mathbf{w} \\ &\text{subject to} \;\; \mathbf{w}^H\mathbf{a}_{\mathbf{d}} = 1, \end{aligned} \tag{12}$$

can be obtained, and the solution is given by:

$$\mathbf{w}_{\text{MVDR}} = \frac{\hat{\mathbf{P}}^{-1}\mathbf{a}_{\mathbf{d}}}{\mathbf{a}_{\mathbf{d}}{}^H\hat{\mathbf{P}}^{-1}\mathbf{a}_{\mathbf{d}}}, \tag{13}$$

where $\hat{\mathbf{P}}$ is an estimate of matrix $\mathbf{P}$ and $\mathbf{a}_{\mathbf{d}}$ is the steering vector in the desired source DOA. Unlike the ATF $\mathbf{h}$, the steering vector $\mathbf{a}_{\mathbf{d}}$ is dependent only on the array configuration; it is the transfer function between the source and the array in a free field. Therefore, $\mathbf{a}_{\mathbf{d}}$ can be calculated (or measured) once, for a given DOA, and then can be used for any acoustic condition. The beamformer in (13) is called the minimum variance distortionless response (MVDR) beamformer [19, p. 163], but, in this case, the distortionless response is only for the direct sound.

If matrix $\mathbf{P}$ cannot be estimated for some reason, then, by assuming that the undesired microphone signal is spatially white noise, namely $\mathbf{P} \sim \mathbf{I}_I$, where $\mathbf{I}_I$ is a unit matrix of size $I \times I$, the beamformer in (13) can be rewritten as:

$$\mathbf{w}_{\text{Bar}} = \frac{\mathbf{a}_{\mathbf{d}}}{\|\mathbf{a}_{\mathbf{d}}\|^2}, \tag{14}$$

which is also known as the Bartlet beamformer [33]. The beamformers presented in (13) and (14), in addition to the ideal solution from (9), will be used to estimate the desired source signal, and the results of each will be theoretically and experimentally analyzed later, where the ideal solution from (9) will be set as an upper bound for the other beamformers.

### *B. Acoustic transfer function and desired microphone signal estimation*

To estimate the ATF $\mathbf{h}$, the following formulation is used [34]:

$$\hat{\mathbf{h}}_{\text{opt}} = \frac{S_{s\mathbf{x}}}{S_{ss}}, \tag{15}$$

where $S_{s\mathbf{x}}$ and $S_{ss}$ are the cross-spectrum of the signals $s$ and the observed signal $\mathbf{x}$, and the auto-spectrum of $s$, respectively.

Under the assumption that $s$ and $\mathbf{n}$ are uncorrelated, the estimator in (15) leads to a correct estimate of the ATF, namely, $\hat{\mathbf{h}}_{\text{opt}} \approx \mathbf{h}$. Unfortunately, in the considered acoustic conditions the desired source signal $s$ is not available. Therefore, the estimate of the desired source signal is used to approximate the estimate in (15) as follows:

$$\hat{\mathbf{h}} = \frac{S_{\hat{s}\mathbf{x}}}{S_{\hat{s}\hat{s}}}, \tag{16}$$

where $S_{\hat{s}\mathbf{x}}$ and $S_{\hat{s}\hat{s}}$ are the cross-spectrum of the signals $\hat{s}$ and the observed signal $\mathbf{x}$, and the auto-spectrum of $\hat{s}$, respectively. By using the source signal estimate from (11), and the observed signal from (2), the cross-spectrum of the desired source signal estimate and the observation vector, $S_{\hat{s}\mathbf{x}}$, is given by:

$$S_{\hat{s}\mathbf{x}} = S_{\hat{s}\mathbf{d}} + S_{\hat{s}\mathbf{n}}. \tag{17}$$

By using matrix $\mathbf{P}$ from (3), the explicit definition of $\mathbf{d}$ from (2), and the explicit expression of the desired source signal estimate given in (11), the cross-spectra $S_{\hat{s}\mathbf{d}}$, and $S_{\hat{s}\mathbf{n}}$ are given by:

$$S_{\hat{s}\mathbf{d}} = S_{ss}(\mathbf{w}^H\mathbf{h})^*\mathbf{h}, \tag{18}$$

and

$$S_{\hat{s}\mathbf{n}} = \mathbf{P}\mathbf{w}, \tag{19}$$

where $S_{ss}$ is the auto spectrum of the desired source signal. By substituting (18) and (19) into (17), $S_{\hat{s}\mathbf{x}}$ is rewritten as:

$$S_{\hat{s}\mathbf{x}} = S_{ss}(\mathbf{w}^H\mathbf{h})^*\mathbf{h} + \mathbf{P}\mathbf{w}. \tag{20}$$

Using (11) and (3), the auto-spectrum of the desired source signal estimate, $S_{\hat{s}\hat{s}}$, is given by:

$$S_{\hat{s}\hat{s}} = S_{ss}|\mathbf{w}^H\mathbf{h}|^2 + \mathbf{w}^H\mathbf{P}\mathbf{w}. \tag{21}$$

Therefore, by substituting (20),(21) into (16) the following is obtained:

$$\hat{\mathbf{h}} = \frac{S_{ss}(\mathbf{w}^H\mathbf{h})^*\mathbf{h} + \mathbf{P}\mathbf{w}}{S_{ss}|\mathbf{w}^H\mathbf{h}|^2 + \mathbf{w}^H\mathbf{P}\mathbf{w}}. \tag{22}$$

Later, (22) will be investigated for a simple acoustic scene to provide some insight into this expression.

Using the source signal estimate, $\hat{s}$, and the estimate of the ATF, $\hat{\mathbf{h}}$, the estimate of the desired microphone signal, $\mathbf{d}$, is given by:

$$\mathbf{y} = \hat{s}\hat{\mathbf{h}}. \tag{23}$$

## V. Theoretical performance analysis

In this section a theoretical analysis of the proposed approach, as presented in the previous section, is provided. In order to estimate the desired microphone signals $\mathbf{d}$, the desired source signal estimate from (11) and the estimate of the ATF $\mathbf{h}$ from (22) are substituted into (23). For the general case, the expression from (23) might be complex to analyze. Therefore, a simple acoustic scene, composed of one desired source and one interference, is considered in order to gain insight into the noise attenuation and the signal distortion that the proposed approach provides.

### A. Expansion of the estimated desired microphone signal into distinct components

Under the assumption of a single desired source and a single interference that are located in a reverberant room, the estimates of the desired signal and its corresponding ATF, according to (11) and (22), are given by:

$$\hat{s} = \mathbf{w}^H\mathbf{h}\cdot s + \mathbf{w}^H\mathbf{r}_1\cdot u_1, \tag{24}$$

and

$$\hat{\mathbf{h}} = \frac{\mathbf{h}}{\mathbf{w}^H\mathbf{h}}\cdot\frac{\text{SNR}_{\text{BF}}}{1+\text{SNR}_{\text{BF}}} + \frac{\mathbf{r}_1}{\mathbf{w}^H\mathbf{r}_1}\cdot\frac{1}{1+\text{SNR}_{\text{BF}}}, \tag{25}$$

respectively, where $u_1$ and $\mathbf{r}_1$ were defined in (1), $\text{SNR}_{\text{BF}}$ is the SNR at the output of the beamformer and is given by:

$$\text{SNR}_{\text{BF}} = \frac{S_{ss}|\mathbf{w}^H\mathbf{h}|^2}{\mathbf{w}^H\mathbf{P}\mathbf{w}} = \frac{S_{ss}|\mathbf{w}^H\mathbf{h}|^2}{S_{u_1u_1}|\mathbf{w}^H\mathbf{r}_1|^2}, \tag{26}$$

and $\mathbf{P} = S_{u_1u_1}\mathbf{r}_1\mathbf{r}_1^H$ in this case. As shown in (26), for this specific acoustic scene, the SNR at the beamformer's output can be presented as a multiplication of two components as follows:

