Robust Source Localization and Enhancement With a Probabilistic Steered Response Power Model

Johannes Traa* *Student Member, IEEE*, David Wingate, Noah D. Stein, Paris Smaragdis *Fellow, IEEE*

***Abstract*—Source localization and enhancement are often treated separately in the array processing literature. One can apply Steered Response Power (SRP) localization to determine the sources' Directions-Of-Arrival (DOA) followed by beamforming and Wiener post-filtering to isolate the sources from each other and ambient interference. We show that when there is significant overlap between directional sources of interest in the Time-Frequency (TF) plane, traditional SRP localization breaks down. This may occur, for example, when the array is located near a reflector, significant early reflections are present, or the sources are harmonized. We propose a joint solution to the localization and enhancement problems via a probabilistic interpretation of the SRP function. We formulate optimization procedures for (1) a mixture of single-source SRP distributions (MoSRP) and (2) a multi-source SRP distribution (MultSRP). Unlike in traditional localization, the latter approach explicitly models source overlap in the TF plane. Results shows that the MultSRP model is capable of localizing sources with significant overlap in the TF domain and that either of the proposed methods out-performs standard SRP localization for multiple speakers.**

***Index Terms*—steered response power, source localization, beamforming, blind source separation**

I. Introduction

The array processing literature contains methods [1], [2] for enhancing directional signals in the presence of interferers and noise. Beamforming [3] aims to optimize a spatial filter that isolates a target signal's energy. The simplest example is the Delay-and-Sum (DS) beamformer, which is optimal for a single source in diffuse, additive, white Gaussian noise. This is a data-independent method since the true noise characteristics don't enter into the design of the filter. For general Gaussian noise, the optimal solution is the Minimum-Variance Distortionless Response (MVDR) beamformer. Finally, when several targets and/or interferers are present simultaneously, the Linearly-Constrained Minimum-Variance (LCMV) beamformer [4] provides a means of implementing both distortionless and null constraints. The former protect the target signal while the latter block undesired signals. This assumes that each source's Direction-Of-Arrival (DOA) is known. So, in practice, beamforming is preceded by a localization step.

Beamforming is a linear processor and so has limited source separation capabilities. Combining a beamformer with Wiener post-filtering [5] has been shown to provide adaptive noise reduction for non-stationary signals. However, the source DOAs are still assumed to be known. Some blind beamforming methods ignore the array structure and source DOAs altogether. The Joint Approximate Diagonalization of Eigen-matrices (JADE) algorithm was applied to recover the mixing matrix that describes the channel between sources and sensors [6].

There are many methods for localizing (and tracking) directional sources [7], [8], [9], [10], [11]. Steered Response Power (SRP) localization [12] involves computing the output power of a beamformer steered towards each DOA of interest and locating one or more peaks in the resulting SRP function. A similar approach computes a Generalized Cross-Correlation (GCC) [13] function over time delays. Spectral weighting such as the Phase Transform (PHAT) were explored in [13] for enhancing the SPR and GCC functions. The authors in [12], [14] used the GCC-PHAT approach to efficiently localize speech sources on a hemisphere. A third method centers around the eigenanalysis of the channel correlation matrix. The Multiple Signal Classification (MUSIC) algorithm [15] identifies signal and noise subspaces to form a "pseudo-spectrum" that contains peaks at the source DOAs. This requires a scan over DOA space. The root-MUSIC algorithm [16] avoids this by reducing the localization problem to one of root-finding. A related method, ESPRIT [17], involves a similar analysis, but takes advantage of arrays with a special structure. Several authors have extended these eigen-analysis methods to handle wideband sources [18].

Several authors have described the relationship between the GCC and SRP functions and a probabilistic model of the observed signals [12], [19], [20]. Similarly, SRP localization has been described as a Maximum-Likelihood (ML) problem [21], [22]. In [23], the authors model the observed frequency-domain data vectors as zero-mean Gaussian random variables and use an EM algorithm to learn the covariance parameters of the sources and apply multichannel Wiener filtering to perform source separation. The authors in [24] formulate a DOA-dependent covariance matrix to localize a single source in noisy conditions.

