A Steered Response Power Method for Sound Source Localization With Generic Acoustic Models

Kaspar Müller, Markus Buck, *Member, IEEE*, Simon Doclo, *Senior Member, IEEE*, Jan Østergaard, *Senior Member, IEEE*, and Tobias Wolff

***Abstract*—The steered response power (SRP) method is one of the most popular approaches for acoustic source localization with microphone arrays. It is often based on simplifying acoustic assumptions, such as an omnidirectional sound source in the far field of the microphone array(s), free field propagation, and spatially uncorrelated noise. In reality, however, there are many acoustic scenarios where such assumptions are violated. This paper proposes a generalization of the conventional SRP method that allows to apply generic acoustic models for localization with arbitrary microphone constellations. These models may consider, for instance, level differences in distributed microphones, the directivity of sources and receivers, or acoustic shadowing effects. Moreover, also measured acoustic transfer functions may be applied as acoustic model. We show that the delay-and-sum beamforming of the conventional SRP is not optimal for localization with generic acoustic models. To this end, we propose a generalized SRP beamforming criterion that considers generic acoustic models and spatially correlated noise, and derive an optimal SRP beamformer. Furthermore, we propose and analyze appropriate frequency weightings. Unlike the conventional SRP, the proposed method can jointly exploit observed level and time differences between the microphone signals to infer the source location. Realistic simulations of three different microphone setups with speech under various noise conditions indicate that the proposed method can significantly reduce the mean localization error compared to the conventional SRP and, in particular, a reduction of more than 60% can be archived in noisy conditions.**

***Index Terms*—Beamforming, microphone arrays, distributed microphones, source localization, steered response power (SRP).**

I. Introduction

ACOUSTIC sound source localization is a frequently required task. Beyond source location estimation itself, it is fundamental for various applications, such as teleconferencing, adaptive beamforming, speaker separation, autonomous driving, or robotics. The steered response power (SRP) method is one of the most commonly used methods for source localization with microphone arrays. Its conceptual idea is to use a steered beamformer in order to scan the space for a sound source by observing the steering direction (or position) with maximum beamformer output power. In the 1990 s, Omologo and Svaizer observed that using the phase information of microphone cross power spectra is a useful strategy for time-difference-of-arrival-based source localization [1], [2]. This idea of exploiting microphone cross power spectral densities by means of a steered delay-and-sum beamformer further evolved to the SRP method in its current form as a source localization standard [3], [4]. In particular, the variant applying the phase transform (PHAT) [5] in order to use the cross power spectrum phase asserted itself as the popular SRP-PHAT that is known for its robustness against reverberation [3], [6], [7]. We refer to the standard SRP (including SRP-PHAT) as conventional SRP (CSRP). An extensive literature study reviewing its background and presenting various extensions has been published recently [8]. The CSRP is based on ideal, simplifying acoustic assumptions, namely an omnidirectional point source in the far field of the microphones, free-field propagation and spatially uncorrelated noise [4]. In practice, these assumptions are usually not met [9]. Nevertheless, SRP-PHAT usually performs well with microphone arrays in the far field of a source. Hence, the CSRP is typically used with microphone arrays whose aperture is much smaller than the distance to the sound source – either with a single array for direction-of-arrival (DOA) estimation, or with multiple distributed microphone arrays for source position estimation. In the latter case, the CSRP is often processed for each array individually and weighted and summed up afterwards, or triangulation is used to infer the source location [10], [11], [12]. In various other setups, however, the simple acoustic assumptions are violated to a greater extent. Typical examples are setups with distributed microphones in the near field of a source including the emerging field of wireless acoustic sensor networks, or setups involving directional microphones or sources. Another common application is binaural source localization, for instance with hearing aid devices, where the acoustic head shadow causes frequency-dependent interaural level and time differences that significantly deviate from free-field propagation [13], [14]. In such scenarios, not only phase differences between microphone power spectra can be used to determine the source location but also the power

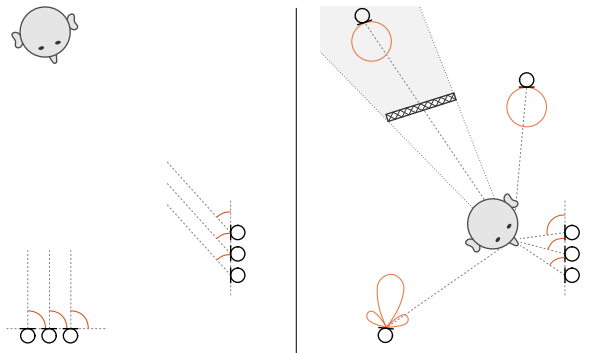
Received 24 March 2025; revised 19 August 2025; accepted 11 September 2025. Date of publication 17 September 2025; date of current version 29 September 2025. This work was supported by the SOUNDS European Training Network – an European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie under Agreement 956369. The associate editor coordinating the review of this article and approving it for publication was Dr. Prasanga Samarasinghe. *(Corresponding author: Kaspar Müller.)*

Kaspar Müller, Markus Buck, and Tobias Wolff are with the Audio AI R&D Department, Cerence AI, 89077 Ulm, Germany (e-mail: kaspar.mueller@cerence.com; markus.buck@cerence.com; tobias.wolff@cerence.com).

Simon Doclo is with the Department of Medical Physics and Acoustics, and the Cluster of Excellence Hearing4all, University of Oldenburg, 26129 Oldenburg, Germany (e-mail: simon.doclo@uni-oldenburg.de).

Jan Østergaard is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg East, Denmark (e-mail: jo@es.aau.dk).

Digital Object Identifier 10.1109/TASLPRO.2025.3611789



(a) Conventional SRP application: one or multiple distributed microphone arrays are in the far field of a sound source.

(b) Exemplary generic application: distributed microphones or arrays are in the near field of a source and might be affected by shadowing effects and different directivities of the source or microphones.

Fig. 1. Exemplary microphone setups for source localization. Unlike in (a), the microphone level differences in (b) contain relevant source position cues.

differences carry relevant source location cues. However, power differences are not exploited by the CSRP (see Fig. 1).

Various approaches have been proposed to improve the localization performance of SRP-PHAT, for instance, with setups involving distributed microphones or sensor networks [15], [16], [17], [18], [19], by alternative prewhitening [20], [21], or by exploiting an auxiliary microphone [22]. Other recent contributions focus on the problem of SRP localization of multiple sound sources [23], [24], [25]. Furthermore, several extension or alternative grid search methods have been proposed to reduce the computational complexity [26], [27], [28], [29], [30]. There are also approaches using measured head-related transfer functions or head models for binaural steered-beamforming localization [13], [31]. Alternative SRP approaches or extensions involving machine learning are proposed, for instance, in [32], [33], [34], [35], [36] to improve localization performance under realistic acoustic conditions. However, to the best of our knowledge there is no signal-processing-based work addressing generic acoustic conditions in general by involving more complex acoustic propagation models and noise characteristics for SRP-based localization with arbitrary microphone constellations.

In this paper, we propose a generalization of the conventional frequency-domain SRP method with regard to the just mentioned aspects: The presented method allows to apply advanced, setup-specific acoustic propagation models. We show that this enables a joint exploitation of phase and level information for localization. Moreover, arbitrary noise characteristics can be taken into account to improve robustness against spatially correlated or inhomogeneous noise. In particular, it is shown that simply replacing the free far-field model of the CSRP by other acoustic models is not optimal with standard SRP beamformers. In order to overcome this limitation, we propose a generalized steered response power (GSRP) beamformer design under consideration of generic acoustic models and noise fields and derive corresponding GSRP beamformers. Moreover, we propose and analyze frequency weightings of the GSRP beamformer output as potential alternatives to the PHAT weighting. The presented methods are evaluated in realistic simulations of different scenarios.

The paper is structured as follows: In Section II, we review the generic mathematical SRP framework and the conventional SRP with its typical applications. Section III introduces the idea behind the generalization of the conventional SRP and shows limitations of standard SRP beamformers in combination with generic acoustic models. The GSRP beamformer design, the resulting beamformers, and frequency weightings are presented in Section IV. The proposed methods are evaluated in comparison to the conventional SRP in Section V.

II. The Steered Response Power Method

This section introduces the signal model and reviews the generic mathematical framework of the SRP method for acoustic source localization[1] as well as its most common variant which we refer to as conventional SRP (CSRP).

A. Signal Model in the Frequency Domain

Throughout this work, we assume that there is only one active target sound source at position $\mathbf{p}_\mathrm{s}\in\mathbb{R}^3$. The noisy sound is recorded by M microphones at different microphone positions $\mathbf{p}_m$, $m\in\{1,\ldots,M\}$. The microphone signals in the frequency domain at angular frequency $\omega = 2\pi f$ are given by

$$\mathbf{y}(\omega) = \mathbf{x}(\omega) + \mathbf{v}(\omega) = \mathbf{h}_\mathrm{s}(\omega)\, s(\omega) + \mathbf{v}(\omega)\,, \qquad (1)$$

where the vector $\mathbf{y}(\omega) = [y_1(\omega),\ldots,y_M(\omega)]^\mathrm{T}$ comprises the M complex-valued microphone signals which consist of the noise-free microphone signals $\mathbf{x}(\omega)$ and the noises $\mathbf{v}(\omega)$. Furthermore, $\mathbf{h}_\mathrm{s}(\omega) = \mathbf{h}(\omega,\mathbf{p}_\mathrm{s}) = [h_1(\omega,\mathbf{p}_\mathrm{s}),\ldots,h_M(\omega,\mathbf{p}_\mathrm{s})]^\mathrm{T}$ is the vector of acoustic transfer functions (ATFs) from the source position $\mathbf{p}_\mathrm{s}$ to the microphone positions $\mathbf{p}_1,\ldots,\mathbf{p}_M$, and $s(\omega)$ is the clean source signal. The ATFs $\mathbf{h}_\mathrm{s}(\omega)$ explicitly include all effects of sound propagation, such as reverberation, the distance-dependent attenuation, shadowing effects and the directivity of the source or the microphones. Additionally, we assume that the source signal and noise signals are independent with zero mean and thus the signal covariance matrix[2] (SCM)

$$\mathbf{\Phi}_{\mathrm{yy}}(\omega) = \mathbb{E}\{\mathbf{y}(\omega)\,\mathbf{y}^\mathrm{H}(\omega)\} = \mathbf{\Phi}_{\mathrm{xx}}(\omega) + \mathbf{\Phi}_{\mathrm{vv}}(\omega)\,, \qquad (2)$$

can be expressed as the sum of the noise-free SCM

$$\mathbf{\Phi}_{\mathrm{xx}}(\omega) = \mathbb{E}\{\mathbf{x}(\omega)\,\mathbf{x}^\mathrm{H}(\omega)\} = \varPhi_{ss}(\omega)\,\mathbf{h}_\mathrm{s}(\omega)\,\mathbf{h}_\mathrm{s}^\mathrm{H}(\omega) \qquad (3)$$

and noise covariance matrix (NCM) $\mathbf{\Phi}_{\mathrm{vv}}(\omega) = \mathbb{E}\{\mathbf{v}(\omega)\mathbf{v}^\mathrm{H}(\omega)\}$, where $\mathbb{E}\{\cdot\}$ is the expectation operator, $\{\cdot\}^\mathrm{H}$ denotes the Hermitian conjugate, and $\varPhi_{ss}(\omega) = \mathbb{E}\{|\mathbf{s}(\omega)|^2\}$ is the power spectral density (PSD) of the source signal. The noise-free SCM in (3) is a rank-one matrix when there is a single sound source. Besides, we assume that $\mathbf{\Phi}_{\mathrm{vv}}(\omega)$ is always invertible.

[1] Note that in this work, source localization explicitly refers to both the DOA estimation and the source position estimation. Throughout this work, we use the more general source position $\mathbf{p}_\mathrm{s}$ as the quantity to estimate from which the source DOA θ_s can be derived.

[2] Often also called spatial covariance matrix or spatial correlation matrix.

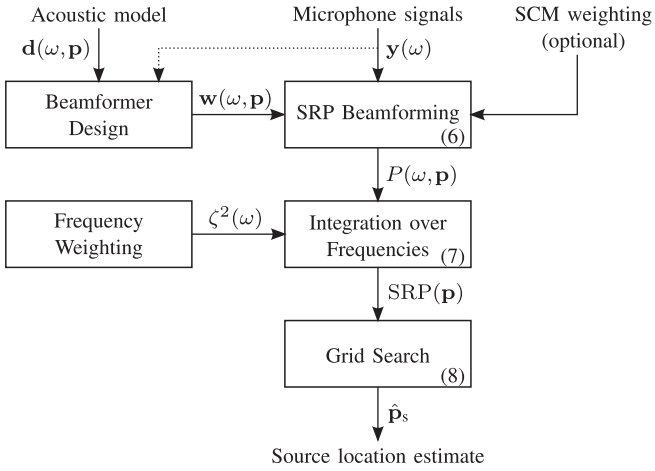


Fig. 2. Structure of the generic steered response power framework.

