Steered Response Power-Based Direction-of-Arrival Estimation Exploiting an Auxiliary Microphone

Klaus Brümann and Simon Doclo

Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all,
Carl von Ossietzky Universität, Oldenburg, Germany
{klaus.bruemann,simon.doclo}@uni-oldenburg.de

Abstract—Accurately estimating the direction-of-arrival (DOA) of a speech source using a compact microphone array (CMA) is often complicated by background noise and reverberation. A commonly used DOA estimation method is the steered response power with phase transform (SRP-PHAT) function, which has been shown to work reliably in moderate levels of noise and reverberation. Since for closely spaced microphones the spatial coherence of noise and reverberation may be high over an extended frequency range, this may negatively affect the SRP-PHAT spectra, resulting in DOA estimation errors. Assuming the availability of an auxiliary microphone at an unknown position which is spatially separated from the CMA, in this paper we propose to compute the SRP-PHAT spectra between the microphones of the CMA based on the SRP-PHAT spectra between the auxiliary microphone and the microphones of the CMA. For different levels of noise and reverberation, we show how far the auxiliary microphone needs to be spatially separated from the CMA for the auxiliary microphone-based SRP-PHAT spectra to be more reliable than the SRP-PHAT spectra without the auxiliary microphone. These findings are validated based on simulated microphone signals for several auxiliary microphone positions and two different noise and reverberation conditions.

Index Terms—Steered-Response Power, Direction-of-arrival, Auxiliary Microphone, Source Localization

I. INTRODUCTION

In many speech processing applications, compact microphone arrays (CMAs) are used to estimate the direction-of-arrival (DOA) of a speech source. A popular DOA estimation method is based on the steered response power (SRP) function [1–5], where the DOA is estimated as the direction maximizing the SRP spectra over frequencies and microphone pairs. Despite its simplicity, the SRP function is rather sensitive to reverberation and spatially coherent noise and has a poor spatial resolution [6].

To improve robustness against noise and reverberation, several frequency weightings have been proposed [7–10]. The most common frequency weighting is the phase transform (PHAT) [11], which removes the influence of the amplitude in the SRP spectra, effectively decorrelating the microphone signals and enhancing the spatial resolution. While the PHAT weighting is optimal in anechoic environments with spatially uncorrelated noise, it has also been demonstrated to work well in realistic noisy and reverberant acoustic environments where

these assumptions are not satisfied [12–14]. Nevertheless, for small inter-microphone distances noise and reverberation typically exhibit a larger spatial coherence than for large inter-microphone distances, especially at low frequencies, negatively affecting the SRP-PHAT spectra and the DOA estimation performance.

In this paper, we assume that an additional microphone is available, which is spatially separated from the CMA. The main objective is to exploit this additional microphone to enhance the estimation accuracy of the SRP spectra in noisy and reverberant environments. In recent years, several source localization methods exploiting an auxiliary microphone have been proposed. While some methods assume that the auxiliary microphone is in the vicinity of the source [15–17], other methods just assume that the auxiliary microphone is spatially separated from the microphone array [18–20]. In both cases the spatial coherence of the noise and reverberation between the microphones of the array and the auxiliary microphone is typically lower than the spatial coherence between the microphones of the array. Similarly to [20], in this paper we propose to compute the SRP-PHAT spectra between the microphones of the CMA based on the SRP-PHAT spectra between the auxiliary microphone and the microphones of the CMA. Assuming a spherically isotropic sound field and a constant power spectrum for the noise and reverberation, we theoretically analyse the influence of the noise and reverberation on the auxiliary microphone-based SRP-PHAT spectrum. We show that the auxiliary microphone-based SRP-PHAT spectrum is more reliable than the conventional SRP-PHAT spectrum without the auxiliary microphone when the distance between the auxiliary microphone and the CMA is larger (depending on the noise and reverberation level).