$$\text{SNR}_{\text{BF}} = \text{SNR}_{\text{in}}\cdot|G|^2, \tag{27}$$

where $\text{SNR}_{\text{in}} = \frac{S_{ss}}{S_{u_1u_1}}$ and $G = \frac{\mathbf{w}^H\mathbf{h}}{\mathbf{w}^H\mathbf{r}_1}$. Therefore, using (24), (25) and (27) leads to the following representation of the desired microphone signal estimate:

$$\begin{aligned}\mathbf{y} = &\underbrace{s\mathbf{h}\cdot\frac{\text{SNR}_{\text{BF}}}{1+\text{SNR}_{\text{BF}}}}_{\mathbf{y}_s} + \underbrace{s\mathbf{r}_1\cdot\frac{G}{1+\text{SNR}_{\text{in}}\cdot|G|^2}}_{\mathbf{y}_{s\text{ phantom}}} \\ &+ \underbrace{u_1\mathbf{h}\cdot\frac{\text{SNR}_{\text{in}}\cdot G}{1+\text{SNR}_{\text{in}}\cdot|G|^2}}_{\mathbf{y}_{u\text{ phantom}}} + \underbrace{u_1\mathbf{r}_1\cdot\frac{1}{1+\text{SNR}_{\text{BF}}}}_{\mathbf{y}_u}.\end{aligned} \tag{28}$$

As can be seen from (28), four components compose $\mathbf{y}$, labeled $\mathbf{y}_s$, $\mathbf{y}_{s\text{ phantom}}$, $\mathbf{y}_{u\text{ phantom}}$ and $\mathbf{y}_u$. The rationale behind the names of the components composing $\mathbf{y}$ stems from their physical interpretation; for instance, $\mathbf{y}_s$ contains the desired source in its original position (due to the appearance of ATF $\mathbf{h}$), whereas, $\mathbf{y}_{s\text{ phantom}}$ is the desired source in the position of the undesired source position (due to the appearance of ATF $\mathbf{r}_1$). These four components are analyzed in the following sections.

### B. Frequency dependent distortion ($\mathbf{y}_s$)

The first component in (28) is composed of the product of the terms $s\mathbf{h}$, and is the desired microphone signal $\mathbf{d}$ and the term $\frac{\text{SNR}_{\text{BF}}}{1+\text{SNR}_{\text{BF}}}$. The term $\frac{\text{SNR}_{\text{BF}}}{1+\text{SNR}_{\text{BF}}}$ is frequency dependent, and may lead to frequency-dependent distortion, in particular when $\text{SNR}_{\text{BF}}$ is low and changes significantly with frequency. Note that $\frac{\text{SNR}_{\text{BF}}}{1+\text{SNR}_{\text{BF}}}$ depends on the beamformer's performance, namely this component becomes closer to 1 as $\text{SNR}_{\text{BF}}$ increases, which leads to better estimation of the desired microphone signals $\mathbf{d}$ from this component.

### C. Distortion with spatial characteristic ($\mathbf{y}_{s\,\text{phantom}}$)

The second component in (28), namely $\mathbf{y}_{s\,\text{phantom}}$, is a scaled version of the desired source signal, but relocated to the undesired source position. Therefore, this component is considered to be a spatial distortion. It is clear from (28) that when $\text{SNR}_{\text{BF}} \gg 1$ and $|G| > 1$, this component is proportional to $\frac{1}{G}$, implying that this component diminishes as $G$ increases.

### D. Spatially distorted noise component ($\mathbf{y}_{u\,\text{phantom}}$)

The third component in (28), namely $\mathbf{y}_{u\,\text{phantom}}$, is a scaled version of the undesired source, but relocated to the desired source position. Therefore, this component is considered to be an artificial interference, because it does not represent a real source in the room. It is clear from (28) that when $\text{SNR}_{\text{BF}} \gg 1$ and $|G| > 1$, this component is proportional to $\frac{1}{G}$, implying that this component also diminishes as $G$ increases.

### E. Attenuated noise component ($\mathbf{y}_u$)

The fourth component in (28), namely $\mathbf{y}_u$, is a scaled version of the undesired source from its original direction. This component diminishes as $\text{SNR}_{\text{BF}}$ increases.

### F. Noise reduction and signal distortion trade-off

In many traditional methods for noise attenuation, such as those mentioned in Sec. I, there is a trade-off between noise reduction and signal distortion, namely, as the noise reduction increases the distortion level also increases and vice versa. This trade-off is investigated in this section for the proposed approach.

Assuming $\text{SNR}_{\text{BF}} \gg 1$ and $|G| \gg 1$, the source signal estimate from (24) reduces to:

$$\hat{s} \approx s \cdot \mathbf{w}^H \mathbf{h}. \tag{29}$$

Note that even when $\text{SNR}_{\text{BF}}$ is high the estimate of $s$ is distorted by $\mathbf{w}^H\mathbf{h}$. Now, using $\hat{s}$ in (29) to estimate the ATF $\mathbf{h}$ with (16) leads to:

$$\hat{\mathbf{h}} \approx \frac{\mathbf{h}}{\mathbf{w}^H\mathbf{h}}. \tag{30}$$

This result also shows that the estimate of $\mathbf{h}$ is distorted even under high $\text{SNR}_{\text{BF}}$. Next, by substituting (29) and (30) into (23), the following estimate of the desired microphone signal $\mathbf{d}$, assuming that $\mathbf{w}^H\mathbf{h}$ has an inverse (i.e., $\mathbf{w}^H\mathbf{h} \neq 0$) is given:

$$\mathbf{y} \approx s \cdot \mathbf{w}^H\mathbf{h} \cdot \frac{\mathbf{h}}{\mathbf{w}^H\mathbf{h}} = s\mathbf{h} = \mathbf{d}. \tag{31}$$

This is a very important result, which shows that when $\text{SNR}_{\text{BF}}$ is high, the estimate of the desired signal at the microphone is approximately distortionless, even though the estimates of $s$ and $\mathbf{h}$ may carry significant distortion. Furthermore, according to subsections V-B, V-C, V-D, and V-E, it can be deduced that there is no trade-off between noise reduction and signal distortion for this specific acoustic scene, namely, higher noise attenuation at the beamformer's output leads to improved performance in terms of both noise reduction and signal distortion. These insights provide the motivation to find the beamformer which maximizes $\text{SNR}_{\text{BF}}$.

### G. Limitations

Clearly, alongside the advantages of the proposed approach, there are some limitations. Observing (28), the limitations of the proposed approach can be outlined:

- To approximately achieve the result from (31) $G$ needs to have a high value. According to the definition of $G$, to get that high value, spatial separation between the ATFs of the desired and the undesired sources is required. This result predicts that the proposed approach will have limited noise attenuation in the case of diffuse noise.
- In the proposed approach, it is assumed that the desired signal $\mathbf{d}$ is given by the multiplication of the desired source signal, $s$, and the corresponding ATF, $\mathbf{h}$. Later in this paper, the ATF will be estimated in the STFT domain using time averaging (for each frequency). In order to achieve the multiplicative transfer function approximation, the length of the STFT window should be large enough [29]. Therefore, a limit on the time duration of the room impulse response, is required as prior information for the proposed approach.

These two limitations can cause errors and artifacts when the required conditions are violated, and their influence should be further investigated in future work.

## VI. Frequency smoothing for enhancing ATF estimation

In the current work, the ATF was estimated using averaging in the STFT domain by assuming the multiplicative transfer function (MTF) approximation [29]. As shown in [29], the ATF estimate may suffer from numerical errors. Later in this paper, these errors will be shown to have a distinct effect on the signal perception. In this section, these errors are investigated, and a perceptually effective way to reduce them is suggested.