B. Generic SRP Framework and Conventional SRP

The structure of the generic mathematical SRP framework is shown in Fig. 2. In SRP, a beamformer is steered towards multiple candidate points to scan the space for a sound source. The beamformer output for point $\mathbf{p}$ is

$$z(\omega,\mathbf{p}) = \mathbf{w}^{\mathrm{H}}(\omega,\mathbf{p})\,\mathbf{y}(\omega) \tag{4}$$

with beamformer weights $\mathbf{w}(\omega,\mathbf{p}) = [w_1(\omega,\mathbf{p}),\ldots,w_M(\omega,\mathbf{p})]^{\mathrm{T}}$. The typical beamformer in the CSRP is a delay-and-sum (DS) beamformer [3], [4] with[3]

$$\mathbf{w}_{\mathrm{DS}}(\omega,\mathbf{p}) = \mathbf{d}_{\mathrm{ff}}(\omega,\mathbf{p}) = \left[e^{-j\omega T_1(\mathbf{p})},\ldots,e^{-j\omega T_M(\mathbf{p})}\right]^{\mathrm{T}}, \tag{5}$$

where $\mathbf{d}_{\mathrm{ff}}(\omega,\mathbf{p})$ is the acoustic free far-field model that models the time of flight $T_m(\mathbf{p})$ of the actual, unknown ATF $\mathbf{h}(\omega,\mathbf{p})$ from a position $\mathbf{p}$ to each microphone position $\mathbf{p}_m$. However, there are also publications proposing other beamformer approaches for SRP, such as the minimum power distortionless response[4] (MPDR) beamformer [32], [38], target beamforming or null-steering beamforming [31]. The PSD of the SRP beamformer output (i.e., the *steered response power*) is

$$P(\omega,\mathbf{p}) = \mathbb{E}\{|z(\omega,\mathbf{p})|^2\} = \mathbf{w}^{\mathrm{H}}(\omega,\mathbf{p})\,\mathbf{\Phi}_{\mathrm{yy}}(\omega)\,\mathbf{w}(\omega,\mathbf{p})\,. \tag{6}$$

In addition, various contributions, e.g., [3], [6], [39], [40], [41], [42], propose a weighting of the SCM $\mathbf{\Phi}_{\mathrm{yy}}(\omega)$. The most common weighting in the CSRP is the phase transform (PHAT) weighting [5] which scales each element of the SCM to magnitude one. In combination with the DS beamforming, this weighting yields the popular SRP-PHAT which is known for its robustness against reverberation [3], [6], [7], [43]. A broadband SRP value is determined by integrating the SRP beamformer PSDs $P(\omega,\mathbf{p})$ over all frequencies:

$$\mathrm{SRP}(\mathbf{p}) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\zeta^2(\omega)\;P(\omega,\mathbf{p})\;d\omega\,, \tag{7}$$

where $\zeta^2(\omega)$ is a generic, real-valued, positive frequency weighting factor. Finally, the SRP is computed at a grid of possible source locations (which we call SRP map) and the source position estimate is the position with maximum SRP output (7), i.e.,

$$\hat{\mathbf{p}}_\mathrm{s} = \arg\max_{\mathbf{p}}\;\mathrm{SRP}(\mathbf{p})\,. \tag{8}$$

III. SRP With Generic Acoustic Models

The CSRP is not capable of exploiting microphone signal level differences for localization and extracts all source position cues purely from the observed TDOAs between the microphones. This is mainly because the CSRP is based on the acoustic free far-field model of (5) which purely models delays. But also the PHAT weighting removes level information of the microphone covariances so that only phase differences remain to determine the source location. While this is suitable in typical CSRP applications with microphone arrays in the far field of the sound source, exploiting level differences might be highly desirable in generic microphone setups (cf. Fig. 1). This can be achieved by considering both level and phase information in the acoustic model. For instance, for sound source localization with distributed microphones, the acoustic free near-field model [44] can be beneficial. It is given by

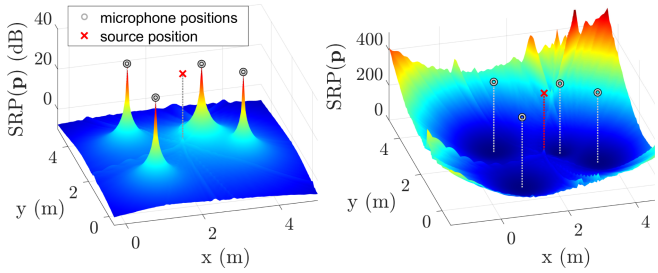
$$\mathbf{d}_{\mathrm{nf}}(\omega,\mathbf{p}) = \left[\frac{1}{4\pi r_1(\mathbf{p})}\,e^{-j\omega T_1(\mathbf{p})},\ldots,\frac{1}{4\pi r_M(\mathbf{p})}\,e^{-j\omega T_M(\mathbf{p})}\right]^{\mathrm{T}} \tag{9}$$

with $T_m(\mathbf{p}) = r_m(\mathbf{p})/c$, where $r_m(\mathbf{p}) = \|\mathbf{p}-\mathbf{p}_m\|$ is the distance between $\mathbf{p}$ and microphone m, and c is the speed of sound. It exploits that the signals in microphones in the vicinity of the sound source have a significantly higher level than in distant microphones. Moreover, one can reduce the mismatch between the acoustic model and the actual acoustical conditions by using generic acoustic models for SRP. For example, for binaural source localization with hearing aid microphones, using measured or modeled head-related transfer functions (HRTFs) as acoustic model is preferable over a free-field model. This is because by using HRTFs the acoustic shadowing effect of the head is taken into account which causes frequency-dependent interaural level and time differences that significantly differ from free-field assumptions, especially for lateral sound sources [13], [14]. However, considering generic acoustic models for SRP is not straightforward, as is shown with the following example.

Let us consider a simple simulated scenario with four distributed microphones in the free field (no reverberation). An omnidirectional sound source in the middle of the four microphones emits white noise (desired signal). In order to consider near-field effects, it would be intuitive to simply apply the free near-field model of (9) instead of the free far-field model of (5) in the conventional DS beamforming formulation of (5), i.e., $\mathbf{w}_{\mathrm{DS,nf}}(\omega,\mathbf{p}) = \mathbf{d}_{\mathrm{nf}}(\omega,\mathbf{p})$. However, we can see that the magnitude of $\mathbf{d}_{\mathrm{nf}}(\omega,\mathbf{p})$ in (9) goes to infinity if $\mathbf{p}$ approaches one

[3] In some publications, e.g., in [37], the DS beamformer weights are scaled by $1/M$. This, however, does not affect the SRP localization result.

[4] Note that the authors of [32], [38] refer to it as minimum variance distortionless response (MVDR) beamformer. However, we use the denotation MPDR to distinguish from MVDR as also done, for instance, in [37].



(a) SRP with DS beamforming. (b) SRP with MVDR beamforming.

Fig. 3. SRP maps when employing the acoustic free near-field model of (9) with a DS (5) or MVDR (13) beamformer. The simulated sound source (red cross) is located between four distributed microphones (gray circles). In this exemplary scenario, SRP localization with a DS or MVDR beamformer fails.

of the microphone positions $\mathbf{p}_m$, since $r_m(\mathbf{p})$ in the denominator of the m-th element of the vector in (9) goes to zero and thus

$$\lim_{\mathbf{p}\to\mathbf{p}_m} \|\mathbf{d}_{\text{nf}}(\omega,\mathbf{p})\| = \infty \quad \forall\, m \in \{1,\ldots,M\}\,. \tag{10}$$

Conversely, we can see that the magnitude of $\mathbf{d}_{\text{nf}}(\omega,\mathbf{p})$ diminishes for an increasing distance $r_m(\mathbf{p})\to\infty$ between $\mathbf{p}$ and all microphone positions $\mathbf{p}_m$:

$$\lim_{r_m(\mathbf{p})\to\infty} \|\mathbf{d}_{\text{nf}}(\omega,\mathbf{p})\| = 0 \quad \forall\, m \in \{1,\ldots,M\}\,. \tag{11}$$

With (6), we see that the near-field DS beamformer output power $P_{\text{DS,nf}}(\omega,\mathbf{p}) = \mathbf{d}_{\text{nf}}^{\text{H}}(\omega,\mathbf{p})\,\mathbf{\Phi}_{\text{yy}}(\omega)\,\mathbf{d}_{\text{nf}}(\omega,\mathbf{p})$ also goes to infinity if $\mathbf{p}$ gets close to a microphone position $\mathbf{p}_m$, and to zero for very distant positions, respectively. This behavior is visualized in Fig. 3(a) (in logarithmic scale).

Hence, it becomes obvious that the DS beamformer would require an appropriate normalization. Let us therefore consider an SRP beamformer with a distortionless response constraint, such as MVDR or MPDR, which normalizes the beamformer output PSD so that it equals the source PSD [37]. Under the assumption of spatially uncorrelated and homogeneous noise, the NCM and its inverse are diagonal matrices with

$$\mathbf{\Phi}_{\text{vv}}(\omega) = \sigma_v^2(\omega)\,\mathbf{I}\,, \quad \text{and} \quad \mathbf{\Phi}_{\text{vv}}^{-1}(\omega) = \frac{1}{\sigma_v^2(\omega)}\,\mathbf{I}\,, \tag{12}$$

where $\sigma_v^2(\omega)$ is the noise power and $\mathbf{I}$ is the identity matrix. With this, the MVDR beamformer simplifies to

$$\mathbf{w}_{\text{MVDR}}(\omega,\mathbf{p}) = \frac{\mathbf{\Phi}_{\text{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})}{\mathbf{d}^{\text{H}}(\omega,\mathbf{p})\,\mathbf{\Phi}_{\text{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})} \overset{(12)}{=} \frac{\mathbf{d}(\omega,\mathbf{p})}{\|\mathbf{d}(\omega,\mathbf{p})\|^2}\,, \tag{13}$$

which can be interpreted as a normalized DS beamformer. With (10) and (11), we can see that the MVDR beamformer weights of (13) behave in the opposite way to the previous DS example: The magnitude of the MVDR weights (and thus also its output PSD) goes to zero if $\mathbf{p}$ approaches a microphone position $\mathbf{p}_m$, while it goes to infinity for increasingly distant positions. It can be shown that this also holds for MPDR and for MVDR with arbitrary (invertible) NCMs. This behavior is visualized in Fig. 3(b).

This simple example shows that a correct SRP localization with DS or a distortionless response beamformer, such as MVDR, is not guaranteed with a near-field model – even in the noise-free case with no mismatch between the acoustic model and the simulated ATFs, i.e., $\mathbf{d}(\omega,\mathbf{p}) = \mathbf{h}(\omega,\mathbf{p})$. In fact, this is not an exclusive problem of the near-field model but rather it affects any acoustic model that incorporates source-position-dependent level information and not purely phase information. Moreover, also with a far-field model that ignores level information, the influence of noise on the localization result cannot be controlled with the conventional SRP. For these reasons, we propose a novel beamforming method for SRP in the following section that addresses the above mentioned problems.

IV. Generalized Steered Response Power Method

In this section, a generalized steered response power (GSRP) method for sound source localization is introduced. It generalizes the conventional SRP (see Section II-B) in the sense that it allows for using generic acoustic models by means of a specific beamformer design to avoid the effects shown in the previous section (see Fig. 3). Furthermore, generic inhomogeneous noise characteristics are taken into account. In addition, we propose appropriate frequency weightings and briefly address some practical aspects of GSRP.

A. GSRP Beamformer Design Criteria

In order to deduce an appropriate GSRP beamformer design, we start from the following obvious SRP localization objective: The SRP map $\text{SRP}(\mathbf{p})$ must have a unique global maximum at the source position $\mathbf{p}_{\text{s}}$ so that $\arg\max_{\mathbf{p}} \text{SRP}(\mathbf{p}) = \mathbf{p}_{\text{s}}$. With (7), this objective can be translated into a narrowband equivalent: The SRP beamformer output PSD $P(\omega,\mathbf{p})$ shall have a global maximum at the source position $\mathbf{p} = \mathbf{p}_{\text{s}}$ at each frequency ω, i.e.,

$$P(\omega,\mathbf{p}_{\text{s}}) \geq P(\omega,\mathbf{p}) \quad \forall\, \mathbf{p}_{\text{s}},\ \mathbf{p},\ \omega\,. \tag{14}$$

Note that we accept that the global maximum might not be unique in each frequency ω which can be, for instance, due to spatial aliasing. However, if multiple global maxima of $P(\omega,\mathbf{p})$ exist, their respective locations are frequency-dependent and thus vary over frequencies except for the commonly required, frequency-independent maximum at $\mathbf{p}_{\text{s}}$. Therefore, integrating over all frequencies in (7) usually leads to a unique global SRP maximum at $\mathbf{p}_{\text{s}}$.

Under the assumption of statistically independent source and noise signals (cf. (2)), we can split the beamformer output PSD of (6) into two PSD terms $P^{(\mathbf{x})}(\omega,\mathbf{p})$ and $P^{(\mathbf{v})}(\omega,\mathbf{p})$, which we refer to as *source response* and *noise response*, respectively:

$$P(\omega,\mathbf{p}) = \underbrace{\mathbf{w}^{\text{H}}(\omega,\mathbf{p})\,\mathbf{\Phi}_{\text{xx}}(\omega)\,\mathbf{w}(\omega,\mathbf{p})}_{\text{source response } P^{(\mathbf{x})}(\omega,\mathbf{p})} + \underbrace{\mathbf{w}^{\text{H}}(\omega,\mathbf{p})\,\mathbf{\Phi}_{\text{vv}}(\omega)\,\mathbf{w}(\omega,\mathbf{p})}_{\text{noise response } P^{(\mathbf{v})}(\omega,\mathbf{p})}\,. \tag{15}$$

We treat these two terms separately in the following GSRP beamformer design. Let us first consider the noise-free case, i.e., we only consider the source response $P^{(\mathbf{x})}(\omega,\mathbf{p})$. The condition of (14) then requires that the source response $P^{(\mathbf{x})}(\omega,\mathbf{p})$ has a

global maximum at $\mathbf{p} = \mathbf{p}_\mathrm{s}$ at each frequency. This defines a first GSRP beamformer design criterion as

$$\boxed{\begin{array}{ll}\textbf{Criterion No 1} & P^{(\mathbf{x})}(\omega, \mathbf{p}_\mathrm{s}) \geq P^{(\mathbf{x})}(\omega, \mathbf{p}) \\ max.\, source\, response \forall\, \mathbf{p},\ \mathbf{p}_\mathrm{s},\ \text{with } \Phi_{ss}(\omega) > 0\,. & \end{array}} \tag{16}$$

We refer to this as the *maximum source response criterion*, where $\Phi_{ss}(\omega) > 0$ is to avoid the trivial solution of $P^{(\mathbf{x})}(\omega, \mathbf{p}_\mathrm{s}) = P^{(\mathbf{x})}(\omega, \mathbf{p}) = 0$.