Based on simulated microphone signals for a speech source in a reverberant room with multi-talker babble noise, we validate these findings for several positions of the auxiliary microphone. Simulation results show that for most positions of the auxiliary microphone a notable improvement in terms of DOA estimation accuracy is achieved, even in challenging acoustical conditions.

II. SRP-BASED DOA ESTIMATION EXPLOITING AN AUXILIARY MICROPHONE

We consider an acoustic scenario with a single speech source in a noisy and reverberant environment, recorded with
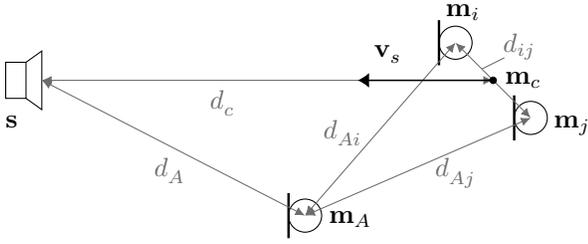
Fig. 1: Exemplary arrangement of a source at position $\mathbf{s}$, closely spaced microphones at positions $\mathbf{m}_i$ and $\mathbf{m}_j$, and an auxiliary microphone at position $\mathbf{m}_A$.

a CMA consisting of M microphones. The position of the source is denoted by $\mathbf{s}$ and the position of the j-th microphone is denoted by $\mathbf{m}_j$ (see Fig. 1).

In the continuous-time Fourier transform domain, the reverberant and noisy j-th microphone signal $Y_j(\omega)$ can be decomposed as

$$Y_j(\omega) = X_j^D(\omega) + \underbrace{X_j^R(\omega) + N_j(\omega)}_{U_j(\omega)}, \quad (1)$$

where ω denotes the radial frequency, $X_j^D(\omega)$ denotes the direct path component and $U_j(\omega)$ denotes the undesired component, comprising of the reverberation component $X_j^R(\omega)$ and the noise component $N_j(\omega)$. Assuming a free field acoustic scenario, the direct path component is equal to the source signal $S(\omega)$ with a phase shift and attenuation, i.e.,

$$X_j^D(\omega) = \frac{1}{Ad_j} S(\omega) \exp(-\jmath\omega d_j/\nu), \quad (2)$$

with $\jmath = \sqrt{-1}$, $A = 4\pi$ the attenuation factor of a point source, $d_j = ||\mathbf{s} - \mathbf{m}_j||$ the distance between the source and the j-th microphone, and ν the speed of sound. Assuming that the source is in the far field of the CMA, the direct path components in the i-th and j-th microphones are related as

$$X_j^D(\omega) = X_i^D(\omega) \exp(\jmath\omega\tau_{ij}(\mathbf{v}_s)), \quad (3)$$

where $\tau_{ij}(\mathbf{v}_s) = -\mathbf{v}_s^{\mathrm{T}}(\mathbf{m}_i - \mathbf{m}_j)/\nu$ denotes the TDOA between the i-th and j-th microphones. The time-difference of arrival (TDOA) is defined relative to the source DOA vector $\mathbf{v_s} = [\cos(\theta_s), \sin(\theta_s)]^{\mathrm{T}}$, where θ_s denotes the azimuthal angle of the source, relative to the centroid of the CMA. The SRP function [1–5] is defined as

$$\varphi(\mathbf{v}) = \sum_{(i,j)\,:\,i>j} \int_{-\infty}^{\infty} \psi_{ij}(\omega) \exp(\jmath\omega\tau_{ij}(\mathbf{v}))d\omega, \quad (4)$$

with DOA vector $\mathbf{v} = [\cos(\theta), \sin(\theta)]^{\mathrm{T}}$. The weighted SRP spectrum between the i-th and j-th microphones is defined as

$$\psi_{ij}(\omega) = W_{ij}(\omega) \underbrace{\mathbb{E}\{Y_i(\omega)Y_j^*(\omega)\}}_{\tilde{\psi}_{ij}(\omega)}, \quad (5)$$

with $\mathbb{E}\{\cdot\}$ the expected value operator, $\{\cdot\}^*$ the complex conjugate operator, $\tilde{\psi}_{ij}(\omega)$ the (unweighted) SRP spectrum and $W_{ij}(\omega)$ a weighting determining the contribution of the SRP spectrum in each frequency to the integral in (4).