Time-frequency (TF) masking [25] is known to outperform linear filtering (e.g. beamforming) methods for enhancing wideband sources. This approach was originally motivated by the disjointness of speech over the TF plane [26]. The trade-off is the introduction of musical noise artifacts in the enhanced/separated signals. Despite this, much work has been done on source localization and separation using TF features [27], [28]. In [29], the authors localize the sources with

Manuscript received ... ..., 2015.

J. Traa is a PhD student in the ECE department at the University of Illinois at Urbana-Champaign (UIUC) (traa2@illinois.edu).

D. Wingate and N. Stein are researchers at Lyric Labs, Analog Devices (David.Wingate@analog.com, Noah.Stein@analog.com).

P. Smaragdis holds a joint faculty position in ECE and CS at UIUC and works with Adobe Systems, Inc. (paris@illinois.edu)

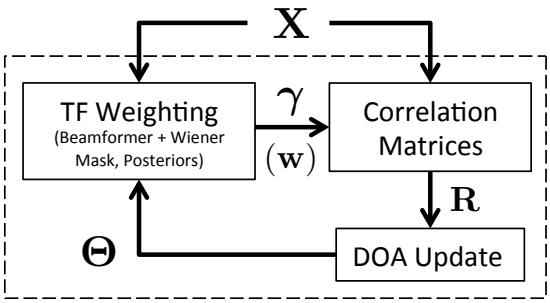

Fig. 1: Proposed iterative approach. A probabilistic SRP model is combined with time-frequency masking to perform blind source localization and separation in the presence of non-directional interference. Observed data $\mathbf{X}$ collected with microphones is used to optimize source DOAs $\boldsymbol{\Theta}$ using channel correlation information $\mathbf{R}$, Wiener mask weights $\boldsymbol{\gamma}$, and possibly source-specific TF weights $\mathbf{w}$.

visual cues and separate them by applying data-independent beamforming followed by TF masking. And in [30], a TF masking approach is proposed for tracking acoustic sources with a Rao-Blackwellized particle filter.

In this paper, we interpret the SRP function as a likelihood and propose two methods for maximizing it as a function of the source DOAs. One uses a mixture of single-source SRPs and the other uses an SRP that explicitly models the presence of multiple sources. We show that the latter is robust to situations where the sources have a significant amount of overlap in the TF plane. This occurs when, for example, early reflections are present [31] (e.g. the array is located near a reflective surface) or the sources are harmonized as when instrumentalists play in unison. We apply TF masking [25], [27] to isolate TF bins that correspond to the directional signals of interest. In this way, the localization, separation, and Wiener post-filtering steps are merged into one. The flow diagram for the proposed approach is shown in Fig. 1. This work is motivated in part by unified approaches that have been shown to out-perform sequential approaches for automatic speech recognition [32]. Indeed, we show that the proposed method outperforms a standard sequential localization approach.

The contributions of this paper are:

- A discussion of the probabilistic interpretation of the Steered Response Power (SRP) function.
- An SRP model that explicitly considers the presence of multiple simultaneous sources.
- A joint, iterative method for source localization and enhancement/separation applied to a mixture of SRPs model and the proposed multi-source model.
- Experiments demonstrating the benefits of the proposed approach for source localization and separation.

II. Data Model

Consider the convolutive time-domain model of a single source recorded under noisy conditions at M channels:

$$x_m[t] = a_m[t] * s[t] + n_m[t] \quad , \tag{1}$$

where $x_m[t]$ is the recorded sample at the m^{th} channel, $a_m[t]$ is the Room Impulse Response (RIR) between the source and m^{th} channel, $n_m[t]$ is a Gaussian noise process, and $*$ denotes convolution. We apply the Short-Time Fourier Transform (STFT) to the mixed, noisy signal x_m to de-couple the signal components across frequency and time. Thus, at time frame $t \in [1, T]$ and frequency $f \in [1, F]$, we have:

$$\mathbf{x}_{ft} = \mathbf{a}_f \, s_{ft} + \mathbf{n}_{ft} \quad , \tag{2}$$

where $\mathbf{x}_{ft} \in \mathbb{C}^M$ is an observed data vector, $\mathbf{a}_f \in \mathbb{C}^M$ is the mixing vector, $s_{ft} \in \mathbb{C}$ is the source coefficient, and $\mathbf{n}_{ft} \in \mathbb{C}^M$ contains the noise coefficients. For simplicity, we assume that $\mathbf{n}_{ft} \sim \mathcal{N}\left(\mathbf{0}, \sigma_f^2 \mathbf{I}\right)$ and that the source and noise coefficients are statistically independent.