We first assume, for simplicity, that the desired source signal is available. This means that (15) can be approximated in practice by employing the MTF approximation and applying averaging in the STFT domain. Therefore, the observation signal ($\mathbf{x}$ from (2)) and the desired source signal ($s$ from (11)) are presented in the STFT domain [35] as follows:

$$\mathbf{x}_{p,k} = \mathbf{d}_{p,k} + \mathbf{n}_{p,k}, \tag{32}$$

where $p$ represents the time frame index and $k$ represents the frequency index. Then, assuming that there are $N_t$ available time frames for each frequency index, the observation matrix:

$$\mathbf{x}_k = [\mathbf{x}_{1,k}, \mathbf{x}_{2,k}, ..., \mathbf{x}_{N_t,k}] = \mathbf{d}_k + \mathbf{n}_k, \tag{33}$$

where $\mathbf{d}_k = [\mathbf{d}_{1,k}, \mathbf{d}_{2,k}, ..., \mathbf{d}_{N_t,k}]$ and $\mathbf{n}_k = [\mathbf{n}_{1,k}, \mathbf{n}_{2,k}, ..., \mathbf{n}_{N_t,k}]$, and the desired source signal vector:

$$\mathbf{s}_k = [s_{1,k}, s_{2,k}, ..., s_{N_t,k}]^T, \tag{34}$$

are defined for each frequency index $k$. Here, $s_{p,k}$ denotes the source signal in the STFT domain. Finally, the numerical approximation of (15) for each frequency index $k$ is given by [29]:

$$\tilde{\mathbf{h}}_k = \frac{\mathbf{x}_k \mathbf{s}_k^*}{\|\mathbf{s}_k\|_2^2}, \tag{35}$$

where the numerator is the estimate of $S_{s\mathbf{x}}$, the denominator is the estimate of $S_{ss}$ and $^*$ represents the complex conjugate. By substituting the explicit expression of (33) into (35) the following is obtained:

$$\tilde{\mathbf{h}}_k = \frac{\mathbf{d}_k \cdot \mathbf{s}_k^*}{\|\mathbf{s}_k\|_2^2} + \frac{\mathbf{n}_k \cdot \mathbf{s}_k^*}{\|\mathbf{s}_k\|_2^2}. \tag{36}$$

The first component of (36), namely $\frac{\mathbf{d}_k \cdot \mathbf{s}_k^*}{\|\mathbf{s}_k\|_2^2}$, is the estimate of (15), which may suffer from error due to the limited window length of the STFT and due to other numerical errors. The second component of (36) is the cross term between $s$ and $\mathbf{n}$, which diminishes because $s$ and $\mathbf{n}$ are assumed to be uncorrelated. However, because of the finite number of time frames this cross term will not be completely eliminated. Therefore, (35) may also suffer from this cross term residual error. In total (35) can be approximated by:

$$\tilde{\mathbf{h}}_k \approx \mathrm{Em}_k \cdot \mathbf{h}_k + \mathbf{Er}_k, \tag{37}$$

where $\mathrm{Em}_k$ is a multiplicative error representation of the first term in (36), $\mathbf{Er}_k$ represents the cross term residual error and $\mathbf{h}_k$ represents the true ATF at frequency index $k$. These errors also occur when the MTF approximation is used to estimate (16).

In terms of spatial perception, when (37) is multiplied by the desired source signal to estimate $\mathbf{d}$, the cross term residual error that is multiplied by the desired source signal will be shown to have a significant effect that dramatically deteriorates the auditory experience. Moreover, this phenomenon may damage the spatial cues available in the estimate of $\mathbf{d}$. The cross term residual error becomes less significant as the number of available time frames increases [29].

In order to overcome the effect of the cross term residual error for signals with a short time duration, the following assumptions are used. First, the cross term residual error, $\mathbf{Er}$, at the frequency index $k$ is assumed to be uncorrelated with $\mathbf{Er}$ at the different frequency indices. The second assumption is based on the perceptual insight that frequency smoothing of the ATF may not lead to a significant change in auditory perception [36]. Therefore, frequency smoothing is used as follows: for $K = 2 \cdot L_F + 1$ frequency indices around the frequency index $k$, the band observation matrix:

$$\mathbf{X}_k = [\mathbf{x}_{(k-L_F)}, \mathbf{x}_{(k-L_F+1)}, ..., \mathbf{x}_{(k+L_F)}], \tag{38}$$

and the band desired signal vector:

$$\mathbf{S}_k = [\mathbf{s}_{(k-L_F)}^T, \mathbf{s}_{(k-L_F+1)}^T, ..., \mathbf{s}_{(k+L_F)}^T]^T, \tag{39}$$

are defined, where $L_F$ is the number of frequency indices which are used in the band. When $k < L_F$ or $k > L_F$ only the available frequency indices are used, namely, $1, 2, ..., k + L_F$ and $k - L_F, k - L_F + 1, ..., F$, respectively, where $F$ is the number of frequency indices that are available. Then, the estimate of the ATF at the frequency index $k$ using frequency smoothing is given by:

$$\tilde{\mathbf{h}}_k = \frac{\mathbf{X}_k \mathbf{S}_k^*}{\|\mathbf{S}_k\|_2^2}. \tag{40}$$

The solution given in (40) may be understood as a weighted arithmetic mean of the ATF $\mathbf{h}$ in the frequency band $K$ around the frequency index $k$. As a larger $K$ is used, the cross term residual error decreases, but, because the ATF is averaged over more frequencies, the multiplicative error may increase. As mentioned above, for some values of $K$ this frequency smoothing may not deteriorate the auditory perception. On the other hand, due to the extra averaging over the additional $K-1$ frequency indices, the cross term residual error may be attenuated.

Therefore, in order to implement the proposed approach, first the source signal is estimated using a beamformer, as presented in Sec. IV-A, and then, the ATF is estimated as suggested in (40) for a specific value of $K$. Finally, these two estimates are multiplied as follows:

$$\mathbf{y} = \hat{s} \cdot \tilde{\mathbf{h}}_{\hat{s}}, \tag{41}$$

where $\hat{s}$ represents the desired source signal estimate and $\tilde{\mathbf{h}}_{\hat{s}}$ is the ATF estimation as presented in (40), where instead of using the true desired source signal, its estimate is used (the frequency index $k$ was omitted for brevity). The estimate in (41) is an approximation of (28) (for the considered acoustic scene), which, besides the noise and the distortion components that were presented in (28), also suffers from multiplicative and cross term residual errors as presented above.

## VII. Objective measures of performance

In this section, objective measures are formulated in order to evaluate the proposed approach.

First, a measure which calculates the distortion of a vector in comparison with a reference vector is proposed in order to evaluate the distortion in the estimated source signal and ATF, as mentioned in the previous sections. Given vector $\mathbf{z}$ and a reference vector $\mathbf{z}_{ref}$, where both are presented in the frequency domain, the normalized average over $F$ frequency points of the squared Euclidean distance between these vectors is defined as:

$$\mathit{Dist}(\mathbf{z}, \mathbf{z}_{ref}) = \frac{\sum_{f=1}^{F} \left\|\mathbf{z} - \mathbf{z}_{ref}\right\|^2}{\sum_{f=1}^{F} \left\|\mathbf{z}_{ref}\right\|^2}. \tag{42}$$

As *Dist* becomes smaller, the distortion level in $\mathbf{z}$ is reduced, which means that $\mathbf{z}$ becomes more similar to $\mathbf{z}_{ref}$.

The first measure defined using *Dist* is the total distortion of the desired microphone signal estimate $\mathbf{y}$. Recall from Sec. VI that (41) suffers from theoretical distortion due to $\mathbf{y}_s$ and $\mathbf{y}_{s\,\text{phantom}}$, and practical distortion due to the multiplicative and cross term residual errors as detailed in Sec. VI. In order to quantify this distortion, the desired microphone signal, $\mathbf{d}$, is taken as a reference. As the computation of $\mathbf{y}$ involves beamforming (in the first stage of estimating $s$, see Sec. IV-A), this distortion measure is defined for a given beamformer. First, the total distortion of the MVDR is given by:

$$\mathrm{TD}_{\mathrm{MVDR}} = \mathit{Dist}\big(\hat{s} \cdot \tilde{\mathbf{h}}_{\hat{s}}|_{u_1=0}, \mathbf{d}\big), \tag{43}$$

where $\hat{s}$ is computed using an MVDR beamformer as in (13). The total distortion of the Bartlet based processing, namely $\mathrm{TD}_{\mathrm{Bartlet}}$, is defined similarly.