The source DOA vector is then estimated as the DOA vector maximizing the SRP function in (4), i.e.,

$$\hat{\mathbf{v}}_s = \underset{\mathbf{v}}{\operatorname{argmax}} \; \varphi(\mathbf{v}). \quad (6)$$

In this paper, we consider the commonly used PHAT weighting $W_{ij}^{\mathrm{PHAT}}(\omega) = 1/|\tilde{\psi}_{ij}(\omega)|$ in (5), yielding the SRP-PHAT spectrum $\psi_{ij}^{\mathrm{PHAT}}(\omega)$. Assuming the direct path component $X_j^D(\omega)$ and the undesired component $U_j(\omega)$ to be uncorrelated in each microphone, the SRP-PHAT spectrum is given by

$$\psi_{ij}^{\mathrm{PHAT}}(\omega) = \frac{\tilde{\psi}_{ij}^D(\omega) + \tilde{\psi}_{ij}^U(\omega)}{|\tilde{\psi}_{ij}^D(\omega) + \tilde{\psi}_{ij}^U(\omega)|}, \quad (7)$$

with the SRP spectrum of the direct path component $\tilde{\psi}_{ij}^D(\omega) = \mathbb{E}\{X_i^D(\omega)X_j^{D,*}(\omega)\}$ and the SRP spectrum of the undesired component $\tilde{\psi}_{ij}^U(\omega) = \mathbb{E}\{U_i(\omega)U_j^*(\omega)\}$. Assuming the undesired component to be spatially uncorrelated, i.e., $\tilde{\psi}_{ij}^U(\omega) = 0$ for $i \neq j$, and using (3), the SRP-PHAT spectrum in (5) is given by $\psi_{ij}^{\mathrm{PHAT}}(\omega) = \exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s))$, for which it can be easily shown that $\hat{\mathbf{v}}_s = \mathbf{v}_s$.

Even though in practice both assumptions (i.e., uncorrelated direct path and undesired components, spatially uncorrelated undesired component) are typically not satisfied, it has been shown that the SRP-PHAT method works rather well in real noisy and reverberant environments [12–14]. However, in the presence of spatially coherent noise and reverberation, the performance may suffer, especially when using CMAs, for which the spatial coherence between the microphones is typically larger than for arrays with large inter-microphone distances.

Instead of considering alternative weightings that are more robust to noise and reverberation [7–10], in this paper we propose to compute the SRP-PHAT spectrum exploiting the availability of an additional microphone at an unknown position $\mathbf{m}_A$ (see Fig. 1), which is spatially separated from the CMA (analysis in the following section). Similarly to [20], but now in the frequency domain instead of the time domain, we propose to compute the auxiliary microphone-based SRP-PHAT spectrum $\psi_{ij}^{\mathrm{PHAT\text{-}A}}(\omega)$ between the i-th and j-th microphones by multiplying the SRP-PHAT spectrum $\psi_{iA}^{\mathrm{PHAT}}(\omega)$ between the i-th microphone and the auxiliary microphone with the SRP-PHAT spectrum $\psi_{Aj}^{\mathrm{PHAT}}(\omega)$ between the auxiliary microphone and the j-th microphone, i.e.,

$$\boxed{\psi_{ij}^{\mathrm{PHAT\text{-}A}}(\omega) = \psi_{iA}^{\mathrm{PHAT}}(\omega)\psi_{Aj}^{\mathrm{PHAT}}(\omega)} \quad (8)$$