For a point source in an anechoic environment, we write $\mathbf{a}_f$ explicitly in terms of the source DOA as the unit steering vector:

$$\mathbf{a}_f(\boldsymbol{\phi}) = \frac{1}{\sqrt{M}} \exp\left(j \frac{2\pi l_f}{c} \mathbf{m}^\top \boldsymbol{\phi}\right) \quad , \tag{3}$$

where $\boldsymbol{\phi} \in \mathbb{R}^3$ is the source's unit DOA vector, $\mathbf{m} \in \mathbb{R}^{3 \times M}$ is the matrix of M sensor positions, l_f is the center frequency of the f^{th} band, and c is the speed of sound.

When $K > 1$ sources are present, we can write:

$$\mathbf{x}_{ft} = \mathbf{A}_f(\boldsymbol{\Phi}) \, \mathbf{s}_{ft} + \mathbf{n}_{ft} \quad , \tag{4}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_{1:K}\}$ denotes the set of source DOAs, $\mathbf{s}_{ft} \in \mathbb{C}^K$ is a vector of source coefficients, and we have defined the steering matrix:

$$\mathbf{A}_f(\boldsymbol{\Phi}) = \begin{bmatrix} \mathbf{a}_f(\boldsymbol{\phi}_1) & \cdots & \mathbf{a}_f(\boldsymbol{\phi}_K) \end{bmatrix} \in \mathbb{C}^{M \times K} \quad . \tag{5}$$

III. Classical array processing

In this section, we describe how the propagation models in (2) and (4) are used for narrowband beamforming and localization. We can easily extend this to the broadband case by assuming that all frequency bands are mutually independent.

A. Beamforming

Linear spatial filters are often used for enhancing directional signals. They can be described by a weight vector $\mathbf{w}_f \in \mathbb{C}^M$ that is used to estimate a source coefficient s_{ft} via $\widehat{s}_{ft} = \mathbf{w}_f^H \mathbf{x}_{ft}$. The optimal $\mathbf{w}_f$ minimizes the expected output power of the filter:

$$P = \mathrm{E}\left[\, |\widehat{s}_{ft}|^2 \,\right] = \mathbf{w}_f^H \, \mathbf{R}_f \, \mathbf{w}_f \quad , \tag{6}$$

without distorting the desired signal at DOA $\boldsymbol{\phi}$:

$$\widehat{\mathbf{w}}_f = \underset{\mathbf{w}_f}{\operatorname{argmin}} \; \mathbf{w}_f^H \mathbf{R}_f \, \mathbf{w}_f \quad , \tag{7}$$

$$\text{s.t.} \quad \mathbf{a}_f^H(\boldsymbol{\phi}) \, \mathbf{w}_f = 1 \quad . \tag{8}$$

where we have defined the channel correlation matrix:

$$\mathbf{R}_f = \mathrm{E}\left[\mathbf{x}_{ft} \, \mathbf{x}_{ft}^H\right] \quad . \tag{9}$$

The solution is the well-known MVDR beamformer:

$$\widehat{\mathbf{w}}_f^{MVDR} = \frac{\mathbf{R}_f^{-1}\mathbf{a}_f(\boldsymbol{\phi})}{\mathbf{a}_f^H(\boldsymbol{\phi})\,\mathbf{R}_f^{-1}\mathbf{a}_f(\boldsymbol{\phi})} \quad . \tag{10}$$