Now we take into account the noise response $P^{(\mathbf{v})}(\omega, \mathbf{p})$ in (15). We propose a second beamformer design criterion that requires that the noise response is constant for all $\mathbf{p}$:

$$\boxed{\begin{array}{ll}\textbf{Criterion No 2} & P^{(\mathbf{v})}(\omega, \mathbf{p}) = C \quad \forall\, \mathbf{p}\,, \\ constant\, noise\, response \forall\ \text{invertible } \mathbf{\Phi}_{\mathrm{vv}}(\omega)\,, & \end{array}} \tag{17}$$

where C is a positive, real-valued constant. This *constant noise response* criterion implies that the SRP map in the noise-only case shall be flat. It is intuitive that the condition of (14) is fulfilled if both GSRP beamformer design criteria (16) and (17) are met because the maximum source response criterion ensures a global maximum of $P^{(\mathbf{x})}(\omega, \mathbf{p})$ at position $\mathbf{p}_\mathrm{s}$ and the constant noise response criterion yields a position-independent, constant offset of the beamformer output PSD in (15).

B. GSRP Beamformer Derivation: The MVCNR Beamformer

We consider the following generic linear beamformer formulation as approach for the GSRP beamformer derivation:

$$\mathbf{w}(\omega, \mathbf{p}) = \alpha(\omega, \mathbf{p})\, \mathbf{A}\, \mathbf{h}(\omega, \mathbf{p})\,, \tag{18}$$

where $\mathbf{A}$ is a complex-valued Hermitian, positive-definite matrix and $\alpha(\omega, \mathbf{p})$ is a complex-valued scaling factor. Specifically, (18) is a generalized formulation of optimum beamformers including maximum signal-to-noise ratio (max-SNR), minimum mean-square error (MMSE), MVDR and MPDR beamformers [37], [45], but it also includes general data-independent beamformers [45] such as DS beamformers. We choose this approach from the family of optimum beamformers because the beamformers that are commonly used in the SRP context (cf. Section II-B) can also be assigned to this family. Hence, the beamformer of (18) can be interpreted as a generalization of existing SRP beamformers.

Criterion No 1: First, we consider the maximum source response criterion of (16). With $\mathbf{\Phi}_{\mathrm{xx}}(\omega) = \Phi_{ss}(\omega)\, \mathbf{h}_\mathrm{s}(\omega)\, \mathbf{h}_\mathrm{s}^{\mathrm{H}}(\omega)$ of (3), the source response in (15) becomes

$$P^{(\mathbf{x})}(\omega, \mathbf{p}) = \Phi_{ss}(\omega)\, \left|\mathbf{w}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{h}_\mathrm{s}(\omega)\right|^2 . \tag{19}$$

By inserting the previous equation into (16), dividing both sides of the inequality by $\Phi_{ss}(\omega)$ and taking the square root, we can reformulate the maximum source response criterion as

$$\textbf{Criterion No 1 (equivalent)} \quad \left|\mathbf{w}^{\mathrm{H}}(\omega, \mathbf{p}_\mathrm{s})\, \mathbf{h}_\mathrm{s}(\omega)\right| \geq \left|\mathbf{w}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{h}_\mathrm{s}(\omega)\right| \tag{20}$$

for all $\mathbf{p}$ and $\mathbf{p}_\mathrm{s}$. In fact, this formulation shows that the maximum source response criterion is independent of the source signal.

Now we can insert the general beamformer formulation of (18) into the maximum source response criterion of (20):

$$\left|\alpha(\omega, \mathbf{p}_\mathrm{s})\right| \left|\mathbf{h}_\mathrm{s}^{\mathrm{H}}(\omega)\, \mathbf{A}\, \mathbf{h}_\mathrm{s}(\omega)\right| \geq \left|\alpha(\omega, \mathbf{p})\right| \left|\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{A}\, \mathbf{h}_\mathrm{s}(\omega)\right| . \tag{21}$$

In the Appendix, it is shown that (21) – and thus the maximum source response criterion – is satisfied for every $\mathbf{p}$ by choosing

$$\alpha(\omega, \mathbf{p}) = \frac{\zeta(\omega)}{\sqrt{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{A}\, \mathbf{h}(\omega, \mathbf{p})}}\,, \tag{22}$$

where $\zeta(\omega)$ is any positive, real-valued scalar. With this, the general beamformer formulation of (18) results as

$$\mathbf{w}(\omega, \mathbf{p}) = \zeta(\omega)\, \frac{\mathbf{A}\, \mathbf{h}(\omega, \mathbf{p})}{\sqrt{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{A}\, \mathbf{h}(\omega, \mathbf{p})}}\,. \tag{23}$$

This beamformer fulfills the maximum source response criterion of (16) for any Hermitian, positive-definite $\mathbf{A}$.

Criterion No 2: In the next step, we must ensure that the constant noise response criterion of (17) is fulfilled. To this end, we insert (23) into the noise response $P^{(\mathbf{v})}(\omega, \mathbf{p}) = \mathbf{w}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{\Phi}_{\mathrm{vv}}(\omega)\, \mathbf{w}(\omega, \mathbf{p})$, and substitute it into (17). Thus,

$$P^{(\mathbf{v})}(\omega, \mathbf{p}) = \zeta^2(\omega)\, \frac{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{A}^{\mathrm{H}}\, \mathbf{\Phi}_{\mathrm{vv}}(\omega)\, \mathbf{A}\, \mathbf{h}(\omega, \mathbf{p})}{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{A}\, \mathbf{h}(\omega, \mathbf{p})} = C \tag{24}$$

must hold for all $\mathbf{p}$. It can be seen that this criterion is met by choosing $\mathbf{A} = \mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)$ as (24) then reduces to

$$P^{(\mathbf{v})}(\omega, \mathbf{p}) = \zeta^2(\omega)\,. \tag{25}$$

The resulting beamformer thus has a *constant noise response* over all $\mathbf{p}$. With this, we have derived a GSRP beamformer which fulfills both beamformer criteria (16) and (17) for generic ATFs $\mathbf{h}(\omega, \mathbf{p})$ and generic invertible NCMs $\mathbf{\Phi}_{\mathrm{vv}}(\omega)$:

$$\boxed{\mathbf{w}_{\mathrm{MVCNR}}(\omega, \mathbf{p}) = \zeta(\omega)\, \frac{\mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}(\omega, \mathbf{p})}{\sqrt{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\, \mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}(\omega, \mathbf{p})}}\,.} \tag{26}$$

We call this beamformer *minimum variance constant noise response (MVCNR)* beamformer, where the naming is explained in the following section.

C. Discussion of the MVCNR Beamformer

The MVCNR beamformer resembles the MVDR beamformer, however, with the main difference of the square root in the denominator in (26). In this section, we will briefly identify the relation between MVCNR and MVDR.

When we steer the MVCNR beamformer towards the source position $\mathbf{p}_\mathrm{s}$, its source response according to (19) is

$$P_{\mathrm{MVCNR}}^{(\mathbf{x})}(\omega, \mathbf{p}_\mathrm{s}) = \Phi_{ss}(\omega)\, \zeta^2(\omega)\, \mathbf{h}_\mathrm{s}^{\mathrm{H}}(\omega)\, \mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}_\mathrm{s}(\omega)\,. \tag{27}$$

From a comparison of (19) and (27) we can infer that

$$\mathbf{w}_{\mathrm{MVCNR}}^{\mathrm{H}}(\omega, \mathbf{p}_\mathrm{s})\, \mathbf{h}_\mathrm{s}(\omega) = \zeta(\omega)\, \sqrt{\mathbf{h}_\mathrm{s}^{\mathrm{H}}(\omega)\, \mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}_\mathrm{s}(\omega)}\,. \tag{28}$$

This reveals a relevant difference from MVDR, which is based on the distortionless response constraint $\mathbf{w}^{\mathrm{H}}(\omega,\mathbf{p}_{\mathrm{s}})\,\mathbf{h}_{\mathrm{s}}(\omega)=1$. Hence, MVCNR is not distortionless but its output PSD is scaled to ensure that it reaches a global maximum at $\mathbf{p}_{\mathrm{s}}$ according to (16) and (17).[5] In turn, MVCNR is not scaling-invariant with regard to the noise power (in contrast to MVDR). These aspects also become visible when rewriting (26) as

$$\mathbf{w}_{\mathrm{MVCNR}}(\omega,\mathbf{p}) = \zeta(\omega)\sqrt{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}\;\underbrace{\frac{\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}}_{\mathbf{w}_{\mathrm{MVDR}}(\omega,\mathbf{p})}. \tag{29}$$

It is noteworthy that the MVCNR beamformer of (26) can also be derived via minimum variance optimization using Lagrange multipliers (similar to the derivation of the MVDR beamformer, e.g., in [37]) with (28) as linear constraint, i.e.,

$$\min_{\mathbf{w}(\omega,\mathbf{p})}\ \mathbf{w}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{vv}}(\omega)\,\mathbf{w}(\omega,\mathbf{p})$$
$$\text{subject to}\ \ \mathbf{w}^{\mathrm{H}}(\omega,\mathbf{p})\,\mathbf{h}(\omega)=\zeta(\omega)\,\sqrt{\mathbf{h}^{\mathrm{H}}(\omega)\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{h}(\omega)}\,. \tag{30}$$

Due to this similarity to MVDR, however with a constant noise response instead of a distortionless signal response, we call it *minimum variance constant noise response*. Its output PSD according to (6) is

$$P(\omega,\mathbf{p})=\zeta^2(\omega)\frac{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{h}(\omega,\mathbf{p})}. \tag{31}$$

Interestingly, a similar term was derived independently from this work via a deterministic maximum likelihood approach for binaural DOA estimation with hearing aids in [31].

D. Simplification for Uncorrelated and Homogeneous Noise

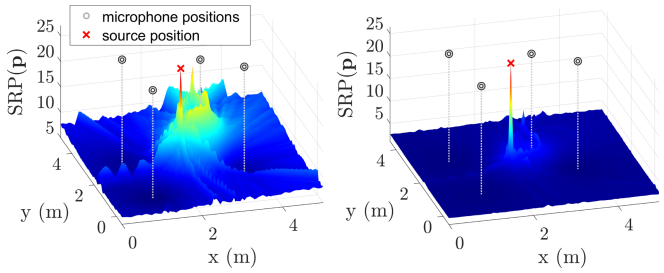
Under the assumption of spatially uncorrelated and homogeneous noise with $\boldsymbol{\Phi}_{\mathrm{vv}}(\omega)=\sigma_v^2(\omega)\,\mathbf{I}$ according to (12), the MVCNR beamformer weights of (26) simplify to

$$\boxed{\mathbf{w}_{\mathrm{NMF}}(\omega,\mathbf{p})=\frac{\zeta(\omega)}{\sigma_v(\omega)}\,\frac{\mathbf{h}(\omega,\mathbf{p})}{\|\mathbf{h}(\omega,\mathbf{p})\|}}\,. \tag{32}$$

We can recognize this as a unit-length *normalized matched filter (NMF)* [46] that is scaled by the positive, real-valued factor $\zeta(\omega)/\sigma_v(\omega)$. In [31], an un-scaled version of this beamformer is proposed as target beamformer amongst others in the context of binaural DOA estimation. Its output PSD is

$$P_{\mathrm{NMF}}(\omega,\mathbf{p})=\frac{\zeta^2(\omega)}{\sigma_v^2(\omega)}\,\frac{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)\,\mathbf{h}(\omega,\mathbf{p})}{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\,\mathbf{h}(\omega,\mathbf{p})}\,. \tag{33}$$

[5]Note that in special cases, other SRP beamformers might also satisfy the GSRP beamformer criteria of (16) and (17). For instance, one can show that MVDR and also DS fulfills the criterion No 1 (16) in the acoustic far field, and criterion No 2 (17) if, in addition, the noise is spatially uncorrelated and homogeneous. However, in general, this is not the case.



(a) SRP with MVCNR beamforming. (b) SRP with MPCNR beamforming.

Fig. 4. SRP maps when employing the acoustic free near-field model of (9) with the proposed MVCNR (a) or MPCNR (b) beamformer. The simulated signals equal those used in Fig. 3. In contrast to DS or MVDR, the maximum of the SRP maps with MVCNR or MPCNR coincide with the source position.

E. GSRP Beamformer Derivation Continued: The MPCNR

The MVCNR beamformer is not a unique solution for the GSRP beamformer design problem described in Section IV-A. In order to show this, we refer to the fact that MVDR and MPDR theoretically coincide $\mathbf{w}_{\mathrm{MVDR}}(\omega,\mathbf{p}_{\mathrm{s}})=\mathbf{w}_{\mathrm{MPDR}}(\omega,\mathbf{p}_{\mathrm{s}})$ without model errors, i.e., $\mathbf{d}(\omega,\mathbf{p}_{\mathrm{s}})=\mathbf{h}(\omega,\mathbf{p}_{\mathrm{s}})$, when they are steered towards $\mathbf{p}_{\mathrm{s}}$ [37], [47]. This property can be used to derive the minimum power pendant of the MVCNR beamformer. Specifically, the respective beamformer weights can be determined by replacing the MVDR weights $\mathbf{w}_{\mathrm{MVDR}}(\omega,\mathbf{p})$ in (29) by those of an MPDR beamformer, i.e.,

$$\mathbf{w}_{\mathrm{MPCNR}}(\omega,\mathbf{p}) = \zeta(\omega)\sqrt{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}\;\underbrace{\frac{\boldsymbol{\Phi}_{\mathrm{yy}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}{\mathbf{h}^{\mathrm{H}}(\omega,\mathbf{p})\boldsymbol{\Phi}_{\mathrm{yy}}^{-1}(\omega)\mathbf{h}(\omega,\mathbf{p})}}_{\mathbf{w}_{\mathrm{MPDR}}(\omega,\mathbf{p})}. \tag{34}$$

In accordance with MVCNR, we call this beamformer *minimum power constant noise response* (MPCNR). Similar as in (30), this solution can also be derived via minimum power optimization with $\min_{\mathbf{w}(\omega,\mathbf{p})}\ \mathbf{w}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)\,\mathbf{w}(\omega,\mathbf{p})$ subject to (28). Moreover, one can show that the MPCNR beamformer also fulfills the GSRP beamformer condition of (14).[6] However, preliminary results showed that localization with MPCNR in practice is highly prone to model errors. This also relates to the fact that MPDR is known to be less robust against versatile perturbations such as model mismatches compared to MVDR [37], [47]. For this reason, we refrain from a systematic analysis of the MPCNR beamformer in the remainder of this paper to focus on the more robust variants MVCNR and NMF.