Assuming the undesired component to be spatially uncorrelated, it can easily be shown that the auxiliary microphone-based SRP-PHAT spectrum is equivalent to the SRP-PHAT spectrum without the auxiliary microphone, i.e., $\psi_{ij}^{\mathrm{PHAT\text{-}A}}(\omega) = \exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s))$. However, for spatially coherent noise and reverberation, the SRP-PHAT spectra $\psi_{iA}^{\mathrm{PHAT}}(\omega)$ and $\psi_{Aj}^{\mathrm{PHAT}}(\omega)$ between the auxiliary microphone and the microphones of the CMA are assumed to be less affected by noise and reverberation than the SRP-PHAT spectrum $\psi_{ij}^{\mathrm{PHAT}}(\omega)$ due to the spatial separation of the auxiliary microphone from the CMA. This will be analysed and validated in more detail in the following sections.

III. ANALYSIS OF AUXILIARY MICROPHONE-BASED SRP-PHAT SPECTRUM

The main objective of this section is to theoretically analyse the influence of the undesired component on the auxiliary microphone-based SRP-PHAT spectrum in (8) compared to the conventional SRP-PHAT spectrum in (7). To this end, we define a distortion measure and analyse its contribution across frequencies for different positions of the auxiliary microphone and different power ratios between the source and the undesired component.

Since the source is assumed to be in the far field of the CMA, we assume that $d_i = d_j = d_c$, with d_c the distance between the source and the centroid of the CMA (see Fig. 1). In addition, for the analysis in this section we assume a spherically isotropic sound field [21] for the undesired component (noise and reverberation). In the next section, the findings will be validated with realistic microphone signals. Using the aforementioned assumptions, the SRP-PHAT spectrum in (7) can be written as

$$\psi_{ij}^{\text{PHAT}}(\omega) = \frac{\frac{\phi_S(\omega)}{(Ad_c)^2} \exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s)) + \phi_U(\omega)\operatorname{sinc}(\frac{\omega d_{ij}}{\nu})}{|\frac{\phi_S(\omega)}{(Ad_c)^2} \exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s)) + \phi_U(\omega)\operatorname{sinc}(\frac{\omega d_{ij}}{\nu})|},$$
(9)

with $\phi_S(\omega) = \mathbb{E}\{|S(\omega)|^2\}$ and $\phi_U(\omega) = \mathbb{E}\{|U(\omega)|^2\}$ the power spectral densities of the source and undesired component, respectively. By defining the source-to-undesired ratio (SUR) at the source position as $\text{SUR}(\omega) = \phi_S(\omega)/\phi_U(\omega)$, the expression in (9) can be written as

$$\psi_{ij}^{\text{PHAT}}(\omega) = \frac{\exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s)) + \overbrace{\frac{(Ad_c)^2}{\text{SUR}(\omega)}\operatorname{sinc}\left(\frac{\omega d_{ij}}{\nu}\right)}^{D_{ij}(\omega)}}{|\exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s)) + \frac{(Ad_c)^2}{\text{SUR}(\omega)}\operatorname{sinc}\left(\frac{\omega d_{ij}}{\nu}\right)|},$$
(10)

with $D_{ij}(\omega)$ the distortion to the optimal phase shift $\exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s))$. This distortion depends on the SUR, the distance d_c between the source and the centroid of the CMA, and the distance d_{ij} between the i-th and j-th microphones.

Similarly to (10), we can also rewrite the auxiliary microphone-based SRP-PHAT spectrum in (8) as

$$\psi_{ij}^{\text{PHAT-A}}(\omega) = \frac{\exp(-\jmath\omega\tau_{iA}(\mathbf{v}_s)) + \frac{A^2 d_c d_A}{\text{SUR}(\omega)}\operatorname{sinc}(\frac{\omega d_{Ai}}{\nu})}{|\exp(-\jmath\omega\tau_{iA}(\mathbf{v}_s)) + \frac{A^2 d_c d_A}{\text{SUR}(\omega)}\operatorname{sinc}(\frac{\omega d_{Ai}}{\nu})|} \cdot \ldots$$
$$\frac{\exp(-\jmath\omega\tau_{Aj}(\mathbf{v}_s)) + \frac{A^2 d_c d_A}{\text{SUR}(\omega)}\operatorname{sinc}(\frac{\omega d_{Aj}}{\nu})}{|\exp(-\jmath\omega\tau_{Aj}(\mathbf{v}_s)) + \frac{A^2 d_c d_A}{\text{SUR}(\omega)}\operatorname{sinc}(\frac{\omega d_{Aj}}{\nu})|},$$
(11a)