If we assume (in the derivation) that $\mathbf{R}_f = \mathbf{I}$, this reduces to the data-independent Delay-and-Sum (DS) beamformer:

$$\widehat{\mathbf{w}}_f^{DS} = \mathbf{a}_f(\boldsymbol{\phi}) \quad , \tag{11}$$

with corresponding empirical output power:

$$P_{DS} = \mathbf{a}_f^H(\boldsymbol{\phi})\,\widehat{\mathbf{R}}_f\,\mathbf{a}_f(\boldsymbol{\phi}) \quad , \tag{12}$$

where:

$$\widehat{\mathbf{R}}_f = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{ft}\,\mathbf{x}_{ft}^H \quad , \tag{13}$$

is an empirical approximation of (9). The DS beamformer can be pre-computed, while the MVDR beamformer provides better noise suppression when $\mathbf{R}_f \neq \mathbf{I}$. They are generalized to the multi-source case via the multiply-constrained optimization problem:

$$\widehat{\mathbf{w}}_f = \underset{\mathbf{w}_f}{\operatorname{argmin}}\ \mathbf{w}_f^H\mathbf{R}_f\,\mathbf{w}_f \quad , \tag{14}$$

$$\text{s.t.} \quad \mathbf{A}_f^H(\boldsymbol{\Phi})\,\mathbf{w}_f = \mathbf{u} \quad , \tag{15}$$

where $\mathbf{u} \in \mathbb{C}^K$ contains the desired gains for all DOAs. The solution is sometimes refered to as the Linearly-Constrained Minimum-Variance (LCMV) beamformer [4]:

$$\widehat{\mathbf{w}}_f^{LCMV} = \mathbf{R}_f^{-1}\mathbf{A}_f(\boldsymbol{\Phi})\left[\mathbf{A}_f^H(\boldsymbol{\Phi})\,\mathbf{R}_f^{-1}\mathbf{A}_f(\boldsymbol{\Phi})\right]^{-1}\mathbf{u} \quad . \tag{16}$$

B. Steered Response Power (SRP) Localization

Beamformers can be used if the source DOA(s) $\boldsymbol{\Phi}$ are known. SRP localization [12] aims to identify $\boldsymbol{\Phi}$ by searching for peaks in the output power of a single-source beamformer at hypothesized DOAs $\boldsymbol{\theta}$. We can see this for the DS beamformer and one source by writing (12) as:

$$P_{DS}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}|\,\mathbf{a}_f^H(\boldsymbol{\theta})\,\mathbf{x}_{ft}|^2 \tag{17}$$

$$= \frac{1}{T}\sum_{t=1}^{T}|\,\mathbf{a}_f^H(\boldsymbol{\theta})\,\mathbf{a}_f(\boldsymbol{\phi})\,|^2\,|s_{ft}|^2 + C(f,T) \tag{18}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}|s_{ft}|^2 + C(f,T) \quad , \tag{19}$$

where $C(f,T) \xrightarrow{T\to\infty} \sigma_f^2$ is due to the additive noise term and equality holds when $\boldsymbol{\theta} = \boldsymbol{\phi}$. The Phase Transform (PHAT) [13] involves setting all the magnitudes of the components of $\mathbf{x}_{ft}$ to 1 and helps to emphasize the peaks in $P_{DS}(\boldsymbol{\theta})$.

When multiple sources are present, one might search for multiple peaks in the SRP [12]. There are three issues with this approach. First, this scan over DOA space may be infeasible when computation power is limited or we desire a high spatial resolution. Efficient strategies for searching for peaks in the SRP function have been proposed to work around this issue [33], [34]. Second, we expect the peaks to be poorly localized at low frequencies and for closely-spaced microphones because the steering vectors are hard to distinguish under those conditions. And third, if the source coefficients are simultaneously large in magnitude, the SRP function is distorted by cross-terms.

A more effective approach scans over all DOA sets $\boldsymbol{\Phi}$ using an LCMV beamformer and locates the peak output power. However, this leads to a combinatorial problem. If we discretize the DOA search space into D look directions, we must scan over D^K $\boldsymbol{\Phi}$'s. We remedy this by modeling the multi-source SRP function as a continuous likelihood function parametrized by $\boldsymbol{\Phi}$ and maximizing it with gradient ascent.

