A ROBUST REGION-BASED NEAR-FIELD BEAMFORMER

Jorge Martinez[1], *Nikolay Gaubitch*[1] *and W. Bastiaan Kleijn*[1,2]

[1] Circuits and Systems, Delft University of Technology, The Netherlands
[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

In this paper, a broadband region-based near-field beamforming algorithm is proposed and demonstrated for acoustic applications. We use an eigenfilter structure with a minimum-energy cost function based on desired and undesired near-field regions. Robustness is thus achieved by focusing on signals generated from desired zones in space while rejecting signals from undesired zones. This construction leads to a linear matrix pencil formulated in terms of the array gain to these desired and undesired zones. We include a far-field model as part of the rejection zones that further improves performance in reverberant environments. We demonstrate the robustness of the algorithm in simulated and real scenarios.

Index Terms— near-field beamformer, generalized eigenvalue problem, robust beamformer, microphone arrays, reverberation.

1. INTRODUCTION

Beamforming is an important technique in microphone-array technologies [1]. Examples include hands-free speech communication systems such as car speakerphones, teleconference systems, hearing-aids, mobile telephony, and voice-commanded systems such as smart TVs, navigation systems and personal mobile assistants. In these applications broadband algorithms are necessary since the target acoustic signals have wide bandwidth. The sources and interferers are well within the near-field of the array; far-field designs perform poorly in comparison to near-field designs [1–3]. Moreover the acoustic channel is subject to severe multipath interference in the form of reverberation. Standard beamforming algorithms assume the positions of the microphones and the acoustical sources to be known precisely. However inaccuracies in the positions cause performance degradation in non-robust designs [4, 5], [6, Chap. 1].

Broadband near-field beamforming robust against position errors has gathered relatively little attention in the literature, major propositions are [7–10]. The algorithm in [7] uses a low-rank approximation of the near-field signal space in a spatial region with a linearly constrained minimum variance (LCMV) design. Despite its relatively low-complexity, the complexity of the algorithm increases with the size of the desired spatial region and the bandwidth of the target signals. Moreover, the degrees of freedom of the LCMV are used to provide robustness; reserving degrees of freedom for interference rejection is costly. Ser *et al.* [8], proposed a refining approach for the LCMV design. Assuming the actual target position is close to the focal point, a search is performed in the vicinity of the focal point that maximizes the output power of the beamformer using a non-linear optimization procedure. The resulting algorithm has high complexity as a function of the size of the search space. Furthermore, if the actual target is not found in the search space then beamforming results in target-signal cancellation. In [9], Chen *et al.* propose a mini-max optimal approach to provide robustness against many kinds of parameter mismatch. Although the solution is found by an efficient second order cone programming solver, one can still argue that the complexity is high for applications with low computational power such as hearing-aids. Two propositions are given in [10] based on *diagonal loading* (see e.g. [6]). One requires a reference steering vector not known *a-priori* to calculate an error cost function. The other directly assumes a reference error. None of the aforementioned methods has an explicit model to account for reverberation (e.g., in the form of mixed near-field and far-field interference), relaying only on their near-field performance. This is an important issue as it has been shown that an optimal near-field beam pattern does not guarantee an optimal far-field response unless explicitly enforced [11]. More importantly, the performance of the above mentioned algorithms has not previously been evaluated in real-life scenarios.

We propose a novel robust near-field beamformer for microphone arrays that incorporates a far-field reverberation model to further improve the performance in real scenarios. Instead of focusing on a single or a discrete set of points in space, desired and rejection regions are defined and properly emphasized using spatial weighting functions. Further, a reverberation model in terms of an *isotropic* far-field is added as part of the interference. An eigenfilter structure [12] is then established, using the response power ratio of the desired to the rejection regions as cost function. This construction leads to a generalized eigenvalue problem on the space of beamformer weights (a linear matrix-pencil). An ordered sequence of orthonormal beamformer weights can then be found that forms a subspace. The beamformer weights in this subspace have a response that focuses on the desired zones, attenuating near-field signals in the undesired zones and reverberation in the form of an isotropic far-field. This approach can then be seen as a generalization to a mixed near-field far-field case of the *maximum-energy* method [13]. To evaluate the robustness of the algorithm we present results in simulated and real scenarios.