For comparison, Fig. 4 shows the SRP maps of the same simulated scenario as in Fig. 3 using MVCNR and MPCNR (with $\zeta(\omega)=1$) with the acoustic near-field model of (9). Unlike with DS or MVDR, the SRP maps with MVCNR or MPCNR have a global maximum an the true source position.

[6]The (not straightforward) proof is omitted here because this work focuses on the more practical MVCNR and NMF beamformers.

F. Frequency Weighting

Whereas the previous sections dealt with the GSRP design of (14) at each frequency ω independently, this section proposes and analyzes frequency weightings for the presented GSRP beamformers. In the generic SRP framework (see Fig. 2), the frequency weighting is introduced as positive, real-valued factor $\zeta^2(\omega)$ in (7). We can also recognize this scaling factor in the output PSDs of the GSRP beamformers in (31) and (33). Hence, $\zeta^2(\omega)$ can be used to scale the contribution of each frequency to the broadband SRP result. Below, we propose frequency weightings specifically for MVCNR which, however, apply in the same way for the simplified NMF.

With (25) and (27), the output PSD of the MVCNR beamformer in (15), when steered towards $\mathbf{p}_\mathrm{s}$, yields

$$P_{\mathrm{MVCNR}}(\omega, \mathbf{p}_\mathrm{s}) = \zeta^2(\omega) \left[\Phi_{ss}(\omega)\, \mathbf{h}_\mathrm{s}^\mathrm{H}(\omega)\, \mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}_\mathrm{s}(\omega) + 1\right]. \tag{35}$$

This can be rewritten with the trace operator $\mathrm{tr}(\cdot)$ and (3) as

$$P_{\mathrm{MVCNR}}(\omega, \mathbf{p}_\mathrm{s}) = \zeta^2(\omega) \left[\mathrm{tr}\left(\mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{\Phi}_{\mathrm{xx}}(\omega)\right) + 1\right], \tag{36}$$

which is due to the circular shift invariance of the trace operator, i.e., $\mathbf{h}_\mathrm{s}^\mathrm{H}(\omega)\, \mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}_\mathrm{s}(\omega) = \mathrm{tr}(\mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{h}_\mathrm{s}^\mathrm{H}(\omega)\, \mathbf{h}_\mathrm{s}(\omega))$.

1) SNR Weighting: Let us first consider a fixed, signal- and frequency-independent weight such as $\zeta_{\mathrm{SNR}}^2(\omega) = 1/M$, where the subscript refers to the signal-to-noise ratio (SNR). Inserting into (36) yields

$$P_{\text{MVCNR-SNR}}(\omega, \mathbf{p}_\mathrm{s}) = \frac{1}{M}\, \mathrm{tr}\left(\mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{\Phi}_{\mathrm{xx}}(\omega)\right) + \frac{1}{M}. \tag{37}$$

This term can be interpreted as a lower-limited, narrowband-SNR-dependent weighting since the scaling depends on the noise-free SCM $\mathbf{\Phi}_{\mathrm{xx}}(\omega)$ and the inverse NCM $\mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)$. In particular, when assuming spatially uncorrelated and homogeneous noise according to (12), equation (37) simplifies to

$$P_{\text{MVCNR-SNR}}(\omega, \mathbf{p}_\mathrm{s}) \overset{(12)}{=} \frac{\mathrm{tr}(\mathbf{\Phi}_{\mathrm{xx}}(\omega))}{\mathrm{tr}(\mathbf{\Phi}_{\mathrm{vv}}(\omega))} + \frac{1}{M} = \mathrm{SNR}(\omega) + \frac{1}{M}, \tag{38}$$

where $\mathrm{tr}(\mathbf{\Phi}_{\mathrm{vv}}(\omega)) = M\, \sigma_v^2$ and the narrowband SNR is

$$\mathrm{SNR}(\omega) = \frac{\mathbb{E}\{\mathbf{x}^\mathrm{H}(\omega)\, \mathbf{x}(\omega)\}}{\mathbb{E}\{\mathbf{v}^\mathrm{H}(\omega)\, \mathbf{v}(\omega)\}} = \frac{\mathrm{tr}(\mathbf{\Phi}_{\mathrm{xx}}(\omega))}{\mathrm{tr}(\mathbf{\Phi}_{\mathrm{vv}}(\omega))}. \tag{39}$$

The contribution of each frequency thus is directly related to its narrowband SNR. This is also visualized in Fig. 5 where $P_{\text{MVCNR-SNR}}(\omega, \mathbf{p}_\mathrm{s})$ of (38) is plotted (dotted line) for spatially uncorrelated and homogeneous noise with various SNRs.

2) Spectral Flattening: The signal or noise PSD levels can vary highly over frequencies, for instance, because of the frequency sparsity of speech or due to colored noise. As a consequence, the SNR weighting might induce a highly non-uniform contribution of different frequencies where only a few single frequencies with highest narrowband SNR predominate the SRP result. In order to prevent this, we propose a spectral flattening that equalizes the contribution of each frequency to the broadband SRP result. This can be achieved, for instance, by enforcing $P_{\mathrm{MVCNR}}(\omega, \mathbf{p}_\mathrm{s}) = 1$ for all ω regardless of the narrowband SNR

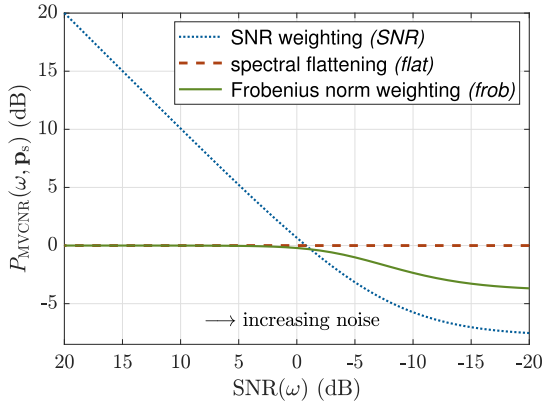


Fig. 5. Comparison of the GSRP beamformer output PSD $P_{\mathrm{MVCNR}}(\omega, \mathbf{p}_\mathrm{s})$ of (36) for all presented frequency weightings as a function of the narrowband SNR of (39) with spatially uncorrelated and homogeneous noise. Simulation details: $\mathbf{\Phi}_{\mathrm{xx}}(\omega)$ is a randomly generated $M \times M$ Hermitian rank-one matrix with $M = 6$; $\mathbf{\Phi}_{\mathrm{vv}}(\omega)$ as in (12) and scaled according to $\mathrm{SNR}(\omega)$.

(see Fig. 5, dashed line). With (36), this directly yields the weighting

$$\zeta_{\mathrm{flat}}^2(\omega) = \left(\mathrm{tr}\left(\mathbf{\Phi}_{\mathrm{vv}}^{-1}(\omega)\, \mathbf{\Phi}_{\mathrm{xx}}(\omega)\right) + 1\right)^{-1}. \tag{40}$$

3) Frobenius Norm Weighting: We propose the following weighting as practical simplification of the spectral flattening:

$$\zeta_{\mathrm{frob}}^2(\omega) = \frac{\sigma_v^2(\omega)}{\|\mathbf{\Phi}_{\mathrm{yy}}(\omega)\|_F}, \tag{41}$$

where $\sigma_v^2(\omega) = \mathbb{E}\{\mathbf{v}^\mathrm{H}(\omega)\, \mathbf{v}(\omega)\}/M = \mathrm{tr}(\mathbf{\Phi}_{\mathrm{vv}}(\omega))/M$ is the average noise power and $\|\mathbf{A}\|_F = \sqrt{\sum_{m=1}^{M} \sum_{n=1}^{N} |a_{mn}|^2}$ denotes the Frobenius norm of an $M \times N$ matrix $\mathbf{A}$ with elements a_{mn}. For spatially uncorrelated, homogeneous noise according to (12), it can be shown that $\zeta_{\mathrm{frob}}^2(\omega) \approx \zeta_{\mathrm{flat}}^2(\omega)$ if $\mathrm{SNR}(\omega) > 0$ dB, while $\zeta_{\mathrm{frob}}^2(\omega)$ asymptotically approaches $\zeta_{\mathrm{flat}}^2(\omega)/\sqrt{M}$ at noise-dominated frequencies. In other words, each frequency has the same contribution at high SNR while highly noisy frequencies are attenuated up to a factor of $\sqrt{M}$. This also becomes visible in Fig. 5 (solid line).

Note that when applying the Frobenius norm weighting to the NMF beamformer, its output PSD (33) simplifies to

$$P_{\text{NMF-frob}}(\omega, \mathbf{p}) = \frac{\mathbf{h}^\mathrm{H}(\omega, \mathbf{p})}{\|\mathbf{h}(\omega, \mathbf{p})\|} \frac{\mathbf{\Phi}_{\mathrm{yy}}(\omega)}{\|\mathbf{\Phi}_{\mathrm{yy}}(\omega)\|_F} \frac{\mathbf{h}(\omega, \mathbf{p})}{\|\mathbf{h}(\omega, \mathbf{p})\|}, \tag{42}$$

which does not require a noise power estimate anymore. One might observe a similarity between the Frobenius norm weight in the previous equation and the element-wise PHAT weighting [5] used in SRP-PHAT. In fact, the Frobenius norm weighting in (42) can be interpreted as a power-difference-preserving counterpart of the PHAT weighting, which yields an equalization of the SCM Frobenius norm over all frequencies while preserving the power differences of the microphone PSDs in the SCM. As a consequence, the Frobenius norm weight coincides with the PHAT weighting (except for a scaling by $1/M$) if no power differences are observed between the microphones (e.g., in a compact microphone array in the free far-field of the source).

TABLE I
OVERVIEW OF THE PROPOSED GENERALIZED SRP BEAMFORMERS

SRP beamformer	Beamformer weights $\mathbf{w}(\omega,\mathbf{p})$		Output PSD $P(\omega,\mathbf{p})$	
Minimum variance constant noise response *(MVCNR)*	$\zeta(\omega)\,\frac{\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})}{\sqrt{\mathbf{d}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})}}$	cf. (26)	$\zeta^2(\omega)\,\frac{\mathbf{d}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})}{\mathbf{d}^{\mathrm{H}}(\omega,\mathbf{p})\,\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})}$	cf. (31)
Normalized matched filter *(NMF)*	$\frac{\zeta(\omega)}{\sigma_v(\omega)}\,\frac{\mathbf{d}(\omega,\mathbf{p})}{\|\mathbf{d}(\omega,\mathbf{p})\|}$	cf. (32)	$\frac{\zeta^2(\omega)}{\sigma_v^2(\omega)}\,\frac{\mathbf{d}^{\mathrm{H}}(\omega,\mathbf{p})\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)\mathbf{d}(\omega,\mathbf{p})}{\mathbf{d}^{\mathrm{H}}(\omega,\mathbf{p})\mathbf{d}(\omega,\mathbf{p})}$	cf. (33)

TABLE II
PROPOSED FREQUENCY WEIGHTINGS FOR THE GENERALIZED SRP

Method	Frequency weight $\zeta^2(\omega)$	
SNR weighting *(SNR)*	$\frac{1}{M}$	
Spectral flattening *(flat)*	$\frac{1}{\mathrm{tr}(\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\boldsymbol{\Phi}_{\mathrm{yy}}(\omega))-M+1}$	cf. (40)
Frobenius norm weighting *(frob)*	$\frac{\sigma_v^2(\omega)}{\|\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)\|_F}$	cf. (41)

G. GSRP Beamforming in Practice

This section briefly discusses relevant aspects of the practical application of the above-presented GSRP method for acoustic source localization.

Acoustic Models: In practice, the ATFs $\mathbf{h}(\omega,\mathbf{p})$ in the proposed beamformers are unknown and thus are replaced by an acoustic model $\mathbf{d}(\omega,\mathbf{p})$. The respective beamformer weights and output PSDs are listed in Table I. Moreover, it is worth mentioning that the output PSDs of the GSRP beamformers are invariant with regard to a (frequency-dependent) complex-valued scaling of $\mathbf{d}(\omega,\mathbf{p})$. As a consequence, the beamformer output PSD is identical regardless of whether applying ATFs or relative transfer functions (RTFs) as acoustic model $\mathbf{d}(\omega,\mathbf{p})$ since RTFs are scaled ATFs.

Frequency Weightings: In (40), we introduced the spectral flattening which depends on the inverse NCM and the noise-free SCM. While the NCM can be estimated directly, e.g., during pauses of the desired signal, the unknown noise-free SCM can not be estimated directly. We propose to approximate $\boldsymbol{\Phi}_{\mathrm{xx}}(\omega)=\boldsymbol{\Phi}_{\mathrm{yy}}(\omega)-\boldsymbol{\Phi}_{\mathrm{vv}}(\omega)$ using (2), which allows us to rewrite $\mathrm{tr}(\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\boldsymbol{\Phi}_{\mathrm{xx}}(\omega))$ as $\mathrm{tr}(\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\boldsymbol{\Phi}_{\mathrm{yy}}(\omega))-M$. The resulting frequency weight is listed in Table II.