Another variation of the measure *Dist* is used to assess the distortion due to the practical estimation of the ATF as

mentioned in Sec. VI. Further, as presented in Sec. VI, the ATF estimate using the MTF approximation may suffer from multiplicative and cross term residual errors. Therefore, the model of $\mathbf{y}$ from 28 may suffer from additional errors. In order to determine the source of the error, a reference method is used that assumes that the source is given, and the ATF is estimated by using (35). Then, the desired microphone signals for this case are given by:

$$\mathbf{y}_{\text{known } s} = s \cdot \tilde{\mathbf{h}} = s \cdot \text{Em} \cdot \mathbf{h} + s \cdot \mathbf{Er}. \tag{44}$$

Therefore, in order to assess the distortion in $\mathbf{y}_{\text{known } s}$ due to the practical estimation process of the ATF, the following measure is used:

$$\text{TD}_{\text{known } s} = \textit{Dist}(\mathbf{y}_{\text{known } s}, \mathbf{d}). \tag{45}$$

Then, in order to segregate between the contributions of the different errors (**Er** and Em), the following measures are used:

$$\text{Mult\_error} = \textit{Dist}(s \cdot \text{Em} \cdot \mathbf{h}, \mathbf{d}), \tag{46}$$

and

$$\begin{aligned} \text{Res\_error} &= \textit{Dist}(\mathbf{d} + s \cdot \mathbf{Er}, \mathbf{d}) \\ &= \frac{\sum_{f=1}^{F} \|s \cdot \mathbf{Er}\|^2}{\sum_{f=1}^{F} \|\mathbf{d}\|^2}. \end{aligned} \tag{47}$$

Finally, in order to quantify the noise reduction in the processed signal using the MVDR based processing, the following noise gain measure is used:

$$\text{NG}_{\text{MVDR}} = \frac{\sum_{f=1}^{F} \left\| \hat{s} \cdot \tilde{\mathbf{h}}_{\hat{s}}|_{s=0} \right\|^2}{\sum_{f=1}^{F} \|\mathbf{n}\|^2}, \tag{48}$$

where $\hat{s}$ is computed using an MVDR beamformer as in (13). The noise gain for the processed signals using Bartlet based processing, namely $\text{NG}_{\text{Bartlet}}$, is calculated similarly. The noise component of $\mathbf{y}_{\text{known } \mathbf{h}}$ from (9) is used as a reference method for the noise gain performance of the proposed approach:

$$\text{NG}_{\text{known } \mathbf{h}} = \left. \frac{\sum_{f=1}^{F} \|\mathbf{y}_{\text{known } \mathbf{h}}\|^2}{\sum_{f=1}^{F} \|\mathbf{n}\|^2} \right|_{s=0}. \tag{49}$$

As NG becomes smaller, the noise components are better attenuated.

## VIII. Simulation study

In this section, a comprehensive Monte Carlo simulating, validating some of the theories using the objective measures of performance, is reported. Some prior information about the acoustic scene and microphone array will be used to prevent errors from auxiliary algorithms:

- The DOA of the desired source is assumed to be given. In practice, for the case of a reverberant environment, the DOA should be estimated, for example, using [37], [38]
- The noise covariance matrix is assumed to be given. In practice, this matrix should be estimated. In the case of stationary noise source, voice activity detector (VAD) can be used to detect speech absence time segments, and then time averaging can be applied to estimate this matrix (as presented in [1]). In the case of a non-stationary interfering source, e.g. a speaker, association methods can be used to detect the time segments of the interfering speaker [31].
- The time duration of the room impulse response is assumed to be given. In practice it can be estimated using [39], [40].

Each acoustic scene in this study was constructed as presented next. In addition, all the values of the parameters that will be presented next are reported in Table. I.

### A. Setup

The acoustic scene simulated using the image method [41] consists of a desired point source, with source signal $s(t)$, and an interfering point source, with source signal $u(t)$ (where $t$ represents the time index), which are located in a rectangular room with reverberation time $T_{60}$ and critical distance $r_c$. A microphone array is also located in the room, and the ATFs $\mathbf{h}$ and $\mathbf{r}$, representing the transfer functions from the desired source and the interference, respectively, to the microphone array, are also simulated. The length of $\mathbf{h}$ in the time domain is denoted as $\mathcal{L}_{\mathbf{h}}$. The microphone array is located at $(x_0, y_0, z_0)$ in the room, with microphone signals sampling the sound pressure with a sampling frequency of 16 kHz. The microphone array, the desired source and the interference have the same position on the $z$-axis ($z = 1.7\,m$), and the locations of the sources in the $x-y$ plane are presented by using the couple $(r_s, \Phi_s)$ and $(r_u, \Phi_u)$, which represent the distance and the angle of the sources relative to the microphone array's position, as shown in Fig. 1. Two additional noise sources were also simulated. The first one is sensor noise, simulated as white noise, with $\tilde{\mathbf{n}}_{sn}$ representing the microphone noise signals in the time domain. The second one is reverberant noise, i.e., noise whose reverberant part is much more dominant than its direct part as they are captured at the array's microphones. This reverberant noise is generated by a pink noise source that is located at the point $(x, y, z)$ in the room that is most distant from the origin, where $\tilde{\mathbf{n}}_{pn}$ represents the reverberant noise signals as they are captured at the microphone array in the frequency domain.

### B. Methodology

Having generated microphone signals as presented in Sec. VIII-A, the spatial correlation matrices are computed as follows. The matrix $\mathbf{P}$ from (3) for every frequency $f$ can be described as composed of three matrices: $\mathbf{P} = \mathbf{P}_{pn} + \mathbf{P}_{sn} + \mathbf{P}_u$, where matrices $\mathbf{P}_{sn}$, $\mathbf{P}_{pn}$ and $\mathbf{P}_u$ are the correlation matrices of the microphone sensor noise, the reverberant noise and the interference at frequency $f$, respectively, and are assumed to be uncorrelated. In order to estimate the matrices $\mathbf{P}_{sn}$ and $\mathbf{P}_{pn}$ for each frequency $f$, each entry of the vectors $\tilde{\mathbf{n}}_{sn}$ and $\tilde{\mathbf{n}}_{pn}$ is presented in the time-frequency domain (using a Hanning window of length 512 samples, $50\%$ overlap, FFT of length $\mathcal{L}_2$ and sampling frequency of 16 kHz). Then, for each frequency $f$, the outer products of these vectors are averaged over all the available time frames (assuming 5 seconds are available for averaging). The matrix $\mathbf{P}_u$ is calculated using the

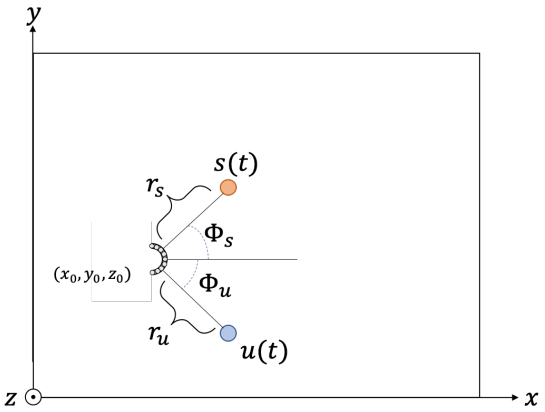


Fig. 1. Top projection of the acoustic scene, showing the two sources and the semi circular array.

model $\mathbf{P}_u = S_{u_1u_1} \cdot \mathbf{r}_1 \mathbf{r}_1^H$ for each frequency $f$, where $S_{u_1u_1}$ is the average auto spectrum of the undesired source signal calculated in the STFT domain with similarity as mentioned above.