2. ROBUST NEAR-FIELD BEAMFORMING

Let an acoustic process, expressed in the frequency domain and taking place at spatial position $\mathbf{x}_s \in \mathbb{R}^3$, be denoted by $S(\mathbf{x}_s, \omega)$, where $\omega = 2\pi f$ represents angular frequency, with f the frequency in Hz. We consider the observation of the process with a microphone array comprising M microphones at positions $\mathbf{x}_m \in \mathbb{R}^3$. The M input signals are combined using a broadband beamformer. The weight functions of the beamformer, $w_m(\omega)$, are arranged in a vector $\mathbf{w}(\omega) \in \mathbb{C}^M$. The propagation function (i.e. free-field Green's function) from the source position $\mathbf{x}_s$ to the microphone position $\mathbf{x}_m$

This work was supported by Google Inc.

is given by [14, p. 311], [15, p. 253], [16, p. 51]

$$v(\mathbf{x}_s, \mathbf{x}_m, \omega) = \frac{\mathrm{e}^{j\omega\|\mathbf{x}_s - \mathbf{x}_m\|/c}}{4\pi\|\mathbf{x}_s - \mathbf{x}_m\|}, \quad (1)$$

where $\|\cdot\|$ is the Euclidean norm and c is the speed of sound. A propagation vector, say $\mathbf{v}(\omega) \in \mathbb{C}^M$, can now be formed by stacking the propagation functions from the source to the microphones,

$$\mathbf{v}(\mathbf{x}_s, \omega) = [v(\mathbf{x}_s, \mathbf{x}_1, \omega), \ldots, v(\mathbf{x}_s, \mathbf{x}_M, \omega)]^{\mathsf{T}}, \quad (2)$$

where $^{\mathsf{T}}$ denotes vector or matrix transposition. Using this notation, the overall beamformer response can be written as

$$Y(\omega) = \int_{\mathcal{V}} \mathbf{w}^{\mathsf{H}}(\omega)\mathbf{v}(\mathbf{x}_s, \omega)S(\mathbf{x}_s, \omega)d\mathbf{x}_s + \mathbf{w}^{\mathsf{H}}(\omega)\mathbf{n}(\omega), \quad (3)$$

where $\mathbf{n}(\omega) \in \mathbb{C}^M$ is a noise vector given by a process uncorrelated with $S(\mathbf{x}_s, \omega)$ (e.g., given by non-acoustic noise like quantization or thermal noise), $\mathcal{V} \subset \mathbb{R}^3$ denotes the zone where the acoustic process is defined, and the superscript $^{\mathsf{H}}$ denotes Hermitian transposition.

Let us consider a region in space, $\mathcal{A}$, that contains acoustic signals we want to emphasize, and a region $\mathcal{B}$ that contains signals we want to suppress. We define a pair of spatial weighting/windowing functions, $g_{\mathcal{A}}(\mathbf{x}_s)$ and $g_{\mathcal{B}}(\mathbf{x}_s)$ respectively, that define and properly weight the two zones. Using this notation, the beamformer response of the emphasized zone is written as

$$Y_{\mathcal{A}}(\omega) = \mathbf{w}^{\mathsf{H}}(\omega)\int_{\mathbb{R}^3} \mathbf{v}(\mathbf{x}_s, \omega)S(\mathbf{x}_s, \omega)g_{\mathcal{A}}(\mathbf{x}_s)d\mathbf{x}_s + \mathbf{w}^{\mathsf{H}}(\omega)\mathbf{n}(\omega). \quad (4)$$