SNR-Dependent NCM Regularization: The proposed MVCNR beamformer (Table II) indicates a noise-dependent linear transformation of the acoustic model vector $\mathbf{d}(\omega,\mathbf{p})$ in the term $\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)\,\mathbf{d}(\omega,\mathbf{p})$. In practice, this can be unfavorable in the low-noise case as $\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}(\omega)$ might induce a considerable transformation of the model $\mathbf{d}(\omega,\mathbf{p})$ even though the noise component in the signals actually is negligible. To this end, we propose a simple regularization of the NCM by means of adding a scaled identity matrix:

$$\boldsymbol{\Phi}_{\mathrm{vv,reg}}(\omega)=\boldsymbol{\Phi}_{\mathrm{vv}}(\omega)+\epsilon_{\mathrm{reg}}\,\sigma_y^2(\omega)\,\mathbf{I}\,, \tag{43}$$

where ϵ_{reg} is a small, positive, real-valued regularization factor and $\sigma_y^2(\omega)=\mathrm{tr}(\boldsymbol{\Phi}_{\mathrm{yy}}(\omega))/M$ is the average microphone signal power. For low noise with $\sigma_v^2(\omega)\ll\sigma_y^2(\omega)$, the regularized NCM approaches a scaled identity matrix (i.e., only a scaling and no transformation of $\mathbf{d}(\omega,\mathbf{p})$) whereas the NCM is virtually unmodified in the high-noise case.

Spatial Aliasing, Spatial Sampling and Signal Bandwidth: Spatial aliasing is a common problem of beamformer-based localization [48]. It becomes especially severe with distributed microphone setups as the aliasing frequency reduces with increasing microphone distance. However, not only spatial aliasing needs to be considered but there is also a relation between the signal bandwidth and the spatial SRP resolution, i.e., the spatial resolution of the SRP grid search in (8): A lower spatial SRP resolution reduces the upper frequency of the usable signal bandwidth [49]. Considering these effects might be crucial in practice, especially for setups involving larger microphone distances. However, a detailed analysis of these aspects is beyond the scope of this paper.

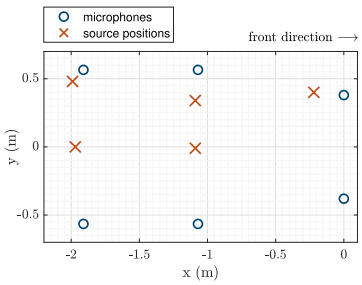
V. EVALUATIONS

A. Simulation Setups

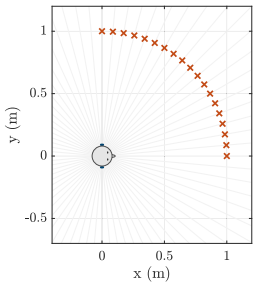
The proposed generalized SRP beamforming approaches are evaluated in comparison to the conventional SRP method in three scenarios with different microphone constellations with speech as source signal.

1) Car Setup – Speaker Position Estimation With Distributed Microphones in a Car: The first scenario uses six distributed omnidirectional microphones in the car roof in a minivan (three seat rows with two distributed microphones per row) with reverberation time $T_{60}\approx 90$ ms. We evaluated five different speaker positions with frontal head orientation. The microphones are elevated 30 cm above the mouth of the speaker. The geometry of the setup is plotted in Fig. 6(1). For each speaker position, we generated six-seconds-long microphone signals (three seconds of stationary noise followed by three seconds of noisy speech) of two female and two male speakers with driving noise at different speeds. The noise-free speech signals are simulated using clean speech snippets of the Clarity Speech Corpus [50], which are convolved with measured in-car room impulse responses (RIRs) from [51]. The RIRs were captured using a mouth simulator with frontal orientation. The used dataset also contains multichannel driving noise recordings at stationary speeds between 0 km/h and 150 km/h [51].

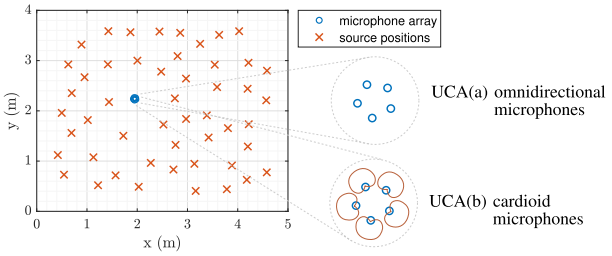
2) HA Setup – DOA Estimation With Binaural Hearing Aid Microphone Arrays in an Office: The second scenario uses two binaural behind-the-ear hearing aid (HA) devices mounted on a head and torso simulator (HATS) in an office room with



(1) Car setup: Speaker position estimation with six distributed microphones in a car (microphones are elevated 30 cm above the source positions).



(2) HA setup: DOA estimation with two behind-the-ear hearing aid devices mounted on a HATS in an office.



(3) UCA setup: DOA estimation with a uniform circular array (UCA) with 5 cm diameter involving (a) five omnidirectional microphones or (b) five outward-oriented cardioid microphones in a simulated room.

Fig. 6. Geometry of the three evaluation setups.

reverberation time $T_{60} \approx 0.4$ s. Each HA device comprises three microphones with a distance of approximately 8 mm (cf. [52] for further details). We evaluated 19 different speaker DOAs (elevation 0°) between an azimuth of 90° (left) and 0° (front) in steps of 5° (see Fig. 6(2)). For each speaker DOA, we generated six-seconds-long microphone signals (three seconds noise only plus three seconds noisy speech) of two female and two male speakers with multi-talker babble noise at different SNRs. The noise-free speech signals are simulated using the same clean speech snippets as in the previous car setup, which are convolved with measured binaural room impulse responses (BRIRs) from [52] (office I). Stationary, diffuse babble noise was simulated according to [53] with the NCM $\mathbf{\Phi}_{\text{vv, diff}}(k) = \sum_{\theta \in [0 \ldots 360°]} \mathbf{h}_{\text{HRTF}}(k,\theta)\, \mathbf{h}_{\text{HRTF}}^{\text{H}}(k,\theta)$, where k is the discrete Fourier transform frequency index and $\mathbf{h}_{\text{HRTF}}(k,\theta)$ are the Fourier-transformed anechoic HRIRs of [52] with DOA azimuth θ (in steps of 5°).

3) UCA Setup – DOA Estimation With a Uniform Circular Array (UCA) in a Simulated Room: The third scenario uses a 5-microphone UCA with 5 cm diameter involving (a) omnidirectional microphones and (b) outward-oriented cardioid microphones in a simulated room with reverberation time $T_{60} = 0.6$ s. This simulated setup was added in order to evaluate the applicability of the proposed GSRP methods in a (a) typical setup of the conventional SRP (i.e., source in far field of the array without relevant source-position-dependent level differences), and in comparison to a (b) setup involving an array of directional microphones (source in far field of the array, but relevant source-position-dependent level differences appear due to the microphone directivities). We evaluated 50 randomly chosen speaker positions in the room (see Fig. 6(3)). For each speaker position, we generated six-seconds-long microphone signals (three seconds noise only plus three seconds noisy speech) of two female and two male speakers in an isotropic noise field (pink noise) at different SNRs. The noise-free speech signals are simulated using the same clean speech snippets as in the car setup, which are convolved with simulated RIRs using [54], [55]. Stationary, diffuse pink noise was simulated with [53] where the spatial coherence function of the microphone array was computed according to [56] incorporating the different microphone directivities of setups (a) and (b).

B. SRP Evaluation Method

We compared the performance of the proposed GSRP beamforming approaches against the conventional SRP by means of time-averaged SRP maps and the mean localization error. In the *car setup*, the absolute localization error is the distance of the source position estimate $\hat{\mathbf{p}}_{\text{s}}$ from the true source position $\mathbf{p}_{\text{s}}$, i.e., $E_{\text{pos}} = \| \hat{\mathbf{p}}_{\text{s}} - \mathbf{p}_{\text{s}} \|$. The absolute localization error in the *HA setup* and the *UCA setup* is the angular deviation $E_{\text{ang}} = |\hat{\theta}_{\text{s}} - \theta_{\text{s}}|$ between the surce DOA estimate $\hat{\theta}_{\text{s}}$ and the ground truth θ_{s}. The mean localization error is computed by averaging the absolute localization error over time and over multiple speakers.

1) Evaluated SRP Algorithms: As baseline, we considered the conventional *SRP-PHAT* and the *CSRP* without PHAT weighting (cf. Section II-B). We compared the baseline against the proposed GSRP algorithms (cf. Table I). In particular, we considered *MVCNR* in combination with the proposed frequency weightings *SNR*, *flat*, *frob* (cf. Table II), and the simpler *NMF* (cf. Table I) in combination with the *SNR* and *frob* weighting. The *flat* weighting was not considered in combination with the *NMF*. This is because the practical benefit of *NMF* over *MVCNR* is that no NCM estimate is required which, however, is required by the *flat* weighting.

2) Applied Acoustic Models: As discussed in Section II-B, the baseline approaches *SRP-PHAT* and *CSRP* are based on the free far-field *(FF)* model of (5). By contrast, GSRP can employ arbitrary acoustic models. For the *car setup* with distributed microphones, we considered the free near-field *(NF)* model $\mathbf{d}_{\text{nf}}(k,\mathbf{p})$ of (9). In addition, we took into account the source directivity by exploiting knowledge of the frontal orientation of the speakers in the car. The directivity of human speech can be modeled [57]. However, we used a dataset of measured, average speech directivities [58], [59]. The respective acoustic model for the GSRP methods in the car setup thus is determined as $\mathbf{d}_{\text{GSRP}}(k,\mathbf{p}) = \mathbf{d}_{\text{nf}}(k,\mathbf{p}) \odot \mathbf{d}_{\text{sd}}(k,\mathbf{p},0°)$, where $\mathbf{d}_{\text{sd}}(k,\mathbf{p},0°)$ is the frequency-dependent average speech directivity of a speaker at position $\mathbf{p}$ with orientation 0° (frontal orientation) towards all microphones and $\odot$ denotes the element-wise Hadamard product. In the car setup, the acoustic models used for CSRP and GSRP thus have the same phase terms (cf. (5) and (9)) but differ in the level.

In the *HA setup*, we used measured anechoic HRTFs of [52] as acoustic model for the evaluated GSRP methods, i.e., $\mathbf{d}_{\text{GSRP}}(k,\theta) = \mathbf{h}_{\text{HRTF}}(k,\theta)$. As discussed in Section III, the HRTFs comprise frequency-dependent interaural level and time

TABLE III
SRP IMPLEMENTATION DETAILS OF THE EVALUATION SETUPS

Sampling frequency f_s	16 kHz
STFT frame size | frame shift | window	512 samples | 256 samples | Hann
SCM recursive smoothing τ_{sm} | α_{sm}	75 ms | 0.2
NCM regularization factor ϵ_{reg}	0.01
	car setup / *HA / UCA setup*
SRP frequency boundaries	100 Hz ... 4 kHz / 100 Hz ... 8 kHz
Spatial SRP resolution	5 cm / 5°

differences that are caused by the acoustic head shadow. Therefore, the HRTFs differ from the CSRP free-field model of (5) in terms of both phase and level. As an intermediate approach, we thus evaluated an additional, phase-corrected SRP-PHAT method. It uses the phase of the HRTFs as acoustic model $d_{\text{HRTF-PHAT},m}(k,\theta) = h_{\text{HRTF},m}(k,\theta)/|h_{\text{HRTF},m}(k,\theta)|$ instead of the free field model phase as proposed in [13].

In the *UCA(a) setup* involving omnidirectional microphone, we used the acoustic free far-field model $\mathbf{d}_{\text{ff}}(\omega,\theta)$ of (5) for both the conventional SRP methods and the GSRP methods. In the *UCA(b) setup* involving cardioid microphones, we considered the microphone directivities *(MD)* in addition to the far-field model for GSRP, i.e., $\mathbf{d}_{\text{GSRP}}(\omega,\theta) = \mathbf{d}_{\text{ff}}(\omega,\theta) \odot \mathbf{d}_{\text{MD}}(\theta)$. The directivity pattern of a cardioid microphone is given by $d_{\text{MD},m}(\theta) = 0.5\,(1+\cos\angle(\phi_m,\theta))$ [56], where ϕ_m reflects the orientation of microphone m, and $\angle(\phi_m,\theta)$ denotes the intermediate angle between ϕ_m and θ.

3) SRP Implementation Details: We implemented the SRP localization in the short-time Fourier transform (STFT) domain using Matlab. All relevant implementation parameters can be found in Table III. The NCM was estimated by averaging over all instantaneous SCM estimates of the noise-only signal part (first three seconds) with $\hat{\mathbf{\Phi}}_{\text{vv}}(k) = 1/L_{\text{n}} \sum_{l=1}^{L_{\text{n}}} \mathbf{y}(k,l)\,\mathbf{y}^{\text{H}}(k,l)$ where $\mathbf{y}(k,l)$ are the STFT signals of frequency bin k and time frame l, and L_{n} is the number of frames of the noise-only signal part. During the second, speech-plus-noise signal part, the SCM $\hat{\mathbf{\Phi}}_{\text{yy}}(k,l)$ was estimated in each frame by recursive smoothing over instantaneous SCM estimates with smoothing time constant τ_{sm} =75 ms (corresponding to smoothing factor $\alpha_{\text{sm}} = 0.2$). The SRP map of (7) and the localization error was computed in each frame with speech activity. Only frequencies above 100 Hz were taken into account (by setting $\zeta^2(\omega) = 0$ in (7) for frequencies below 100 Hz) since no relevant speech energy can be expected below. Furthermore, frequencies in the car setup were upper limited to 4 kHz because incorporating higher frequencies drastically increases the localization error of all evaluated SRP methods. This is due to aliasing effects that mutually limit the usable signal bandwidth with a given spatial SRP map resolution as described in [49], especially if the source is close to the microphones. As a consequence, we must limit the bandwidth to reduce undersampling of the SRP map.