In order to set the level of the microphone sensor noise and the reverberant noise, the following observations are made. First, $\tilde{\mathbf{n}}_{sn}$ satisfies $E\{\tilde{\mathbf{n}}_{sn}\tilde{\mathbf{n}}_{sn}^H\} = \sigma_n^2 \mathbf{I}_I$, where $\sigma_n^2$ is the variance of the noise at each microphone of the array. Second, the correlation matrix of the reverberant noise is a non diagonal matrix in general. Hence, to effectively set the reverberant noise level, only the mean of the matrix diagonal (namely the mean over diag$(E\{\tilde{\mathbf{n}}_{pn}\tilde{\mathbf{n}}_{pn}^H\})$), is used and denoted as $\sigma_{pn}^2$. Following these definitions, three different measures for the noise power ratio were defined: signal to interference ratio (SIR), defined as SIR $= \frac{\sum_{t=1}^{T} s^2(t)}{\sum_{t=1}^{T} u^2(t)}$, signal to sensor noise ratio (SSNR), defined as SSNR $= \frac{\frac{1}{T}\sum_{t=1}^{T} s^2(t)}{\sigma_{sn}^2}$, and signal to reverberant noise ratio (SRNR), defined as SRNR $= \frac{\frac{1}{T}\sum_{t=1}^{T} s^2(t)}{\sigma_{pn}^2}$.

Having defined the levels of all the signal and room parameters for each realization of the Monte Carlo simulation (see details below), the microphone signals can be computed. From the microphone signals, the desired source is estimated using the MVDR, the Bartlet and the oracle beamformers as detailed in (13), (14) and (7), respectively. In the next stage, the ATF is estimated using (40) (with a Hanning window of length $\mathcal{L}_1$ samples, 75 % overlap, FFT of length $\mathcal{L}_2$ and sampling frequency of 16 kHz), where the beamformers' outputs were used instead of $s$. The source estimate and the corresponding estimate of the ATF were multiplied as presented in (23) in order to calculate the estimate of the desired microphone signal $\mathbf{y}$ for the case of the Bartlet and the MVDR beamformers.

In order to compute $\mathbf{y}_{\text{known }\mathbf{h}}$ from (9) the output of the oracle beamformer and the true ATF $\mathbf{h}$ were multiplied. The reference computation $\mathbf{y}_{\text{known }s}$ from (44) was also calculated using (40).

The fixed and the varying parameters for this Monte Carlo simulation are presented in Table. I. The Monte Carlo simulation includes the following varying parameters: three different rooms, four different source positions and two different array configurations, which lead to 24 realizations in total. In each realization the objective measures from Sec. VII were calculated for $K = 2 \cdot L_F + 1$ where $L_F = 1, 2, 3, ..., 30$.

TABLE I
DETAILS AND PARAMETERS OF THE MONTE CARLO SIMULATION.

| Independent variable | Description |
|---|---|
| Desired source signal ($s(t)$) | A 10 seconds speech utterance taken from the TIMIT corpus [42]. |
| Interference signal ($u(t)$) | A 10 seconds speech utterance taken from the TIMIT corpus [42]. |
| SSNR | SSNR $= 30$ dB |
| SIR | SIR $= 0$ dB |
| SRNR | SRNR $= 20$ dB |
| Room parameters | Three different rooms:<br>The first room: dimensions $x \times y \times z = 8\,m \times 6.5\,m \times 3\,m$, $T_{60} = 0.4\,s$, $r_c = 0.86\,m$, $\mathcal{L}_{\mathbf{h}} = 0.4\,s$.<br>The second room: dimensions $x \times y \times z = 6\,m \times 5\,m \times 3\,m$, $T_{60} = 0.6\,s$, $r_c = 0.51\,m$, $\mathcal{L}_{\mathbf{h}} = 0.6\,s$.<br>The third room: dimensions $x \times y \times z = 10\,m \times 8\,m \times 3\,m$, $T_{60} = 0.8\,s$, $r_c = 0.73\,m$, $\mathcal{L}_{\mathbf{h}} = 0.8\,s$ |
| Array position | $(x_0, y_0, z_0) = (2, 2, 1.7)$ m |
| Source positions | Four different source positions: for all the positions $r_s = r_u = 1\,m$ and the angles are: $(\Phi_s, \Phi_u) = (0°, -45°)$, $(\Phi_s, \Phi_u) = (0°, -90°)$, $(\Phi_s, \Phi_u) = (45°, -45°)$ and $(\Phi_s, \Phi_u) = (45°, -90°)$. |
| FFT length $\mathcal{L}_2$ | $\mathcal{L}_2 = 4 \cdot \mathcal{L}_{\mathbf{h}}$ |
| Window length $\mathcal{L}_1$ | $\mathcal{L}_1 = 2 \cdot \mathcal{L}_{\mathbf{h}}$ |
| Array configurations | The first array:<br>spherical microphone array with 36 microphones positioned nearly-uniformly around a rigid sphere with radius $0.042\,m$.<br>The second array:<br>semi-circular array with radius $0.042\,m$ and 10 microphones in free field. |

### C. Distortion analysis

In Fig. 2, the total distortion of the MVDR, the Bartlet and the reference method (which assumes that the desired source is known) are presented for $K = 1$, namely, no frequency smoothing is used. These results are given by averaging over all the realizations of the Monte-Carlo simulation. As presented in this figure, the total distortion of $\mathbf{y}_{\text{known }s}$ is not zero (i.e., not minus infinity in dB). This means that there are some errors in the ATF estimate as mentioned in Sec. VI. Later in this analysis, these errors will be investigated as a function of the frequency smoothing coefficient ($K$). The total distortion of the MVDR almost attains the total distortion of $\mathbf{y}_{\text{known }s}$ for both array configurations. It can therefore be concluded from this result that the total distortion of the MVDR beamformer is dominated by the errors due to the ATF estimation process, rather than the distortion due to the estimation of $s$ (SNR$_{\text{BF}}$ dependent). It can also be seen that the Bartlet beamformer displays relatively high total distortion compared to the total distortion of $\mathbf{y}_{\text{known }s}$. This result implies that the error in the ATF estimation using the output of the Bartlet beamformer is relatively low compared to the distortion due to the estimation of $s$ in this case. In the next sections a deeper investigation is conducted in order to determine the

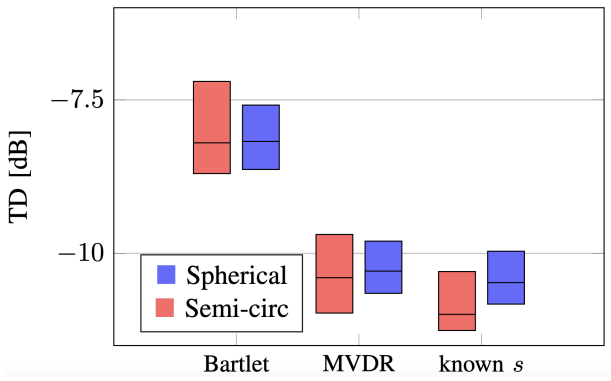


Fig. 2. The total distortion in dB of the different methods, computed using (43) and (45) for both array configurations and averaged over all other Monte Carlo simulation conditions. The median line in the box represents the median; the bottom and top edges represent the 25th and 75th percentiles, respectively.

source of the errors in the MVDR method and the known $s$ reference method.