Make $\mathbf{h}_{\mathcal{A}}(\omega) = \int_{\mathbb{R}^3}\mathbf{v}(\mathbf{x}_s, \omega)S(\mathbf{x}_s, \omega)g_{\mathcal{A}}(\mathbf{x}_s)d\mathbf{x}_s$, the power of the beamformer response is then given by

$$\mathrm{E}\{|Y_{\mathcal{A}}(\omega)|^2\} = \mathrm{E}\left\{\left|\mathbf{w}^{\mathsf{H}}(\omega)\mathbf{h}_{\mathcal{A}}(\omega) + \mathbf{w}^{\mathsf{H}}(\omega)\mathbf{n}(\omega)\right|^2\right\}, \quad (5)$$

where $\mathrm{E}\{\cdot\}$ is the expectation operator. Since the noise process given by vector $\mathbf{n}(\omega)$ is assumed uncorrelated with the acoustic process $S(\mathbf{x}_s, \omega)$, their cross-covariance is zero, and (5) can be rewritten as

$$\mathrm{E}\{|Y_{\mathcal{A}}(\omega)|^2\} = \mathbf{w}^{\mathsf{H}}(\omega)\big(\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)\big)\mathbf{w}(\omega), \quad (6)$$

where $\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega)$ and $\mathbf{R}_{\mathbf{nn}}(\omega)$ are the covariance matrices of the signal vector $\mathbf{h}_{\mathcal{A}}$ and the noise vector $\mathbf{n}$ respectively. Equivalent equations are obtained for the response of the suppressed zone. Let us set up the array gain in terms of the power ratio of the desired to the rejection zones, i.e.,

$$\lambda(\omega) = \frac{\mathrm{E}\{|Y_{\mathcal{B}}(\omega)|^2\}}{\mathrm{E}\{|Y_{\mathcal{A}}(\omega)|^2\}} = \frac{\mathbf{w}^{\mathsf{H}}(\omega)\big(\mathbf{R}_{\mathbf{h}_{\mathcal{B}}\mathbf{h}_{\mathcal{B}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)\big)\mathbf{w}(\omega)}{\mathbf{w}^{\mathsf{H}}(\omega)\big(\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)\big)\mathbf{w}(\omega)}. \quad (7)$$

We would like to find the $\mathbf{w}(\omega)$ that minimizes $\lambda(\omega)$, so that signals in zone $\mathcal{A}$ get emphasized and signals in zone $\mathcal{B}$ get attenuated. To do this we first look for stationary points in $\lambda(\omega)$ as a function of $\mathbf{w}(\omega)$. Let $\mathbf{A} = \mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)$ and $\mathbf{B} = \mathbf{R}_{\mathbf{h}_{\mathcal{B}}\mathbf{h}_{\mathcal{B}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)$. Denoting complex conjugation by *, we set the complex gradient with respect to $\mathbf{w}^*$ of $\lambda(\omega)$ to zero [17],

$$\frac{\partial\lambda(\omega)}{\partial\mathbf{w}^*} \triangleq \left(\frac{\partial\lambda(\omega)}{\partial\mathbf{w}_{\Re}} + j\frac{\partial\lambda(\omega)}{\partial\mathbf{w}_{\Im}}\right)/2 = 0, \quad (8)$$

and $\mathbf{w}_{\Re}$ and $\mathbf{w}_{\Im}$ are the real and imaginary parts of $\mathbf{w}$. Then

$$\frac{\partial\lambda(\omega)}{\partial\mathbf{w}^*} = \frac{\left(\frac{\partial}{\partial\mathbf{w}^*}\mathbf{w}^{\mathsf{H}}\mathbf{B}\mathbf{w}\right)\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w} - \left(\frac{\partial}{\partial\mathbf{w}^*}\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w}\right)\mathbf{w}^{\mathsf{H}}\mathbf{B}\mathbf{w}}{2(\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w})^2}. \quad (9)$$

From (7) we have that $\mathbf{w}^{\mathsf{H}}\mathbf{B}\mathbf{w} = \lambda(\omega)\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w}$, so that (8) becomes

$$\frac{\partial\lambda(\omega)}{\partial\mathbf{w}^*} = \frac{(\mathbf{B}\mathbf{w})\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w} - (\mathbf{A}\mathbf{w})\mathbf{w}^{\mathsf{H}}\mathbf{B}\mathbf{w}}{(\mathbf{w}^{H}\mathbf{A}\mathbf{w})^2} = \frac{\mathbf{B}\mathbf{w}}{\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w}} - \lambda(\omega)\frac{\mathbf{A}\mathbf{w}}{\mathbf{w}^{\mathsf{H}}\mathbf{A}\mathbf{w}} = 0. \quad (10)$$