$$= \frac{\exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s)) + D_{ij}^{\text{A}}(\omega)}{|\exp(-\jmath\omega\tau_{ij}(\mathbf{v}_s)) + D_{ij}^{\text{A}}(\omega)|},$$
(11b)

with d_A the distance between the source and the auxiliary microphone position and $D_{ij}^{\text{A}}(\omega)$ the auxiliary microphone-

based distortion, defined in (12). It is not straightforward to analytically show under which conditions the absolute value of the auxiliary microphone-based distortion $|D_{ij}^{\text{A}}(\omega)|$ in (12) is smaller than the absolute value of the distortion $|D_{ij}(\omega)|$ in (10), leading to a more reliable SRP-PHAT spectrum. However, due to the overall decaying nature of the sinc-function and the fact that for most microphone constellations the distances d_{Ai} and d_{Aj} between the auxiliary microphone and the CMA are larger than the inter-microphone distances d_{ij}, the chances are high that the condition $|D_{ij}^{\text{A}}(\omega)| < |D_{ij}(\omega)|$ is satisfied. To investigate the influence of the constellations, SURs and frequencies on the distortions, we analyse the proportion of frequencies up to ω_0 for which this condition is satisfied, i.e.,

$$P = \frac{1}{\omega_0}\int_0^{\omega_0} \mathcal{H}\left(|D_{ij}(\omega)| - |D_{ij}^{\text{A}}(\omega)|\right) d\omega,$$
(13)

with $\mathcal{H}(x)$ the Heaviside step function, which returns the value $\mathcal{H}(x) = 0$ if $x \leq 0$ and $\mathcal{H}(x) = 1$ if $x > 0$. Since the PHAT weighting normalizes each frequency so that it has an equal contribution in (4), it is assumed that when $P > 0.5$, the auxiliary microphone-based SRP-PHAT function is more suitable for source localization than the SRP-PHAT function without the auxiliary microphone. To analyse the influence of the auxiliary microphone position on the proportion in (13), we consider an exemplary constellation, consisting of a CMA with $M = 2$ microphones and inter-microphone distance $d_{12} = 5$ cm, and a source $d_c = 2$ m from the centroid of the CMA (see Fig. 2).

To average out the influence of the microphone pair orientation, we average the proportion in (13) over 18 different orientations in the x-y plane from $0°$ to $170°$. For simplicity, we assume $\text{SUR}(\omega)$ to be constant for all frequencies, i.e., $\text{SUR}(\omega) = \text{SUR}$, considering 1025 frequencies between 0 and 8 kHz. For different SURs, Fig. 2 shows the influence of the auxiliary microphone position on the averaged proportion of frequencies P_{avg}. The red and blue boundaries at values P_{avg} equal to 0.5 or 0.9 indicate the auxiliary microphone positions for which in 50% or 90% of the frequencies, respectively, the distortion of the auxiliary microphone-based SRP-PHAT spectrum is smaller than the distortion of the SRP-PHAT spectrum without the auxiliary microphone. It can clearly be observed that the averaged proportion P_{avg} depends both on the SUR as well as on the auxiliary microphone position. For the auxiliary microphone to improve the reliability of the SRP-PHAT function, it should be located far enough from the CMA. As the SUR decreases, the size of the bounded areas notably increases. This means that for increasing levels of noise and reverberation, the auxiliary microphone needs to be located farther from the CMA to yield a benefit. It should however be noted that for SUR ≥ 10 dB the auxiliary microphone only needs to be located 20 cm from the centroid of the CMA,