### *D. The trade-off between the multiplicative and the cross term residual errors*

As presented in (37), the estimate of the ATF using the MTF approximation may suffer from multiplicative and cross term residual errors. The results of the Monte Carlo simulation which investigates these errors, assuming that the desired source signal is known, are presented in Fig. 3. As shown in this figure, for $K = 1$, namely, no frequency smoothing is used, the total distortion under the conditions of this experiment is mostly dominated by the cross term residual error, where the multiplicative error is significantly smaller. Therefore, it can be concluded that the total distortion, which is presented in Fig. 2 for the reference method, is mostly dominated by the cross term residual error. As $K$ increases, the weighted arithmetic mean of the ATF $\mathbf{h}$, as detailed in Sec. VI, uses more frequencies for averaging, which may lead to a less accurate ATF estimate. On the other hand, as $K$ increases, the cross term residual error may diminish as also explained in Sec. VI. These phenomena are clearly shown in Fig. 3. The results shown in Fig. 3 are presented for the spherical array, while the results of the semi-circular array are similar, because the ideal and not the estimated source signal is used for both cases. The analysis presented here clearly shows the potential benefit of frequency smoothing for reducing the total distortion in the estimation of the ATF. This is studied further next.

### *E. The effect of frequency smoothing on distortion*

In Fig. 4 the total distortion of the MVDR, the Bartlet and the reference method, where the desired signal is known, are presented as a function of the frequency smoothing parameter $K$. As shown in this figure, the graphs for the different methods have the same trend as a function of $K$, which presents a decrease of the total distortion up to an optimal $K$ ($K = 9$ in this case), and then an increase of the total distortion. This trend implies that the trade-off between the multiplicative error

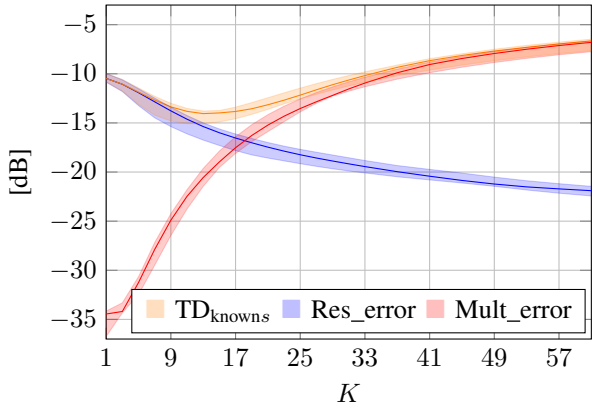


Fig. 3. The measures $\text{TD}_{\text{known}s}$, Res_error and Mult_error represented in dB as a function of $K$ for the spherical array. The results for the semi-circular array are similar. The solid lines represent the median of the Monte Carlo simulation results and the upper and the lower edges of the transparent filled areas represent the 25th and 75th percentiles, respectively.

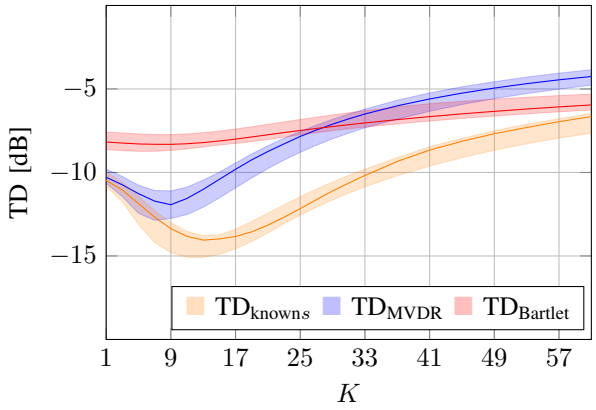


Fig. 4. The total distortion of the Bartlet, MVDR and known $s$ methods presented in dB as a function of $K$ for the spherical array. The results for the semi-circular array are similar.

and the cross term error also occurs in the MVDR and the Bartlet methods. This result reinforces the conclusion that the total distortion of the MVDR method is mostly dominated by the cross term residual error for $K = 1$. For large values of $K$ the figure shows that there is a significant difference between the MVDR method and the known $s$ reference method. This difference could be the result of two effects: the first one is the different multiplicative error in each method, and the second one could stem from the insight that the cross term residual error is not dominant in comparison to the error due to the $\mathbf{y}_{s\,\text{phantom}}$ component, from (28), of the MVDR output, with the latter providing an additional distortion to the MVDR in comparison to the known $s$ reference method. The Bartlet method seems to be more indifferent to the changes in $K$. This could be due to two reasons: the dominant phantom component ($\mathbf{y}_{s\,\text{phantom}}$) as explained above (for small values of $K$), and low multiplicative error. The results shown in Fig. 4 are presented for the spherical array, while the results of the semi-circular array are similar.

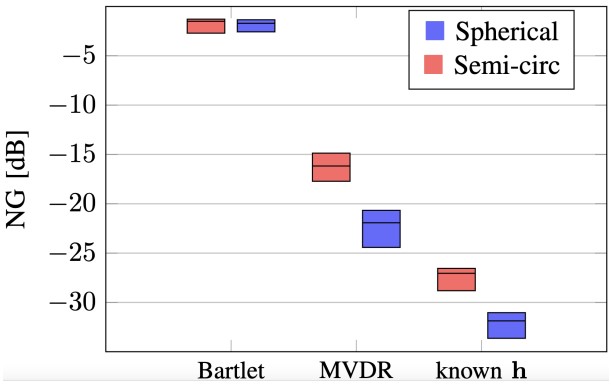


Fig. 5. Monte Carlo simulation results: NG in dB of the different methods.

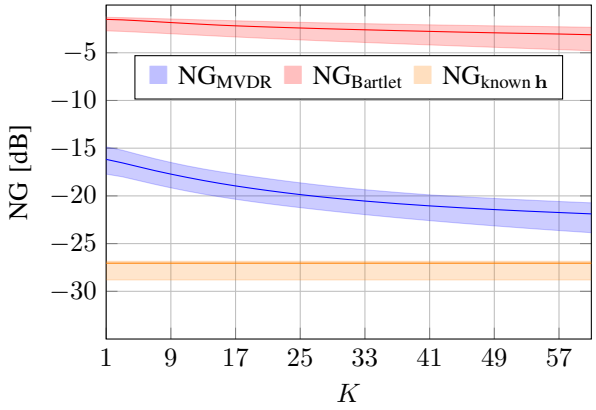


Fig. 6. The noise gain of the Bartlet method, the MVDR method and the reference method that assumes the ATF is known presented in dB as a function of $K$ for the semi-circular array.

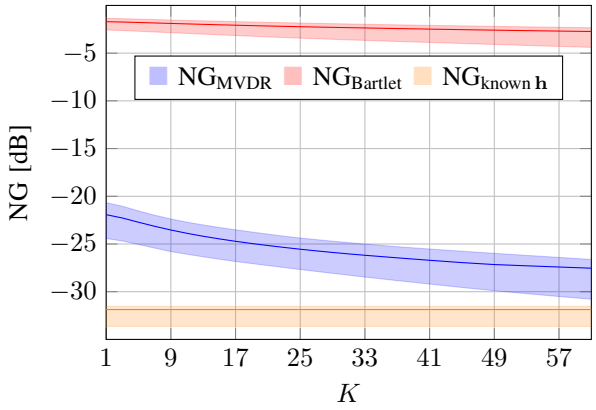


Fig. 7. The noise gain of the Bartlet method, the MVDR method and the reference method that assumes the ATF is known presented in dB as a function of $K$ for the spherical array.

### *F. Noise reduction analysis*

Figure 5 presents the noise gain results for the two array configurations for $K = 1$. As shown in this figure, the Bartlet beamformer barely manages to attenuate the noise. On the other hand, the MVDR beamformer achieves substantial attenuation of the noise components for both arrays. Better attenuation is achieved when using the spherical array due to the larger number of microphones in this array. This can also be seen in the case of $\mathbf{y}_{\text{known }\mathbf{h}}$, where the oracle beamformer (with weights $\mathbf{w}_{\text{oracle}}$) better attenuates the noise in the case of the spherical array. The difference between the noise gain of the MVDR beamformer and $\mathbf{y}_{\text{known }\mathbf{h}}$ is due to two reasons. While the MVDR beamformer includes the components $\mathbf{y}_u$ and $\mathbf{y}_{u\,\text{phantom}}$, as shown in (28), $\mathbf{y}_{\text{known }\mathbf{h}}$ contains only a single noise component that is similar to $\mathbf{y}_{u\,\text{phantom}}$ from (28), leading to less noise in $\mathbf{y}_{\text{known }\mathbf{h}}$. The second reason is the multiplication of the residual noise at the beamformer's output with the cross term residual error from the ATF estimation process. The latter reason will be investigated in the next section.