Substituting for matrices $\mathbf{A}$ and $\mathbf{B}$ we obtain

$$\big(\mathbf{R}_{\mathbf{h}_{\mathcal{B}}\mathbf{h}_{\mathcal{B}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)\big)\mathbf{w}(\omega) = \lambda(\omega)\big(\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega)\big)\mathbf{w}(\omega). \quad (11)$$

This is a linear matrix pencil. Its regularity depends on the sum of the covariance matrices involved and therefore on the selection of the spatial weighting functions $g_{\mathcal{A}}(\mathbf{x}_s)$ and $g_{\mathcal{B}}(\mathbf{x}_s)$, the assumed second order statistics of the input signal $S(\mathbf{x}_s, \omega)$ and of the noise vector $\mathbf{n}(\omega)$. Let us now determine the covariance matrices. Each element of $\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega)$ is given by

$$\begin{aligned}\{\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}\}_{ij} &= \mathrm{E}\{h_{i\mathcal{A}}h^*_{j\mathcal{A}}\} \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} v(\mathbf{x}_s, \mathbf{x}_i, \omega)v^*(\mathbf{x}'_s, \mathbf{x}_j, \omega) \\ &\qquad \times \mathrm{E}\left\{S(\mathbf{x}_s, \omega)S^*(\mathbf{x}'_s, \omega)\right\} g_{\mathcal{A}}(\mathbf{x}_s)g^*_{\mathcal{A}}(\mathbf{x}'_s)d\mathbf{x}_s d\mathbf{x}'_s. \quad (12)\end{aligned}$$

If $S(\mathbf{x}_s, \omega)$ is a zero-mean uncorrelated spatial process (e.g., given by spatially independent sources), then

$$\mathrm{E}\left\{S(\mathbf{x}_s, \omega)S^*(\mathbf{x}'_s, \omega)\right\} = \begin{cases} 0 & \text{if } \mathbf{x}_s \neq \mathbf{x}'_s \\ \sigma_s^2(\mathbf{x}_s, \omega) & \text{otherwise} \end{cases}, \quad (13)$$

where $\sigma_s^2(\mathbf{x}_s, \omega)$ is the spatial variance function of the process. A natural choice is to make $\sigma_s^2(\mathbf{x}_s, \omega)$ uniform, which corresponds to sources being anywhere with equal probability. Their importance within the desired and undesired zones is then determined by the weighting functions. Normalizing the spatial variance (i.e., $\sigma_s = 1$) (12) becomes

$$\{\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}\}_{ij} = \int_{\mathbb{R}^3} v(\mathbf{x}_s, \mathbf{x}_i, \omega)v^*(\mathbf{x}_s, \mathbf{x}_j, \omega)|g_{\mathcal{A}}(\mathbf{x}_s)|^2 d\mathbf{x}_s, \quad (14)$$

with

$$v(\mathbf{x}_s, \mathbf{x}_i, \omega)v^*(\mathbf{x}_s, \mathbf{x}_j, \omega) = \frac{\mathrm{e}^{j(\omega/c)(\|\mathbf{x}_s - \mathbf{x}_i\| - \|\mathbf{x}_s - \mathbf{x}_j\|)}}{16\pi^2\|\mathbf{x}_s - \mathbf{x}_i\|\|\mathbf{x}_s - \mathbf{x}_j\|}. \quad (15)$$

We select spherically symmetric 3-D Gaussian functions as spatial weighting functions. We set

$$g_{\mathcal{A}}(\mathbf{x}_s, \mathbf{x}_c) = \mathrm{e}^{-\|\mathbf{x}_s - \mathbf{x}_c\|^2/(2\sigma_{\mathcal{A}}^2)}/(2\pi\sigma_{\mathcal{A}}^2)^{3/2}, \quad (16)$$

with $\sigma_{\mathcal{A}}$ the root mean square (RMS) width of the Gaussian function. The action range of the spatial window is thus controlled with this parameter. Then (14) becomes,

$$\{\mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}\}_{ij} = \frac{1}{4(2\pi)^5\sigma_{\mathcal{A}}^3} \times \int_{\mathbb{R}^3} \frac{\mathrm{e}^{j(\omega/c)(\|\mathbf{x}_s - \mathbf{x}_i\| - \|\mathbf{x}_s - \mathbf{x}_j\|)}\mathrm{e}^{-\|\mathbf{x}_s - \mathbf{x}_c\|^2/\sigma_{\mathcal{A}}^2}}{\|\mathbf{x}_s - \mathbf{x}_i\|\|\mathbf{x}_s - \mathbf{x}_j\|} d\mathbf{x}_s. \quad (17)$$