$$D_{ij}^{\text{A}}(\omega) = \frac{A^2 d_c d_A}{\text{SUR}(\omega)}\left[\operatorname{sinc}(\omega d_{Ai}/\nu)\exp(-\jmath\omega\tau_{Aj}(\mathbf{v}_s)) + \operatorname{sinc}(\omega d_{Aj}/\nu)\exp(-\jmath\omega\tau_{iA}(\mathbf{v}_s))\right] + \frac{A^4 d_c^2 d_A^2}{\text{SUR}^2(\omega)}\operatorname{sinc}(\omega d_{Ai}/\nu)\operatorname{sinc}(\omega d_{Aj}/\nu)$$
(12)

whereas for SUR $\geq$ 0 dB this increases to 75 cm. Interestingly, incorporating an auxiliary microphone can be beneficial even when it is located farther from the source than the CMA, as long as it is spatially separated from the CMA. Similar results are also obtained for other values of d_{12} and d_c.

IV. EXPERIMENTAL VALIDATION

To experimentally validate the findings from the previous section, in this section the DOA estimation performance of the SRP-PHAT method without and with an auxiliary microphone is compared based on simulated microphone signals. The performance is compared for various positions of the auxiliary microphone and different levels of noise and reverberation.

A. Scenario and Algorithm Parameters

For the simulations, we considered a rectangular room with dimensions 6 m × 6 m × 2.4 m and simulated room impulse responses using the image source method [22], [23], assuming equal reflection coefficients for all walls. We considered a similar constellation as in Section III, now consisting of a CMA with $M = 3$ microphones (to estimate the DOA without ambiguity), centered at $\mathbf{m}_c = [4, 3, 1.75]$ and with inter-microphone distance $d_{ij} = 5$ cm, and a source position $s = [2, 3, 1.75]$ m, located $d_c = 2$ m from the centroid of the CMA. We considered 601 auxiliary microphone positions on a 2D grid in the x-y plane, with an increased spatial resolution close to the CMA (see Fig. 3). To average out the influence of the orientation of the CMA, 12 azimuthal orientations of the CMA in the x-y plane were considered from 0° to 110°. For each orientation, we considered 10 different speech signals of length 3 s, randomly selected from [24] (with equal probability for a male or female speaker), resulting in a total of 120 simulated scenarios for each auxiliary microphone position. To compare directly with two of the plots from Fig. 2 where the SUR was equal to 15 dB and 0 dB, we considered two noise and reverberation conditions with similar broadband SURs. In the first condition, the broadband direct-to-reverberant ratio (DRR) and the broadband reverberant signal-to-noise ratio (RSNR) were both equal to -1.4 dB across the microphones of the CMA, corresponding to a broadband SUR of 14.9 dB at the source position, while in the second condition, the RSNR and the DRR were both equal to -7.2 dB across the microphones of the CMA, corresponding to a broadband SUR of 0.0 dB at the source position. For each condition, the reflection coefficients were determined which achieve the desired average DRR across the microphones, leading to $T_{60} = 0.18$ s for the first condition and $T_{60} = 0.32$ s for the second condition.

The algorithms were implemented at a sampling frequency of 16 kHz and using a short-time Fourier transform with a frame length of 512 samples (corresponding to 32 ms), 50% overlap between frames, and using a square-root-Hann analysis window. The SRP-PHAT spectra in (5) and (8) were computed using recursive averaging, i.e.,

$$\tilde{\psi}_{i,j}[k,l] = \lambda \tilde{\psi}_{i,j}[k,l-1] + (1-\lambda)Y_i Y_j^*[k,l], \quad (14)$$

with frequency bin $k \in \{1, \ldots, K\}$, time frame $l \in \{1, \ldots, L\}$, and smoothing factor $\lambda = 0.98$ (corresponding to an averaging over approximately 0.8 s). The DOA estimate in (6) was obtained by first averaging the SRP-PHAT spectra over all time frames, i.e., $1/L \sum_{l=1}^{L} \psi_{i,j}^{\text{PHAT}}[k,l]$, and then using the averaged SRP-PHAT spectra in (4), where the integral was approximated by a sum over discrete frequency bins k.