### *G. The effect of frequency smoothing on noise reduction*

In Figs. 6 and 7 the noise reduction of the MVDR method, the Bartlet method and the reference method that assumes the ATF $\mathbf{h}$ is known are presented as a function of $K$ for the two arrays. As shown in these figures, the noise reduction that the Bartlet method presents only slightly decreases as a function of $K$. This implies that the components $\mathbf{y}_u$ and $\mathbf{y}_{u\,\text{phantom}}$ of the Bartlet method are more dominant than the cross term residual error in the ATF estimate which is multiplied by the undesired component of (24), when the Bartlet beamformer is used. On the other hand, the MVDR method provides significant noise attenuation for both arrays. As $K$ increases, the noise attenuation of the MVDR, for both array configurations, decreases due to the attenuation of the cross term residual error. This shows that the frequency smoothing can also be useful for reducing noise, in addition to reducing distortion.

## IX. Listening tests

In the previous section, the distortion and the noise reduction that were attained at the output of the proposed method were evaluated using objective measures. In this section, the results of a listening test to evaluate the perceptual quality of the processed signals are reported. This subjective evaluation takes into account spatial aspects as well as spectro-temporal aspects.

### *A. Experimental setup*

The listening test was based on a single acoustic scene which was selected from the multiple scenes generated for the objective analysis in the previous section. The selected scene parameters were chosen from Table I as follows: the second room from Table I, and sources position $r_s = r_u = 1\,m$, $(\Phi_s, \Phi_u) = (-45°, 45°)$. For this acoustic scene, both microphone arrays were used.

### *B. Methodology*

The simulation provides the observed signals $\mathbf{x}$ and the desired microphone signals $\mathbf{d}$ from (2). Then, in order to estimate the desired source signal, the Bartlet, (14), and the MVDR, (13), beamformers were applied. Next, the ATF was estimated using the desired source estimates in (40) with different values of parameter $K$ for each beamformer output.

Three different values of parameter $K$ were used; $K = 1$, which is the conventional method for ATF estimation (no frequency smoothing is used), $K = 9$, which leads to the minimum of the total distortion for both beamformers (according to Fig. 4), and $K = 61$, which was chosen to reduce further the residual estimation error at the cost of increased distortion. The latter was found to be preferable in an informal listening test. In order to evaluate the distortion as a function of the frequency smoothing parameter, only the desired component of the desired source estimate was employed, separated from the estimated signal as in (11). These signals were used in the first listening test, which aimed to study the impact of $K$ on perception. The listening test, based on Recommendation ITU-R BS.1534-1 (MUSHRA, MUltiple Stimuli with Hidden Reference and Anchor) [43], was conducted and included three MUSHRA screens for each microphone array. For each array, the effect of the frequency smoothing parameter, $K$, was investigated. Therefore, the MUSHRA screen included the following signals:

- **Ref**: a binaural signal generated only from the desired microphone signals ($\mathbf{d}$).
- **MVDR $\mathbf{K = 1}$**: a binaural signal generated from the output of the MVDR process without the noise components with $K = 1$ (namely, the binaural reproduction due to $\hat{s} \cdot \tilde{\mathbf{h}}_{\hat{s}}|_{u_1=0}$ with $K = 1$).
- **MVDR $\mathbf{K = 9}$**: a binaural signal generated from the output of the MVDR process without the noise components with $K = 9$.
- **MVDR $\mathbf{K = 61}$**: a binaural signal generated from the output of the MVDR process without the noise components with $K = 61$.

A similar test was conducted for the Bartlet beamformer. In this test no anchor was used.

Next, the desired source estimate was multiplied with the corresponding ATF to reproduce the desired microphone signals, but only for $K = 61$. In addition to the proposed approach two reference methods were also used for this listening test. The first method was the TFS method from [44], where the SNR at each time-frequency bin was estimated as suggested in [22]. The second method was the Bilateral filter from [14]. For both reference methods the required statistics regarding the noise was assumed to be given. Both reference methods were applied to the observed signals in order to estimate the desired microphone signals. Note that the TFS method was used only for the spherical array according to the method's requirements. This test evaluated the quality of all the processed signals. For this evaluation the two reference methods which were mentioned above were used. The MUSHRA screen included the following signals:

- **Ref**: a binaural signal generated only from the desired microphone signals ($\mathbf{d}$).
- **MVDR $\mathbf{K = 61}$**: a binaural signal generated from the output of the MVDR process, $\mathbf{y}$ from (41), with $K = 61$.
- **Bartlet $\mathbf{K = 61}$**: a binaural signal generated from the output of the Bartlet process, $\mathbf{y}$ from (41), with $K = 61$.
- **TFS**: a binaural signal generated from applying the TFS method on the observed signals $\mathbf{x}$.
- **Bilateral**: a bilateral filter was applied to the binaural signals generated from the observed signals $\mathbf{x}$.
- **Anchor**: a binaural signal generated from the unprocessed observed signals $\mathbf{x}$.

In order to reproduce the binaural signals for all the signals which were mentioned above, for the case of the spherical microphone array and for the semi-circular array, the methods from [45] and [46] were used, respectively. For these methods, the head related transfer function (HRTF) compilation of the Neumann KU-100 [26] was used. Then, the binaural signals were equalized for a specific set of headphones that were used. 12 normal hearing subjects participated in this experiment. In total, the listening test included six MUSHRA screens (three screens for each microphone array). All signals were played back using the Matlab (MATLAB R2021a) audio recorder. In all MUSHRA screens the participants were asked to rate the overall quality of the signals with respect to the reference signal, on a scale from 0 to 100. Before rating, the participants performed a training task in order to ensure that the instructions were clearly understood and to familiarize the participants with the stimuli.

### C. Results

In Fig. 8 the results of the listening tests that investigated the effect of the frequency smoothing parameter $K$ are presented for each beamformer and microphone array. While the participants were asked to rate the overall quality of the signals, due to the fact that these signals included only the desired signal components, the actual results present the amount of distortion that each signal contains in terms of auditory perception. As shown in this figure, for both arrays and beamformers, as $K$ increases the signals were rated higher. For both arrays and beamformers the median score of the processed signal with $K = 61$ significantly differs ($p < 0.05$) from the other values of $K$. This result shows that the multiplicative error is more tolerated (in terms of auditory perception) than the cross term residual error.

In Fig. 9 the results of the listening tests which investigate the overall quality of the processed signals are presented. As shown in this figure, for both arrays the MVDR with $K = 61$ achieved a higher score than the other reference methods and the anchor ($p < 0.05$), demonstrating the superior quality of the proposed method.