With similar expressions for $g_{\mathcal{B}}(\mathbf{x}_s, \mathbf{x}_c)$ and $\mathbf{R}_{\mathbf{h}_{\mathcal{B}}\mathbf{h}_{\mathcal{B}}}(\omega)$, with $\sigma_{\mathcal{B}}$ the RMS width of the Gaussian function for the undesired zone. We solve the integrals involved in the calculus of these matrices (14) using a fast Gauss-Hermite numerical quadrature. We model uncorrelated noise with $\mathbf{n}$. It is then natural to make $\mathbf{R}_{\mathbf{nn}} = \sigma_{\mathbf{n}}^2\mathbf{I}$. In practical

applications σ_{n}^2 can be set to an estimate of the variance of quantization and/or internal instrument noise. In this way $\mathbf{R_{nn}}$ represents a parameter to control the white noise gain of the beamformer.

To improve the performance of the beam-former in reverberant environments a far-field isotropic sound field model is included as part of the acoustic interference. Spherically isotropic far-field is modeled as a sound field in the form of *plane-waves arriving from all possible directions* [18, 19]. The tail of a room impulse response (RIR) characterizes most of the subjective reverberant effect in a room, and can be modeled as a dense, homogeneous set of incoming reflections from all possible directions [20, Sec. 4.2]. The inclusion of an isotropic far-field as part of the acoustic interference can then be used to reduce the reverberation effect. The spatial covariance matrix of the microphone signals in presence of a spherically isotropic far-field is a known result [1, p. 66],

$$[\mathbf{R}_{ss}]_{i,j} = N_{ss}(\omega)\mathrm{sinc}\left(\frac{\omega}{c}\|\mathbf{x}_i - \mathbf{x}_j\|\right), \tag{18}$$

where $N_{ss}(\omega)$ is the power spectral density of the interference. In practice this quantity can be estimated from the signal-to-reverberant ratio (SRR) at the microphones [21, Chap. 2], [22].

In summary, the matrices involved in the generalized eigenvalue problem (11) are given by

$$\mathbf{A}(\omega) = \mathbf{R}_{\mathbf{h}_{\mathcal{A}}\mathbf{h}_{\mathcal{A}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega), \tag{19}$$

$$\mathbf{B}(\omega) = \mathbf{R}_{\mathbf{h}_{\mathcal{B}}\mathbf{h}_{\mathcal{B}}}(\omega) + \mathbf{R}_{\mathbf{nn}}(\omega) + \mathbf{R}_{ss}(\omega), \tag{20}$$

and (11) is restated as

$$\mathbf{B}(\omega)\mathbf{w}(\omega) = \lambda(\omega)\mathbf{A}(\omega)\mathbf{w}(\omega). \tag{21}$$

This problem can be solved, e.g., by a generalized Schur decomposition. By selecting only the L smallest eigenvalues based on a threshold, an ordered sequence of orthonormal beamformers (with respect to a weighted inner product) is found that forms a subspace in $\mathbb{C}^M$. By construction, the beamformer weights in this subspace have a response that focuses on the desired spatial zone and attenuate signals in the undesired zone (including reverberation in the form of an isotropic far-field). Further refinement of the beamformer can be made within this subspace to identify or select specific sources located in the desired zone or further processing on the response signal can be done to improve the quality of the acquired acoustic signal. In this work we select the eigenvector with the smallest eigenvalue as optimal beamformer weight vector.

3. EXPERIMENTAL RESULTS

In this section we evaluate the robustness of the proposed beamformer algorithm. To this extent we analyze its performance in simulated scenarios with and without reverberation, and in their real-life counterparts: in an anechoic chamber and in a normal office room.

We compare the proposed beamformer against the minimum-variance distortionless response (MVDR) beamformer. The signals obtained by the closest microphone to the target source are also used as benchmark reference. A comparison against other robust methods found in the literature is left for a future work as to make it fair and insightful we believe that an extensive comparison in a wide range of tests should be provided, placing it outside the scope of this paper.