B. Comparison of DOA Estimation Performance

To compare the DOA estimation performance of the SRP-PHAT method without and with an auxiliary microphone, we evaluated the DOA estimation error

$$\varepsilon = \cos^{-1}\left(\frac{\hat{\mathbf{v}}_s^T \mathbf{v}_s}{||\hat{\mathbf{v}}_s||_2 \cdot ||\mathbf{v}_s||_2}\right), \quad (15)$$

averaged over the 120 simulated scenarios.

For all considered auxiliary microphone positions, Fig. 3 illustrates the mean DOA estimation error, for both noise and reverberation conditions. The mean DOA estimation error for the SRP-PHAT method without an auxiliary microphone was equal to 3.8° in the first condition (RSNR = -1.4 dB, DRR = -1.4 dB), and 34.0° in the second condition (RSNR = -7.2 dB, DRR = -7.2 dB). It can clearly be seen that incorporating an auxiliary microphone in the SRP-PHAT method generally results in a reduced mean DOA estimation error, especially for larger distances between the auxiliary microphone and the CMA. This aligns well with our findings from Section III.

The red-bordered pixels in Fig. 3 indicate positions of the auxiliary microphone for which incorporating the auxiliary microphone did not reduce the mean DOA estimation error. First, it can be seen that in the vicinity of the CMA the spread of red-bordered pixels is larger for the more challenging noise and reverberation condition. This corresponds to the area bounded by the red line in Fig. 2, which increases as the SUR decreases. Second, it can be seen that certain auxiliary microphone positions in the vicinity of walls exhibit a slightly larger mean DOA estimation error when the auxiliary microphone is used. This can be attributed to the limitations of the assumed model in Section III, which assumed a spherically isotropic field for the reverberation without considering early reflections from the walls.

V. CONCLUSIONS

In this paper, we have proposed a method to improve SRP-PHAT-based DOA estimation in CMAs, where noise and reverberation with a high spatial coherence may negatively affect the reliability of the SRP-PHAT spectra. Assuming the availability of a spatially separated auxiliary microphone, we proposed to compute the SRP-PHAT spectra in a different way, using the SRP-PHAT spectra between the auxiliary microphone and the microphones of the CMA. For different SURs, we showed how far the auxiliary microphone needs to be spatially separated from the CMA for the SRP-PHAT spectra to be estimated more reliably with the auxiliary microphone than without. Furthermore, we validated these findings with simulation results based on simulated microphone signals for many different auxiliary microphone positions and two different noise and reverberation conditions, showing the beneficial impact of incorporating an auxiliary microphone for DOA estimation.
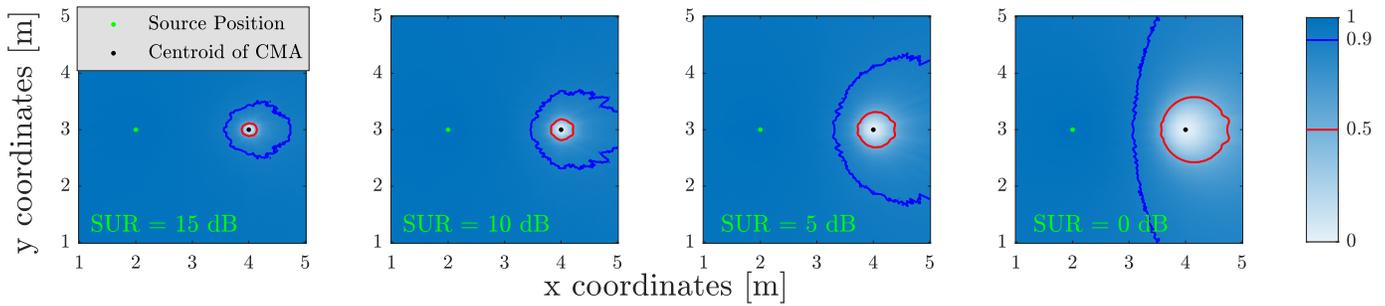
Fig. 2: Averaged proportion of frequencies for which the distortion of the auxiliary microphone-based SRP-PHAT spectrum is smaller than the distortion of the SRP-PHAT spectrum without the auxiliary microphone, where each point in the 2D plane corresponds to a different position of the auxiliary microphone.
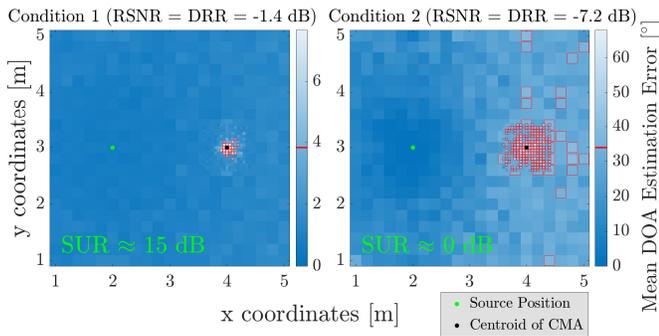