C. Computational Complexity

The computational complexity of the evaluated SRP methods differs in terms of the SRP beamforming method and the frequency weighting. Table IV gives an overview of the most relevant computational differences between the beamforming methods *CSRP* (including *SRP-PHAT*), *NMF* and *MVCNR*. While *NMF-SNR* requires a noise power estimate which scales the beamformer weights of (32), the *MVCNR* method requires the estimation and inversion of the full NCM. These differences in computational complexity are particularly evident in real-time processing where the beamformer weights can be offline precomputed for *CSRP* and *NMF* (except for a scalar multiplication for *NMF-SNR*). By contrast, the *MVCNR* beamformer weights need to be updated in real time once a new NCM estimate is available. Table V shows a comparison of the average processing time per STFT frame (the frame time period is 160 μs) and the processing time factor relative to the *CSRP* method of our Matlab implementation of the *car setup*.

TABLE IV
COMPARISON OF THE COMPUTATIONAL COMPLEXITY OF THE PRESENTED SRP BEAMFORMERS

	CSRP	**NMF**	**MVCNR**
Noise estimation	–	*NMF-SNR*: noise power *NMF-frob*: –	full NCM
NCM inversion	–	–	yes
Noise-dependent beamformer weights	–	*NMF-SNR*: scalar only *NMF-frob*: –	yes

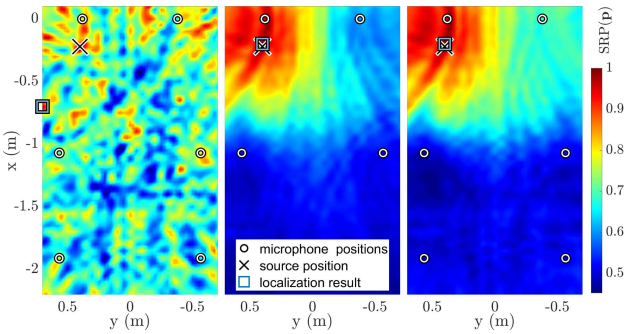
TABLE V
PROCESSING TIME OF MATLAB IMPLEMENTATION (CAR SETUP)

SRP method	**Processing time per frame** absolute / % of frame time period	**Time factor**
CSRP	2.71 µs / 1.70 %	1.00
SRP-PHAT	3.22 µs / 2.01 %	1.19
NMF-SNR	4.47 µs / 2.79 %	1.65
NFM-frob	4.57 µs / 2.86 %	1.69
MVCNR-SNR	6.55 µs / 4.09 %	2.42
MVCNR-flat	6.80 µs / 4.25 %	2.51
MVCNR-frob	6.59 µs / 4.12 %	2.43

D. Results

1) Car Setup: Fig. 7 shows time-averaged SRP heatmaps (i.e., the frame-wise SRP maps $\text{SRP}(\mathbf{p})$ of (7) are averaged over three seconds of noisy speech) of a speaker at the driver position at driving speed 120 km/h for the conventional *SRP-PHAT* and the GSRP methods *MVCNR-frob* and *NMF-frob*. The x- and y-axes indicate the Cartesian x and y coordinates of point $\mathbf{p}$. For better visualization, the SRP heatmaps are normalized to a value of one at the SRP maximum (square marker). The *SRP-PHAT* map is very rough and has multiple sharp peaks. Its maximum (square marker) does not coincide with the true source position. By contrast, in the *MVCNR-frob* and *NMF-frob* maps, a larger area around the source position is elevated whereas areas in the SRP map which are distant from the source are consistently suppressed. The *MVCNR-frob* map shows a better suppression of the region on the right-hand side of the source position (i.e., the front passenger seat in the car) compared to *NMF-frob*. The maximum of both GSRP maps coincides with the source position whereas, in general, the peaks are less sharp than those in the *SRP-PHAT* map.

For a more systematic evaluation of all SRP variants, we computed the mean localization error (MLE). To this end, the



(a) SRP-PHAT *(FF)* (b) MVCNR-frob *(NF)* (c) NMF-frob *(NF)*

Fig. 7. Car setup: Time-averaged SRP maps over 3 s speech from the driver position at 120 km/h driving speed (equivalent to −4 dB SNR averaged over all microphones and 4 dB SNR in the closest microphone, respectively). The acoustic model used in each evaluated SRP method is indicated in parentheses. (FF: free far-field model; NF: free near-field model.).

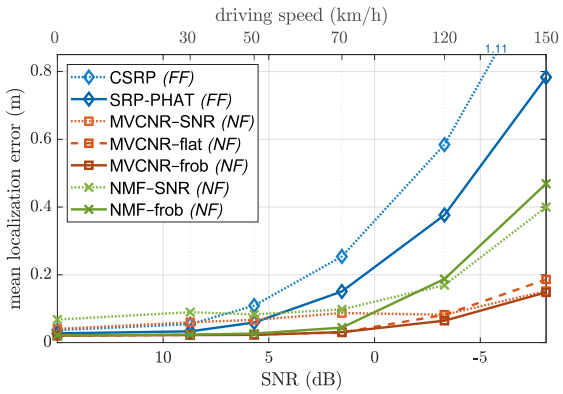


Fig. 8. Car setup: Mean localization error of all evaluated SRP methods over the microphone-averaged SNR. The respective driving speed is specified on the top axis. The acoustic model used in each evaluated SRP method is indicated in parentheses.

absolute localization error $E_{\text{pos}} = \| \hat{\mathbf{p}}_{\text{s}} - \mathbf{p}_{\text{s}} \|$ was computed at each time frame with active speech and averaged over all frames, source positions and speakers. In Fig. 8, the MLE (in meters) is plotted as a function of the driving speed and the respective average SNR over all microphones for all evaluated SRP methods. At SNRs greater than 10 dB, the MLE is low for all evaluated SRP methods. At increasing speed (decreasing SNR), the performance between the different methods is increasingly diverging and the proposed GSRP approaches *MVCNR* and *NMF* in combination with any frequency weighting distinctly outperform the conventional methods *SRP-PHAT* and *CSRP*. The MLE of *MVCNR* and *NMF* is comparable at positive SNRs whereas *MVCNR*, which includes a NCM estimate, outperforms *NMF* at negative SNRs and shows the best overall performance. The *flat* and *frob* weighting with *MVCNR* perform similarly. By contrast, the *SNR* weighting consistently performs worse at high SNRs whereas it is comparable or even preferable over *flat* and *frob* at negative SNRs. At 150 km/h, the MLE is reduced by a factor of five with *MVCNR-frob* compared to *SRP-PHAT*.

In addition to Fig. 8, the upper and lower localization error quartiles are shown in comparison to the MLE for three different driving speeds in Fig. 9 (note the different scaling of the

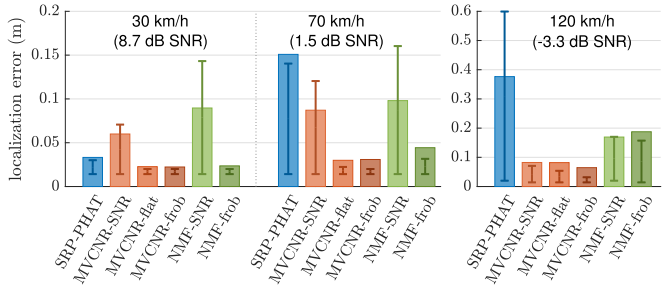


Fig. 9. Car setup: Mean localization error (wide bars) and upper and lower localization error quartiles at different driving speeds (different SNRs).

y-axis for 120 km/h). This plot enables a better differentiability of the performance especially at low speed (high SNR) and, furthermore, it allows to assess the fluctuation of the localization error. The *CSRP* method, which has the greatest overall MLE, is omitted for the purpose of better scaling. At 30 km/h, the MLE and the error quartiles of *SRP-PHAT* are comparable with those of *MVCNR* and *NMF* in combination with the *flat* and *frob* weighting. However, at higher speeds, *SRP-PHAT* is clearly worse than the proposed GSRP methods. In particular, the span of the error quartiles of *SRP-PHAT* is considerably higher which indicates a stronger fluctuation of the localization error. A comparison of the proposed frequency weightings of the GSRP methods shows that the *SNR* weighting generates a greater MLE and error quartile span at high SNRs, whereas it is comparable at 120 km/h with negative SNR.

2) HA Setup: In the second evaluated setup, the source DOA is estimated. This yields a one-dimensional SRP map over the azimuth θ, which we call DOA map to distinguish from the two-dimensional SRP map of the previous evaluation setup. We computed frame-wise DOA maps of 3 s speech snippets at 5 dB SNR for several source directions between $\theta_{\text{s}} = 90°$ and $0°$ with the conventional *SRP-PHAT* using a free far-field model *(FF)*, the modified *SRP-PHAT* method using the phase-transformed HRTF *HRTF-PHAT*, and the GSRP methods *MVCNR-frob* and *NMF-frob* using *HRTFs* as acoustic model. In Fig. 10, the frame-wise DOA maps of each source direction were averaged over time and stacked on top of each other, which generates a two-dimensional DOA heatmap over several source directions θ_{s}. The DOA map maximum of each source direction is marked by a black dot. The gray line indicates the actual source DOA (ground truth). While the *SRP-PHAT (FF)* map shows rather good results for frontal source DOAs between 45° and 0°, the lateral DOAs are mislocalized. By contrast, the *SRP-PHAT* using the *HRTF phase* is better able to resolve lateral source DOAs. However, the localization of DOAs between 80° and 90° is still inaccurate. The plotted GSRP maps show good localization performance over all evaluated DOAs as the SRP map maxima mostly coincide with the source DOAs (gray line). However, the peaks of *MVCNR-frob* and especially *NMF-frob* are broader compared to *SRP-PHAT* whereas the GSRP methods, in particular *MVCNR-frob*, suppress the SRP map regions apart from the main peak to a greater extent. This also reduces the front-back confusion compared to the *SRP-PHAT* maps.

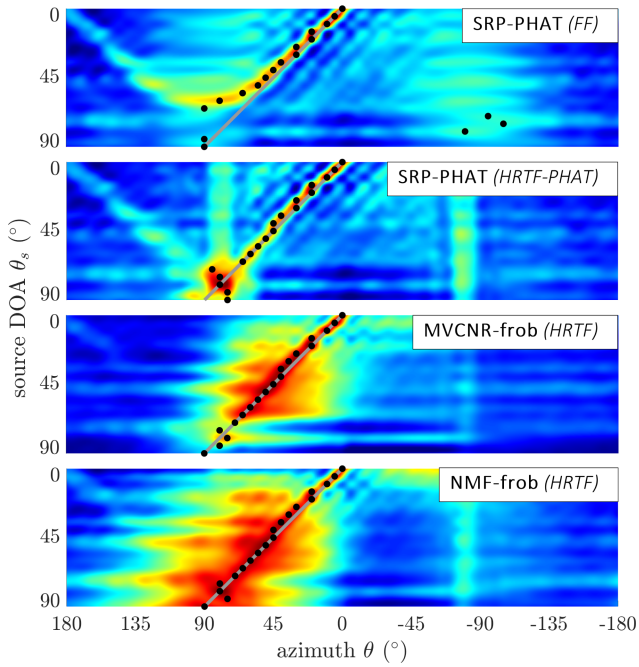


Fig. 10. HA setup: Time-averaged, stacked DOA maps (one-dimensional DOA maps of several source directions θ_s are stacked on top of each other) of 3 s speech snippets at 5 dB SNR. Black dots denote the SRP maximum of each evaluated source DOA θ_s. The acoustic model used in each SRP method is indicated in parentheses. (FF: free far-field model; HRTF: measured HRTFs; HRTF-PHAT: phase-transformed HRTF.)

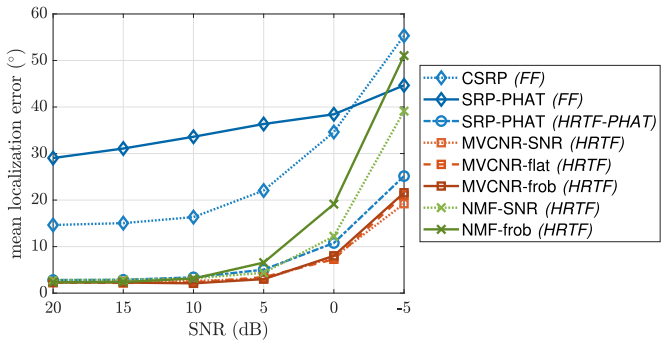
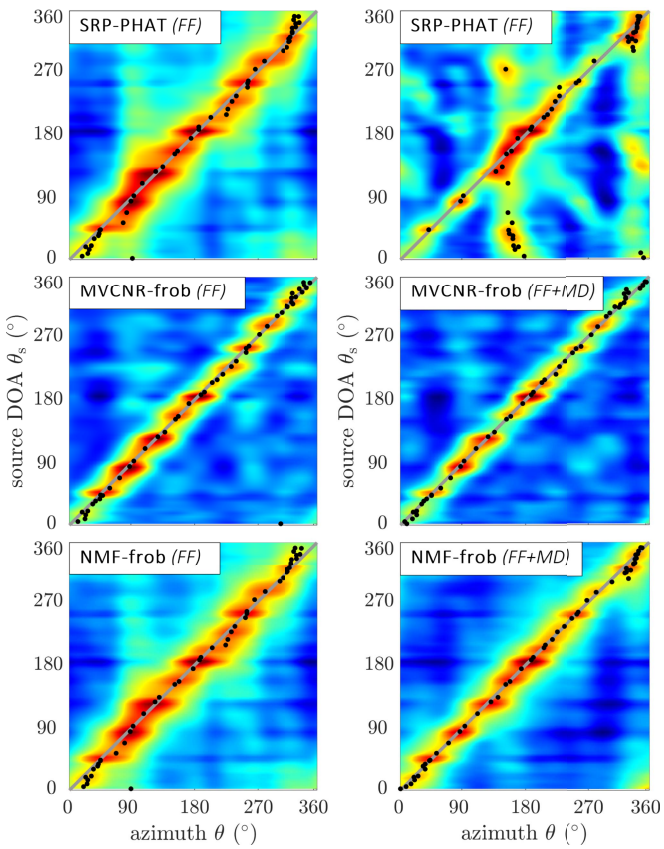


Fig. 11. HA setup: Mean localization error of all evaluated SRP methods over the SNR. The acoustic model used in each evaluated SRP method is indicated in parentheses.