The near-field MVDR beamformer is obtained by minimizing the power of the observed signal vector $\mathbf{z}(\omega) = \mathbf{v}(\mathbf{x}_s, \omega)S(\mathbf{x}_s, \omega)$, subject to the constraint that the signal in the desired look-position is kept undistorted [6]. In our notation $\mathbf{v}(\mathbf{x}_s, \omega)$ (as defined in (2)) represents the look-position vector. The optimal MVDR weights are given by [5],

Fig. 1. Real scenarios where the experiments were conducted. To the left the anechoic chamber of Delft University of Technology. To the right the office room.

$$\mathbf{w}_{\mathrm{mvdr}}(\omega) = \frac{\mathbf{R}_{\mathbf{zz}}^{-1}(\omega)\mathbf{v}(\mathbf{x}_s, \omega)}{\mathbf{v}^{\mathsf{H}}(\mathbf{x}_s, \omega)\mathbf{R}_{\mathbf{zz}}^{-1}(\omega)\mathbf{v}(\mathbf{x}_s, \omega)}, \tag{22}$$

where $\mathbf{R}_{\mathbf{zz}}(\omega)$ is the covariance matrix of the observation vector. MVDR is known to be a non-robust algorithm. Its robustness can, however, be improved if a *regularization* is applied [1, chap. 2]. This consists of adding a weighted identity matrix to the covariance matrix of the observation i.e., using $(\mathbf{R}_{\mathbf{zz}}(\omega) + \epsilon\mathbf{I})$ in the solution (22) instead of $\mathbf{R}_{\mathbf{zz}}(\omega)$, where ϵ is the regularization factor. This regularization trades efficacy for robustness. Unfortunately there is no simple approach to determine the optimal value of ϵ for a given scenario [1, chap. 2]. To keep computational complexity low instead of try and optimize it, in practical applications this factor can be set to a predefined value known to perform well in most cases.

Next we present results in four different scenarios, these are:

1. A simulated scenario with added white Gaussian noise (AWGN) at SNR = 60 dB and no reverberation.
2. A scenario with the same level of AWGN and reverberation simulated with the mirror image source method (MISM) [23], using the implementation by Habets [24]. The reverberation time is set to $T_{60} = 0.1$ s.
3. Measurements performed in the anechoic chamber of Delft University of Technology.
4. Measurements performed in an office room with dimensions 6.85 m by 3.95 m by 3.2 m in the x, y, z directions respectively, and an estimated reverberation time of $T_{60} \approx 0.1$ s.

We use the same scenario parameters and the same sound excerpts in all our simulated and real experiments. We used eight AKG C417 omnidirectional Lavalier condenser microphones, four single-cone loudspeakers with professional-grade drivers and custom-build enclosures. An RME Fireface 800 audio interface was used for recording and playback. Pictures of the real scenarios are given in Fig. 3. From the four acoustic sources the target source signal is selected to be a female speech excerpt. Two interferers are selected to be male speech signals and the third to be music. The speech signals were taken from the TIMIT database [25]. The music excerpt is a fragment of a rock song containing a male voice. The duration of all excerpts is truncated to 7 s. The sampling frequency is set to 16 kHz. The speed of sound is set to $c = 342$ m/s.

For the proposed beamformer we set the desired target region using a Gaussian window (16) with RMS width $\sigma_{\mathcal{A}} = 0.16$. The windows of the three interferers are set to the same RMS width of $\sigma_{\mathcal{B}} = 0.06$. In this way the three-sigma zone that accounts for 99.7%

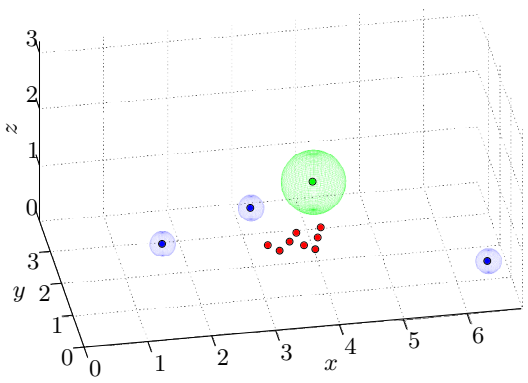

Fig. 2. Configuration used in the experiments. The red dots represent the microphone positions. The green sphere is the three-sigma zone of the Gaussian window enclosing the target source. Equivalently, the blue spheres are the zones of the interferer positions.