## X. Conclusions

In this paper, an approach for spatial signal enhancement was presented. This approach is based on desired source signal estimation and ATF estimation. The theoretical analysis assumed a simplified scenario with a single interfering source and the absence of diffuse or spatially uncorrelated noise. Under this case, the theoretical analysis showed no trade-off between noise reduction and signal distortion. In addition, frequency smoothing has been proposed, and was shown to control the trade-off between the multiplicative error and the cross term residual error in the ATF estimate. While the objective results showed that the error increases with frequency smoothing, the listening test results showed that the

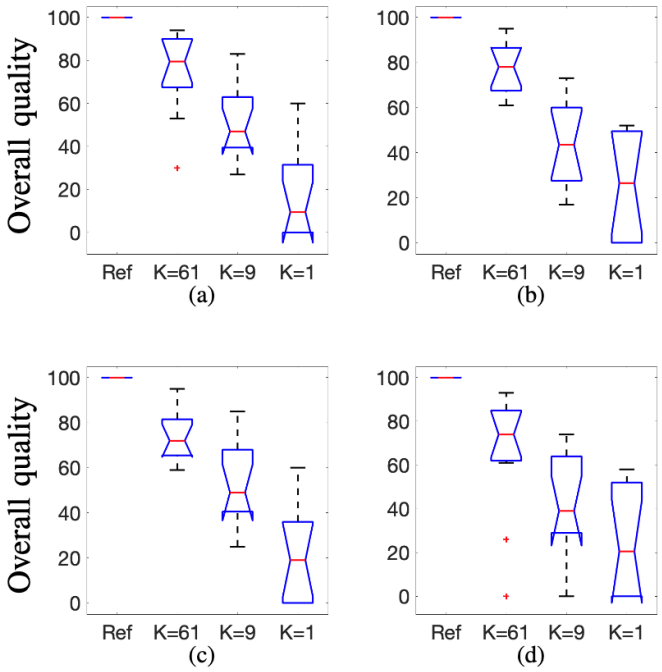


Fig. 8. Ratings as a function of frequency smoothing parameter. (a) MVDR beamformer for the spherical array; (b) Bartlet beamformer for the spherical array; (c) MVDR beamformer for the semi-circular array; (d) Bartlet beamformer for the semi-circular array; Box plot visualization: the median is the middle line; the bottom and top edges represent the 25th and 75th percentiles; the whiskers represent the extreme values, excluding outliers; the notches have been calculated such that boxes with non-overlapping notches have medians which are different at the 95% significance level. Outliers are marked with a red "+".

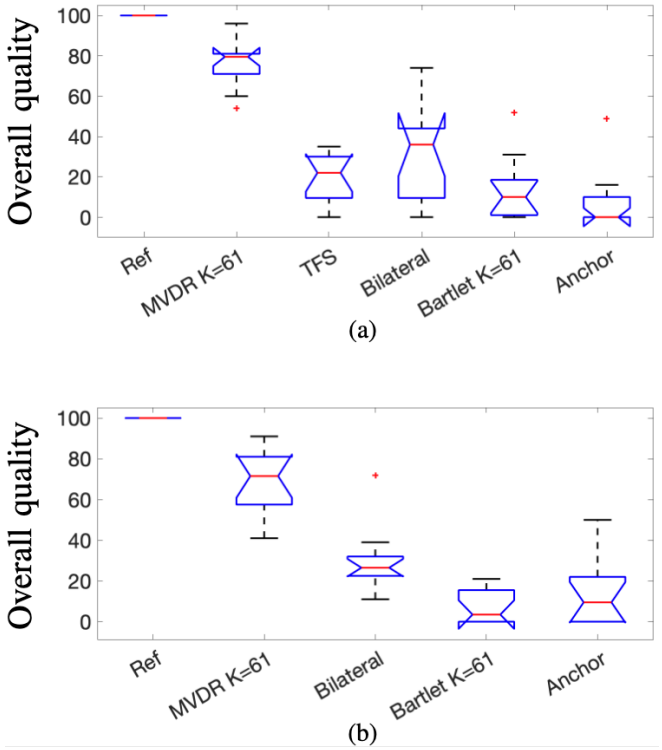


Fig. 9. The results of the overall quality test - (a) presents the results of the spherical array and (b) presents the results of the semi-circular array.

frequency smoothing helps improve the auditory experience. Determination of the frequency smoothing amount for each scenario and frequency index and the effect of errors in noise covariance matrix estimation are proposed for future work.

## REFERENCES


[1] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, 2016.

[2] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2010.

[3] R. Gunther, R. Kazman, and C. MacGregor, "Using 3D sound as a navigational aid in virtual environments," *Behaviour & Information Technology*, vol. 23, no. 6, pp. 435–446, 2004.

[4] S. Mehrotra, W.-g. Chen, Z. Zhang, and P. A. Chou, "Realistic audio in immersive video conferencing," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–4.

[5] V. A. Nguyen, J. Lu, S. Zhao, D. L. Jones, and M. N. Do, "Teleimmersive audio-visual communication using commodity hardware [applications corner]," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 118–136, 2014.

[6] H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[8] ——, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[9] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2162–2176, 2015.

[10] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, 2007.

[11] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Extensions of the binaural MWF with interference reduction preserving the binaural cues of the interfering source," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 241–245.

[12] ——, "Extensions of the binaural MWF with interference reduction preserving the binaural cues of the interfering source," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 241–245.

[13] M. Zohourian and R. Martin, "GSC-based binaural speaker separation preserving spatial cues," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 516–520.

[14] G. Enzner, M. Azarpour, and J. Siska, "Cue-preserving mmse filter for binaural speech enhancement," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[15] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. De Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2011–2023, 2017.

[16] M. Acoustics, "EM32 eigenmike microphone array release notes (v17. 0)," *25 Summit Ave, Summit, NJ 07901, USA*, 2013.

[17] J. Sheaffer, M. Van Walstijn, B. Rafaely, and K. Kowalczyk, "Binaural reproduction of finite difference simulations using spherical array processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2125–2135, 2015.

[18] J. Ahrens and S. Spors, "An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 988–999, 2008.

[19] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[20] N. R. Shabtai and B. Rafaely, "Spherical array beamforming for binaural sound reproduction," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2012, pp. 1–5.

[21] H. Sun, S. Yan, and U. P. Svensson, "Optimal higher order ambisonics encoding with predefined constraints," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 742–754, 2011.

[22] A. Herzog and E. A. Habets, "Direction Preserving Wiener Matrix Filtering for Ambisonic Input-output Systems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 446–450.

[23] ——, "Direction and Reverberation Preserving Noise Reduction of Ambisonics Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2461–2475, 2020.

[24] M. Lugasi and B. Rafaely, "Speech Enhancement Using Masking for Binaural Reproduction of Ambisonics Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1767–1777, 2020.

[25] ——, "Enhancement of Ambisonics signals using time-frequency masking," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.

[26] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

[27] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," *in Proc. 145th Audio Eng. Soc. Convention, New York, NY, USA, 2018, Paper 10129.*, 2020.

[28] C. Borrelli, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "A denoising methodology for higher order Ambisonics recordings," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 451–455.

[29] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[30] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments," *arXiv preprint arXiv:2107.04174*, 2021.

[31] H. Beit-On, M. Lugasi, L. Madmoni, A. Menon, A. Kumar, J. Donley, V. Tourbabin, and B. Rafaely, "Audio Signal Processing for Telepresence Based on Wearable Array in Noisy and Dynamic Scenes," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8797–8801.

[32] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[33] A. Yaro, S. Salisu, A. Umar, and M. Musa, "Multiangulation position estimation performance analysis using a Bartlett's Beamforming Method," *Nigerian Journal of Technology*, vol. 36, no. 4, pp. 1155–1161, 2017.

[34] K. J. Keesman, *System identification: an introduction*. Springer Science & Business Media, 2011.

[35] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 55–69, 1980.

[36] H. G. Hassager, F. Gran, and T. Dau, "The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2992–3000, 2016.

[37] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[38] H. Beit-On and B. Rafaely, "Speaker localization using the direct-path dominance test for arbitrary arrays," in *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*. IEEE, 2018, pp. 1–4.

[39] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010, pp. 1–4.

[40] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[41] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa Timit Acoustic-Phonetic Continuous Speech Corpus CD-ROM {TIMIT}," ,, Tech. Rep., 1993.

[43] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[44] M. Lugasi and B. Rafaely, "Speech enhancement using masking for binaural reproduction of ambisonics signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1767–1777, 2020.

[45] A. Avni and B. Rafaely, "Interaural cross correlation and spatial correlation in a sound field represented by spherical harmonics," in *Proc. Ambisonics Symposium*, 2009.

[46] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-based Binaural Reproduction by Matching of Binaural Signals," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.