Fig. 11 shows the MLE (in degree) of each evaluated SRP method, i.e., the average of the absolute angular deviation $E_{\text{ang}} = |\hat{\theta}_s - \theta_s|$ over all frames with active speech and all source DOAs. The baseline *SRP-PHAT* and *CSRP* with the conventional free field model *(FF)* have a significantly higher MLE than all other methods already at high SNRs. By contrast, the *SRP-PHAT* method involving the *HRTF phase* restores the localization capabilities and performs comparable to the proposed GSRP methods *MVCNR* and *NMF* at SNRs greater than 5 dB. At low SNR, the performance of *MVCNR* is clearly better compared to its simplified counterpart *NMF* which suffers from dominant noise. The *SRP-PHAT* with *HRTF phase* is slightly worse than *MVCNR* at low SNRs. In this setup, no significant differences can be seen between the proposed frequency weightings for

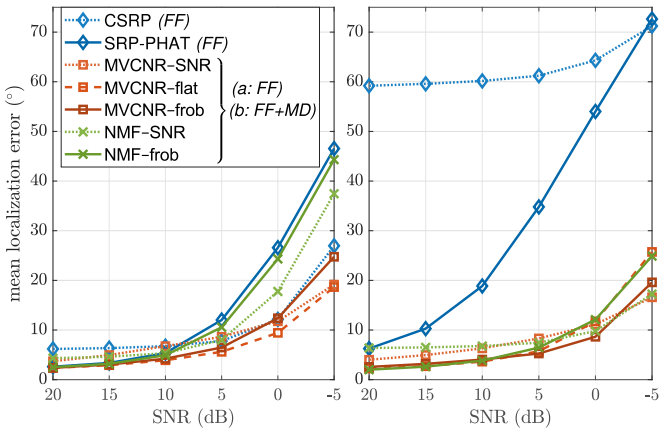


(a) Omnidirectional microphones. (b) Cardioid microphones.

Fig. 12. UCA setup: Time-averaged, stacked DOA maps (one-dimensional DOA maps of several source directions θ_s are stacked on top of each other) of 3 s speech snippets at 5 dB SNR. Black dots denote the SRP maximum of each evaluated source DOA θ_s. The acoustic model used in each SRP method is indicated in parentheses. (FF: free far-field model; FF+MD: free far-field model + cardioid microphone directivities.)

MVCNR. However, for the *NMF*, the *SNR* weighting is advantageous over the *frob* weighting at low SNRs. The proposed *MVCNR* reduces the MLE by 25° to 30° compared to *SRP-PHAT (FF)* at all SNRs.

3) UCA Setup: In the third evaluation setup, we computed the time-averaged DOA maps of 3 s speech snippets at 5 dB SNR for all simulated source positions of the conventional *SRP-PHAT* and the GSRP methods *MVCNR-frob* and *NMF-frob*. Fig. 12 shows the stacked DOA maps of all evaluated source directions of the *UCA(a) setup* with omnidirectional microphones (left) in comparison to the *UCA(b) setup* with directional microphones (right). The SRP maximum of each source DOA is marked by a black dot. The gray line indicates the actual source DOA (ground truth). In the *UCA(a) setup*, the SRP maps of all methods show comparably good results. It is noticeable that the *SRP-PHAT* and *NMF-frob* map look almost identical and show slight DOA deviations at certain angles (e.g., close to $\theta_s = 0°$ or $\theta_s = 90°$). The *MVCNR-frob* map shows good localization accuracy over all evaluated DOAs and regions apart from the source DOA are suppressed more consistently. In the *UCA(a) setup* involving cardioid microphones, the performance of *SRP-PHAT* decreases significantly. One can observe clear sidelobes in the DOA maps,



(a) Omnidirectional microphones. (b) Cardioid microphones.

Fig. 13. UCA setup: Mean localization error of all evaluated SRP methods over the SNR. The acoustic model used in each evaluated SRP method is indicated in parentheses.

e.g., around $\theta = 165°$ or $\theta = 350°$, which cause many DOA mislocalizations. By contrast, the *MVCNR-frob* and *NMF-frob* DOA maps of the *UCA(b) setup* show similar or even slightly better accuracy compared to those of the *UCA(a) setup*.

Fig. 13 shows the MLE (in degree) averaged over all source DOAs of each evaluated SRP method for the *UCA(a) setup* (left) and *UCA(b) setup* (right) over various SNRs. In the *UCA(a) setup* with omnidirectional microphones, all evaluated methods perform good at high SNR. In particular, *SRP-PHAT*, *MVCNR* with *flat* and *frob* weighting, and *NMF-frob* show identical high accuracy for SNR greater than 10 dB. Whereas no relevant differences are observable between *SRP-PHAT* and *NMF-frob* at any SNR, the *MVCNR* methods clearly outperform *SRP-PHAT* and the *NMF* methods at SNRs lower than 5 dB. In the *UCA(b) setup* involving directional microphones, remarkable differences in the performance can be observed between the conventional methods *CSRP* and *SRP-PHAT* and the proposed GSRP methods. The *CSRP* has extremely poor accuracy even at high SNR. The MLE of *SRP-PHAT* at 20 dB SNR is only slightly higher compared to the GSRP methods but increases significantly with lower SNR. The performance of the GSRP methods is comparable, where *MVCNR-frob* shows a slightly lower MLE for positive SNRs and the methods involving the *SNR* weighting are preferable at negative SNR. Compared to the *UCA(a) setup*, the performance of the *MVCNR* methods is similar whereas the *NMF* methods even perform better in the *UCA(b) setup* involving cardioid microphones.

E. Discussion of Results

The results show that the proposed generalized SRP methods *MVCNR* and *NMF* consistently outperform the conventional SRP baseline methods *SRP-PHAT* and *CSRP* in the presented evaluation setups. In the *car setup*, the relevant difference between the conventional SRP methods and the proposed GSRP methods is that the latter exploits observed level differences in addition to TDOAs (acoustic free near-field model) whereas the conventional SRP methods only use TDOA information (free far-field model). Especially at low SNRs, this exploitation of level information with GSRP significantly reduces the localization error in this setup. In the *HA setup*, the proposed GSRP methods also use both microphone level and phase information for localization. However, in this setup, the relevant improvement of GSRP over the conventional SRP, which uses the free far-field model, is the reduced TDOA mismatch between the acoustic model and the observed TDOAs by using measured HRTFs (or the HRTF phase) as acoustic model. These HRTFs implicitly incorporate the shadowing effect of the head which causes significant deviations from the conventional free-field propagation, especially at lateral DOAs. The *UCA setup* helps to further understand the benefit of the GSRP over the conventional SRP methods. The *UCA(a) setup* with an array of omnidirectional microphones is a typical use case of the conventional SRP. In this setup, the same acoustic model is applied to the GSRP methods as to the conventional SRP, which allows to asses the influence of the proposed GSRP frequency weightings and of incorporating the NCM estimate in the *MVCNR* method independently from the acoustic model. As can be expected, the performance of *SRP-PHAT* and *NMF-frob* is almost identical in this scenario since no relevant microphone level differences are observable (cf. Section IV-F3 for more details). By contrast, the *MVCNR* beamformer is able to clearly improve the localization performance at low SNR. The *UCA(b) setup* with an array of five directional microphones significantly impairs the localization performance of the conventional SRP method. These results indicate that the phase information are less reliable in this setup with directional microphones. A likely explanation for this is that phase differences of a microphone pair cannot be assessed if the source DOA is contrary to the orientation of one of the cardioid microphones. For other source DOAs, phase information might be available at high SNR but they are prawn to phase distortions at lower SNR. However, the results show that the GSRP method can compensate for these effects and restores or even slightly improves localization accuracy in the scenario involving directional microphones compared to the *UCA(a) setup*. This is because considering the microphone directivity patterns in the acoustic model of the GSRP methods allows to additionally exploit the source-DOA-dependent level information of the directional microphones for localization.

In all simulated scenarios, *MVCNR* and *NMF* have similar performance at high SNR while *MVCNR* outperforms *NMF* at lower SNR. This is intuitive since *NMF* is a simplification of *MVCNR* assuming spatially uncorrelated and homogeneous noise – and this assumption is increasingly violated with increasing noise which, in fact, is not perfectly spatially uncorrelated and homogeneous. Compared to *SRP-PHAT* and *CSRP*, the proposed GSRP methods not only have a lower MLE but also the fluctuation of the localization error is considerably lower especially at low SNR. This is because the GSRP methods are less prone to TDOA estimation errors as they can additionally exploit level information which might indicate the (coarse) source position even with a noisy TDOA estimate. Furthermore, the results compare the proposed GSRP frequency weightings *SNR*, *flat* and *frob*. The *SNR* weighting scales each frequency depending on its respective narrowband SNR, which might cause

a few single frequency bands with highest SNR to dominate the SRP result. This property seems beneficial in the high-noise case with an SNR close to 0 dB or lower, whereas the *flat* and the simpler *frob* weighting, which equalize the spectral contribution of each frequency, have a better performance at positive SNRs.

VI. Conclusion

In this paper, we have presented a generalization of the conventional SRP method that allows to exploit generic acoustic models and noise characteristics for acoustic sound source localization: the generalized steered response power (GSRP) method. By using an appropriate acoustic model (including ATF or RTF measurements) for a given microphone setup, the proposed approach can jointly exploit the observed microphone level and phase differences to improve the localization performance compared to the conventional SRP which only assesses phase differences. It has been shown that simply replacing the acoustic free far-field model of the conventional SRP method by other acoustic models is not optimal with known SRP beamformer methods, such as delay-and-sum or MVDR. To this end, we propose a novel SRP beamforming design for localization using generic acoustic transfer functions and noise covariance matrices. Based on this GSRP beamforming design, we have derived the *minimum variance constant noise response (MVCNR)* beamformer and its simplification for spatially uncorrelated and homogeneous noise – the *normalized matched filter (NMF)*. Furthermore, different frequency weightings for the presented beamformers have been proposed and analyzed. These frequency weightings are suitable alternatives of the commonly used PHAT weighting for SRP as they, unlike PHAT, preserve the inter-microphone level differences that can be exploited by the GSRP method. Realistic simulations of three different scenarios involving distributed microphones, hearing aid microphone arrays, and arrays with directional microphones under various noise conditions have shown that the mean localization error can be significantly reduced with the proposed methods compared to the conventional SRP. In particular, MVCNR consistently performs well in all evaluated scenarios, whereas NMF suffers under highly noisy conditions. The proposed SRP generalization is especially beneficial for setups where the conventional free far-field assumptions are violated to a greater extent and, moreover, the microphone level differences may contain relevant source location cues. For instance, this is the case in setups with distributed microphones in the near field of the source, or setups involving acoustically shadowed or directional microphones. However, also in a typical scenario of the conventional SRP with a speaker in the far field of a compact circular microphone array, the proposed GSRP methods were able to outperform the SRP-PHAT baseline.

Appendix

Derivation of $\alpha(\omega, \mathbf{p})$ of (22) based on the GSRP beamformer design criterion No 1 (20): The matrix $\mathbf{A}$ in (21) is a Hermitian, positive-definite matrix and therefore can be decomposed with the Cholesky decomposition into $\mathbf{A} = \mathbf{L}\,\mathbf{L}^{\mathrm{H}}$, where $\mathbf{L}$ is a triangular matrix. With this, (21) becomes

$$\left|\alpha(\omega, \mathbf{p}_{\mathrm{s}})\mathbf{h}_{\mathrm{s}}^{\mathrm{H}}(\omega)\mathbf{L}\mathbf{L}^{\mathrm{H}}\mathbf{h}_{\mathrm{s}}(\omega)\right| \geq \left|\alpha(\omega, \mathbf{p})\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\mathbf{L}\mathbf{L}^{\mathrm{H}}\mathbf{h}_{\mathrm{s}}(\omega)\right|. \tag{44}$$

Substituting $\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}}) = \mathbf{L}^{\mathrm{H}}\,\mathbf{h}_{\mathrm{s}}(\omega)$ and $\mathbf{h}'(\omega, \mathbf{p}) = \mathbf{L}^{\mathrm{H}}\,\mathbf{h}(\omega, \mathbf{p})$ into (44) yields

$$\left|\alpha(\omega, \mathbf{p}_{\mathrm{s}})\underbrace{\mathbf{h}'^{\mathrm{H}}(\omega, \mathbf{p}_{\mathrm{s}})\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})}_{\|\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\|^2}\right| \geq \left|\alpha(\omega, \mathbf{p})\mathbf{h}'^{\mathrm{H}}(\omega, \mathbf{p})\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\right|. \tag{45}$$