of the window volume is within 50 cm and 20 cm for the target and interferers respectively. This configuration is depicted in Fig. 3, where a green sphere denotes the three-sigma zone of the Gaussian window enclosing the target source. Equivalently, the blue spheres are the zones of the interferer positions. The microphone positions are denoted by red dots. Note that the target source is neither the closest nor the farthest source to the microphone array. This situation is commonly found in teleconference scenarios, voice-commanded applications like smart TVs, or hands-free systems in cars.

	Novel Beamformer		MVDR $\epsilon = 10^{-3}$		MVDR $\epsilon = 10^{-1}$		Closest mic.
	NE	E	NE	E	NE	E	
sSNR	-1.8	-7.9	7.7	-8.5	4.5	-0.4	-9.8
gSNR	-0.3	-3.8	12.2	-4.3	9.4	4.1	-4.8
PESQ	2.91	1.65	2.49	1.38	2.25	1.94	1.86
STOI	0.96	0.73	0.94	0.75	0.87	0.79	0.50

Table 1. Simulated scenario. No reverberation.

	Novel Beamformer		MVDR $\epsilon = 10^{-3}$		MVDR $\epsilon = 10^{-1}$		Closest mic.
	NE	E	NE	E	NE	E	
sSNR	-7.1	-7.6	-10	-14	-2.5	-4.8	-13
gSNR	-2.5	-2.9	-7.2	-10	1.9	-0.3	-8.7
PESQ	1.72	1.57	1.43	1.13	1.90	1.71	1.88
STOI	0.66	0.60	0.65	0.57	0.79	0.73	0.45

Table 2. Simulated scenario. Reverberation $T_{60} = 0.1$s.

To evaluate the performance we use four standard metrics. Speech intelligibility is assessed by the short-time objective intelligibility measure (STOI) [26], perceptual speech quality is evaluated using PESQ [27], and a raw signal comparison is given by segmental and global SNR. We applied these metrics to the proposed beamformer, the MVDR, and the signal of the closest microphone to the target source. All metrics are calculated with respect to the clean target excerpt. We performed a set of tests without introducing errors in the microphone and sources positions, and a set of test where random errors in all positions were fed into the newly proposed beamformer and the MVDR. The standard deviation of the error is set to 5 cm. A total of 100 repetitions of the experiments with position errors were performed and the average values are reported.

	Novel Beamformer		MVDR $\epsilon = 10^{-3}$		MVDR $\epsilon = 10^{-1}$		Closest mic.
	NE	E	NE	E	NE	E	
sSNR	-7.6	-11	-18	-20	-5.9	-8.5	-11
gSNR	-4.6	-8.8	-13	-15	-2.1	-4.6	-8.0
PESQ	2.15	1.63	2.48	2.89	2.03	1.74	1.35
STOI	0.81	0.67	0.55	0.48	0.85	0.77	0.55

Table 3. Anechoic chamber

	Novel Beamformer		MVDR $\epsilon = 10^{-3}$		MVDR $\epsilon = 10^{-1}$		Closest mic.
	NE	E	NE	E	NE	E	
sSNR	-5.5	-7.4	-32	-33	-14	-16	-13
gSNR	-1.8	-3.7	-27	-28	-10	-12	-9.4
PESQ	1.85	1.62	1.37	1.06	1.55	1.27	1.36
STOI	0.71	0.63	0.46	0.40	0.64	0.57	0.51

Table 4. Office room

For the MVDR we conducted tests with the regularization factor set to $\epsilon = 10^{-3}$; a value that was found to give the best performance under ideal conditions (i.e., the simulated scenario with no reverberation and no position errors). Another set of tests setting $\epsilon = 10^{-1}$ was performed. This value was found to give the best performance in the most challenging scenario (i.e., the office room).