Fig. 3: Mean DOA estimation error of the auxiliary microphone-based SRP-PHAT function compared to the SRP-PHAT function without the auxiliary microphone, for different auxiliary microphone positions in a 2D plane and two different noise and reverberation conditions.

REFERENCES

[1] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Audio, Speech, Language Processing*, no. 3, pp. 288–292, 1997.

[2] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Ph.D. thesis, Brown University, Providence, RI, USA, 2000.

[3] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer, 2010, pp. 307–337.

[4] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, "Multichannel source activity detection, localization, and tracking," in *Audio source separation and speech enhancement*. Wiley, 2018, pp. 47–64.

[5] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on Nyquist-Shannon sampling," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2021, pp. 206–210.

[6] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, 2014.

[7] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[8] H. R. Abutalebi and H. Momenzadeh, "Performance improvement of TDOA-based speaker localization in joint noisy and reverberant conditions," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–13, 2011.

[9] G. Athanasopoulos, "Contributions to acoustic localization for robotic audition," Ph.D. dissertation, Vrije Universiteit Brussel, 2016.

[10] S. Braun, W. Zhou, and E. A. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015, pp. 1–5.

[11] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[12] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, pp. 1–19, 2006.

[13] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?" in *Proc. IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 2565–2568.

[14] J. Velasco, C. J. Martin-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *ELSEVIER Signal Processing*, vol. 119, pp. 209–228, 2016.

[15] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. on Audio, Speech, Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[16] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Trans. on Audio, Speech, Language Processing*, vol. 26, no. 7, pp. 1275–1289, 2018.

[17] U. Kowalk, S. Doclo, and J. Bitzer, "Signal-informed DNN-based DOA estimation combining an external microphone and GCC-PHAT features," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Bamberg, Germany, 2022, pp. 1–5.

[18] D. Fejgin and S. Doclo, "Comparison of binaural RTF-vector-based direction of arrival estimation methods exploiting an external microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, Online, 2021, pp. 241–245.

[19] D. Fejgin and S. Doclo, "Exploiting an external microphone for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. Forum Acusticum*, Turin, Italy, 2023.

[20] K. Brümann and S. Doclo, "Exploiting an external microphone to improve time-difference-of-arrival estimates for Euclidean distance matrix-based source localization," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2023, pp. 16–20.

[21] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, pp. 61–85.

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

[23] E. A. P. Habets, *RIR-Generator*, available at "https://github.com/ehabets/RIR-Generator".

[24] I. Solak, *M-AILABS Speech Dataset*, available at "https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/".