Now, we search for an $\alpha(\omega, \mathbf{p})$ and $\alpha(\omega, \mathbf{p}_{\mathrm{s}})$, respectively, which ensure that this inequality is true for all $\mathbf{p}$ and $\mathbf{p}_{\mathrm{s}}$. To this end, we can use the Hermitian angle between two complex column vectors $\mathbf{a}$ and $\mathbf{b}$, i.e., $\cos_{\mathrm{H}}(\mathbf{a}, \mathbf{b}) = |\mathbf{a}^{\mathrm{H}}\,\mathbf{b}|/(\|\mathbf{a}\|\,\|\mathbf{b}\|)$ with $\cos_{\mathrm{H}}(\mathbf{a}, \mathbf{b}) \in [0, 1]$, to rewrite (45) as

$$\begin{aligned}&\left|\alpha(\omega, \mathbf{p}_{\mathrm{s}})\right| \|\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\|^2 \\ &\geq \left|\alpha(\omega, \mathbf{p})\right| \cos_{\mathrm{H}}\left(\mathbf{h}'(\omega, \mathbf{p}), \mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\right) \|\mathbf{h}'(\omega, \mathbf{p})\| \|\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\|.\end{aligned} \tag{46}$$

When dividing both sides of (46) by $\|\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\|$, we can see that the inequality holds for

$$\alpha(\omega, \mathbf{p}) = \frac{\zeta(\omega)}{\|\mathbf{h}'(\omega, \mathbf{p})\|}, \quad \text{and } \alpha(\omega, \mathbf{p}_{\mathrm{s}}) = \frac{\zeta(\omega)}{\|\mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\|}, \tag{47}$$

respectively, because (46) reduces with (47) to

$$1 \geq \cos_{\mathrm{H}}\left(\mathbf{h}'(\omega, \mathbf{p}), \mathbf{h}'(\omega, \mathbf{p}_{\mathrm{s}})\right), \tag{48}$$

which is true for all $\mathbf{p}$. In (47), $\zeta(\omega)$ is a positive, real-valued scalar. Finally, when substituting $\mathbf{h}'(\omega, \mathbf{p}) = \mathbf{L}^{\mathrm{H}}\,\mathbf{h}(\omega, \mathbf{p})$, the found solution for $\alpha(\omega, \mathbf{p})$ in (47) becomes

$$\alpha(\omega, \mathbf{p}) = \frac{\zeta(\omega)}{\left\|\mathbf{L}^{\mathrm{H}}\,\mathbf{h}(\omega, \mathbf{p})\right\|} = \frac{\zeta(\omega)}{\sqrt{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\,\mathbf{L}\,\mathbf{L}^{\mathrm{H}}\,\mathbf{h}(\omega, \mathbf{p})}}, \tag{49}$$

which can be rewritten by re-composing $\mathbf{L}\,\mathbf{L}^{\mathrm{H}} = \mathbf{A}$ as

$$\alpha(\omega, \mathbf{p}) = \frac{\zeta(\omega)}{\sqrt{\mathbf{h}^{\mathrm{H}}(\omega, \mathbf{p})\,\mathbf{A}\,\mathbf{h}(\omega, \mathbf{p})}}. \tag{50}$$

References

[1] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1994, pp. II/273–II/276.

[2] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, May 1997.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Digital Signal Processing*. Berlin, Germany: Springer, 2001, ch. 8, pp. 157–180.

[4] N. Madhu and R. Martin, "Acoustic Source Localization With Microphone Arrays," in *Advances in Digital Speech Transmission*, Hoboken, NJ, USA: Wiley, Jan. 2008, ch. 6, pp. 135–170.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. TASSP-24, no. 4, pp. 320–327, Aug. 1976.

[6] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 1997, pp. 375–378.

[7] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 1, May 2006, Art. no. 026503.

[8] E. Grinstein et al., "Steered response power for sound source localization: A tutorial review," *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 1, Nov. 2024, Art. no. 59.

[9] S. Gannot, M. Haardt, W. Kellermann, and P. Willett, "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 3–7, Mar. 2019.

[10] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 4, Mar. 2003, Art. no. 860465.

[11] A. Brutti, M. Omologo, and P. Svaizer, "Speaker localization based on oriented global coherence field," in *Proc. INTERSPEECH 2006 - ICSLP, 9th Int. Conf. Spoken Lang. Process.*. Pittsburgh, PA, USA: ISCA, Sep. 2006, pp. 2606–2609.

[12] A. Plinge and G. A. Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 614–618.

[13] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Combined source tracking and noise reduction for application in hearing aids," in *Proc. ITG Conf. Voice Commun.*, 2008, pp. 1–4.

[14] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[15] F. Hummes, J. Qi, and T. Fingscheidt, "Robust acoustic speaker localization with distributed microphones," in *Proc. 19th Eur. Signal Process. Conf.*, 2011, pp. 240–244.

[16] A. Marti, M. Cobos, and J. J. Lopez, "A real-time sound source localization and enhancement system using distributed microphones," *Proc. Audio Eng. Soc. Conv.*, vol. 130, 2011, pp. 152–159.

[17] Y. Huang, J. Tong, X. Hu, and M. Bao, "A robust steered response power localization method for wireless acoustic sensor networks in an outdoor environment," *Sensors*, vol. 21, no. 5, Feb. 2021, Art. no. 1591.

[18] B. Çakmak, T. Dietzen, R. Ali, P. Naylor, and T. Van Waterschoot, "A distributed steered response power approach to source localization in wireless acoustic sensor networks," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2022, pp. 1–5.

[19] B. Çakmak, T. Dietzen, R. Ali, P. Naylor, and T. V. Waterschoot, "Microphone pair selection for sound source localization in massive arrays of spatially distributed microphones," in *Proc. 32nd Eur. Signal Process. Conf.*, Aug. 2024, pp. 251–255.

[20] H. He, X. Wang, Y. Zhou, and T. Yang, "A steered response power approach with trade-off prewhitening for acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 143, no. 2, pp. 1003–1007, Feb. 2018.

[21] Z. Wang, H. He, J. Chen, J. Benesty, and Y. Yu, "A steered response power approach with bilinear prediction-based trade-off prewhitening for speaker localization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2024, pp. 1046–1050.

[22] K. Brümann and S. Doclo, "Steered response power-based direction-of-arrival estimation exploiting an auxiliary microphone," in *Proc. 32nd Eur. Signal Process. Conf.*, Aug. 2024, pp. 917–921.

[23] X. Dang and H. Zhu, "An iteratively reweighted steered response power approach to multisource localization using a distributed microphone network," *J. Acoust. Soc. Amer.*, vol. 155, no. 2, pp. 1182–1197, Feb. 2024.

[24] W.-T. Lai, L. Birnie, X. Chen, A. Bastine, T. D. Abhayapala, and P. N. Samarasinghe, "Source localization by multidimensional steered response power mapping with sparse bayesian learning," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2024, pp. 31–35.

[25] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 3517–3531, 2024.

[26] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 499–508, Sep. 2004.

[27] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[28] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, Oct. 2007, pp. 295–298.

[29] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

[30] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on Nyquist-Shannon sampling," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2021, pp. 206–210.

[31] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[32] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognit. Lett.*, vol. 84, pp. 15–21, Dec. 2016.

[33] D. Salvati, C. Drioli, and G. L. Foresti, "On the use of machine learning in microphone array beamforming for far-field sound source localization," in *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process.*, Sep. 2016, pp. 1–6.

[34] P. Pertila and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 6125–6129.

[35] E. Grinstein, M. Brookes, and P. A. Naylor, "Graph neural networks for sound source localization on distributed microphone networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2023, pp. 1–5.

[36] E. Grinstein, C. M. Hicks, T. v. Waterschoot, M. Brookes, and P. A. Naylor, "The neural-SRP method for universal robust multi-source tracking," *IEEE Open J. Signal Process.*, vol. 5, pp. 19–28, 2024.

[37] H. Van Trees, "Optimum waveform estimation," in *Optimum Array Processing*, H. L. V. Trees, Ed., New York, NY, USA: Wiley, 2002, ch. 6, pp. 428–709.

[38] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Signal Processing in Modern Communication*. Berlin, Germany: Springer, 2010, ch. 12, pp. 307–337.

[39] S. Bedard, B. Champagne, and A. Stephenne, "Effects of room reverberation on time-delay estimation performance," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1994, pp. II/261–II/264.

[40] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, Mar. 2007.

[41] H. R. Abutalebi and H. Momenzadeh, "Performance improvement of tdoa-based speaker localization in joint noisy and reverberant conditions," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, Jan. 2011, Art. no. 621390.

[42] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2015, pp. 1–5.

[43] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 2565–2568.

[44] H. Kuttruff, *Room Acoustics*. Boca Raton, FL, USA: CRC Press, Oct. 2016.

[45] 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.

[46] E. Jan, P. Svaizer, and J. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. ISCAS'95 - Int. Symp. Circuits Syst.*, 1995, pp. 1460–1463.

[47] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. IEEE 26th Conv. Elect. Electron. Engineers Isr.*, Nov. 2010, pp. 416–420.

[48] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, Apr. 2009.

[49] G. Garcia-Barrios, J. M. Gutiérrez-Arriola, N. Sáenz-Lechón, V. J. Osma-Ruiz, and R. Fraile, "Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps," *IEEE Access*, vol. 9, pp. 121549–121560, 2021.

[50] S. Graetzer et al., "Dataset of British english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data Brief*, vol. 41, Apr. 2022, Art. no. 107951.

[51] K. Müller and T. Wolff, "In-car McVAMPIRE–in-car multichannel varying mouth position impulse response dataset," 2024. [Online]. Available: https://zenodo.org/records/12806684

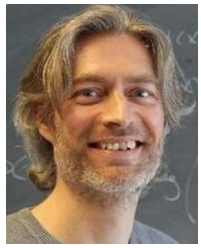
[52] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, Jul. 2009, Art. no. 298605. [Online]. Available: https://medi.uni-oldenburg.de/hrir/

[53] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.
[54] E. Habets, "RIR-generator," 2020. [Online]. Available: https://zenodo.org/records/4117640
[55] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[56] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*. Berlin, Germany: Springer, 2001, pp. 61–85.
[57] K. Müller, B. Çakmak, P. Didier, S. Doclo, J. Østergaard, and T. Wolff, "Head orientation estimation with distributed microphones using speech radiation patterns," in *Proc. 57th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2023, pp. 1404–1409.
[58] S. D. Bellows, C. M. Pincock, J. K. Whiting, and T. W. Leishman, "Average speech directivity," Brigham Young Univ. Scholars Archive, Directiv., Tech. Rep. 1, 2019. [Online]. Available: https://scholarsarchive.byu.edu/directivity/1
[59] T. W. Leishman, S. D. Bellows, C. M. Pincock, and J. K. Whiting, "High-resolution spherical directivity of live speech from a multiple-capture transfer function method," *J. Acoust. Soc. Amer.*, vol. 149, no. 3, pp. 1507–1523, Mar. 2021.

Kaspar Müller received the Dipl.-Ing. degree in electrical and audio engineering from the Technical University of Graz and University of Music and Performing Arts Graz, Graz, Austria, in 2020. He is currently working toward the Dr.-Ing. degree with Cerence AI, Ulm, Germany. From 2020 to 2021, he was a Research Project Member with the Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz. He is currently a Research Scientist with Cerence AI. His research interests include audio signal processing with a focus on microphone array processing, acoustic source localization, and speech enhancement. He was the recipient of the Student Award 2021 of the German Acoustical Society (DEGA) for his Master's thesis.

Markus Buck (Member, IEEE) received the Dipl.-Ing. degree in electrical engineering and the Ph.D. degree from Ulm University, Ulm, Germany, in 1998 and 2004, respectively. From 1998 to 2009, he was with Temic Speech Dialog Systems and Harman/Becker Automotive Systems, Ulm, working on various topics in the field of speech processing. From 2009 to 2019, he was a Research Manager with Nuance Communications, Ulm, leading the technology development in acoustic speech enhancement for hands-free telephony, speech recognition, and in-car communication. Since 2019, he has been with Cerence AI, Ulm. His main research interests include multi-channel signal processing, adaptive filtering and neural network based methods for speech signal processing.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from KU Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with KU Leuven and McMaster University, Canada. From 2007 to 2009, he was the Principal Scientist with NXP Semiconductors, Leuven. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and Scientific Advisor for the Branch Hearing, Speech and Audio Technology HSA of the Fraunhofer Institute for Digital Media Technology IDMT. His research interests include signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo was the recipient of several Best Paper Awards (International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019). He was a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing and is a member of the EAA Technical Committee on Audio Signal Processing. Since 2021, he has been the Senior Area Editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.



Jan Østergaard (Senior Member, IEEE) received the M.Sc. degree from Aalborg University, Aalborg, Denmark, in 1999, and the Ph.D. degree (with cum laude) from the Delft University of Technology, Delft, The Netherlands, in 2007. From 1999 to 2002, he was an R&D Engineer with ETI A/S, Aalborg, and from 2002 to 2003, he was an R&D Engineer with ETI Inc., VA, USA. Between 2007 and 2008, he was a Postdoctoral Researcher with The University of Newcastle, Newcastle, NSW, Australia. He is currently a Full Professor of information theory and signal processing, Head of the Section on AI and Sound, and Head of the Centre on Acoustic Signal Processing Research (CASPR) with Aalborg University. His research interests include in the areas of statistical signal processing, information theory, and acoustic signal processing. He was the recipient of the Danish Independent Research Council's Young Researcher's Award, Best PhD Thesis Award by the European Association for Signal Processing (EURASIP), and fellowships from the Danish Independent Research Council and the Villum Foundations Young Investigator Programme. He is an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY.

Tobias Wolff received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and communications from Signal Processing Group, Technical University of Darmstadt, Darmstadt, Germany, in 2006 and 2011, respectively. In 2005 and 2007, he was a Visiting Researcher with Image Processing Laboratory, University of California Santa Barbara, Santa Barbara, CA, USA, working on subjective perception of video coding artifacts. In 2009, he joined the Department of Speech Signal Enhancement, Nuance Communications Deutschland GmbH, Ulm, Germany. Since 2017, he has been the Principal Researcher with Nuance in the area of multimicrophone acoustic speech enhancement. Since 2019, he has been with Cerence AI in the area of multimicrophone acoustic speech enhancement. His main scientific research interests include beamforming, source separation, and acoustic localization of sound sources.