Tables 1, 2, 3, 4 list the results. The nomenclature NE is used to indicate experiments with no induced position errors, E indicate experiments with induced position errors, sSNR indicates segmental SNR in dB, gSNR indicates global SNR in dB, PESQ is a value (-0.5–4.5), and STOI is a value (0–1).

We can draw the following conclusions from these results. First we confirm that the robustness of the MVDR can be significantly improved if the right regularization factor is used. As mentioned before, to find the optimal value is not a simple task that can be computationally demanding [1, chap. 2]. The MVDR with the best regularization factor performs satisfactorily and even marginally better than our proposed algorithm in all but the office room scenario. This shed light on two important conclusions. First, the performance of an algorithm should be tested on real scenarios, as simulations or even tests in controlled environments like an anechoic chamber can lead to non-realistic conclusions. Second and most importantly we confirm the robustness of our newly proposed beamformer in real-life scenarios.

4. CONCLUSIONS

In this paper we proposed a robust beamformer for acoustic applications. The algorithm uses a novel region-based near-field design combined with a far-field reverberation model. Our results showed that the evaluation of robustness using simulated scenarios can lead to non-realistic conclusions. Most importantly we showed that our proposed algorithm performs robustly in real-life applications.

Acknowledgments

We would like to thank Henry den Bok of Delft University of Technology for helping us setting the experiments in the anechoic chamber.

5. REFERENCES

[1] M. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, June 2001.

[2] J. G. Ryan and R. A. Goubran, "Near-field beamforming for microphone arrays," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Apr 1997, vol. 1, pp. 363–366.

[3] T. D. Abhayapala, R. A. Kennedy, and R. C. Williamson, "Nearfield broadband array design using a radially invariant modal expansion," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 392–403, 2000.

[4] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365 – 1376, Oct. 1987.

[5] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4 –24, Apr. 1988.

[6] H. L. Van Trees, *Optimum Array Processing*, vol. IV of *Detection, Estimation and Modulation Theory*, John Wiley & Sons, Inc., New York, 2002.

[7] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 478–488, 2004.

[8] W. Ser, H. Chen, and Z. L. Yu, "Self-calibration-based robust near-field adaptive beamforming for microphone arrays," *IEEE Trans. Circuits Syst. II*, vol. 54, no. 3, pp. 267–271, Mar. 2007.

[9] H. Chen, W. Ser, and Z. L. Yu, "Optimal design of nearfield wideband beamformers robust against errors in microphone array characteristics," *IEEE Trans. Circuits Syst. I*, vol. 54, no. 9, pp. 1950–1959, 2007.

[10] M. R. Islam, L. C. Godara, and M. S. Hossain, "Robust near field broadband beamforming in the presence of steering vector mismatches," in *Proc. IEEE Wireless and Microw. Technol. Conf. (WAMICON)*, Apr. 2012, pp. 1–6.

[11] D. B. Ward and G. W. Elko, "Mixed nearfield/farfield beamforming: a new technique for speech acquisition in a reverberant environment," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 1997, p. 4 pp.

[12] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, no. 12, pp. 2641 – 2673, 2003.

[13] D. Korompis, K. Yao, and F. Lorenzelli, "Broadband maximum energy array with user imposed spatial and frequency constraints," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1994, vol. iv, pp. IV/529–IV/532 vol.4.

[14] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*, McGraw-Hill, New York, 1968.

[15] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, Academic Press, London, UK, 1999.

[16] J. Ahrens, *Analytic Methods of Sound Field Synthesis*, T-Labs Series in Telecommunication Services. Springer, Berlin, Jan. 2012.

[17] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc. F (Commun., Radar, Signal Process.)*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[18] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[19] T. D. Abhayapala, R. A. Kennedy, and R. C. Williamson, "Isotropic noise modelling for nearfield array processing," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 1999, pp. 11–14.

[20] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, UK, Oct. 2000.

[21] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Signals and Communication Technology. Springer, London, UK, 2010.

[22] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 309–312.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[24] E. A. P. Habets, "Room impulse response generator for MATLAB," RIR_Generator_v2.0_20100920.zip, Sept. 2010.

[25] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Tech. Rep., National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2010, pp. 4214–4217.

[27] 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*, ITU-T, Feb. 2001.