IV. Probabilistic SRP Model

We present a probabilistic model [12] for the observed data vectors $\mathbf{x}_{ft}$ and relate this to the SRP function described in Section III-B. A more detailed account of this probabilistic formulation can be found in Appendix A.

A. SRP Likelihood

The propagation model in (4) corresponds to a Gaussian likelihood for the observed data vectors:

$$\log\mathcal{L}_{ft}(\boldsymbol{\Theta}) = \log\mathcal{N}\left(\mathbf{x}_{ft}\,;\,\boldsymbol{\mu}_{ft}\,,\,\sigma_f^2\,\mathbf{I}\right) \quad . \tag{20}$$

The mean $\boldsymbol{\mu}_{ft}$ encodes the expected value of $\mathbf{x}_{ft}$:

$$\boldsymbol{\mu}_{ft} = \mathrm{E}[\mathbf{x}_{ft}] = \mathbf{A}_f(\boldsymbol{\Theta})\,\mathrm{E}[\mathbf{s}_{ft}|\mathbf{x}_{ft}] \quad , \tag{21}$$

for a hypothesized DOA set $\boldsymbol{\Theta}$. We approximate the expectation with a least-squares estimate:

$$\widehat{\mathbf{s}}_{ft} = \left[\mathbf{A}_f^H(\boldsymbol{\Theta})\,\mathbf{A}_f(\boldsymbol{\Theta})\right]^{-1}\mathbf{A}_f^H(\boldsymbol{\Theta})\,\mathbf{x}_{ft} \quad , \tag{22}$$

which we recognize as the output of a data-independent LCMV beamformer (i.e. $\mathbf{R}_f \propto \mathbf{I}$). Thus, we have:

$$\widehat{\boldsymbol{\mu}}_{ft} = \mathbf{A}_f(\boldsymbol{\Theta})\,\widehat{\mathbf{s}}_{ft} = \mathbf{B}_f(\boldsymbol{\Theta})\,\mathbf{x}_{ft} \quad , \tag{23}$$

where:

$$\mathbf{B}_f(\boldsymbol{\Theta}) = \mathbf{A}_f(\boldsymbol{\Theta})\left[\mathbf{A}_f^H(\boldsymbol{\Theta})\,\mathbf{A}_f(\boldsymbol{\Theta})\right]^{-1}\mathbf{A}_f^H(\boldsymbol{\Theta}) \quad , \tag{24}$$

is a projection matrix. Now we can write (20) as a zero-mean Gaussian likelihood:

$$\log\mathcal{L}_{ft}(\boldsymbol{\Theta}) \propto -\frac{1}{2\sigma_f^2}\mathbf{x}_{ft}^H\,\mathbf{P}_f(\boldsymbol{\Theta})\,\mathbf{x}_{ft} \quad , \tag{25}$$

in terms of a precision matrix:

$$\mathbf{P}_f(\boldsymbol{\Theta}) = \mathbf{I} - \mathbf{B}_f(\boldsymbol{\Theta}) \quad . \tag{26}$$

Aggregating over all t and expanding, we have:

$$\log\mathcal{L}_f(\boldsymbol{\Theta}) \propto -\frac{1}{2\sigma_f^2}\sum_{t=1}^{T}\|\mathbf{x}_{ft}\|_2^2 - \mathbf{x}_{ft}^H\,\mathbf{B}_f(\boldsymbol{\Theta})\,\mathbf{x}_{ft} \quad , \tag{27}$$

which, in the one-source case, simplifies to:

$$\log \mathcal{L}_f(\boldsymbol{\theta}) \propto -\frac{1}{2\sigma_f^2} \sum_{t=1}^{T} \|\mathbf{x}_{ft}\|_2^2 - |\mathbf{a}_f^H(\boldsymbol{\theta})\, \mathbf{x}_{ft}|^2 \quad . \tag{28}$$

This is equivalent to the SRP function defined in (17) as far as identifying the true DOA is concerned. As in [12], we can view beamforming-based localization as a maximum-likelihood problem.

B. Effect of Cross-Talk

To show the effect of cross-talk on the SRP functions, we generated a mixed data vector from two directional sources of comparable magnitude. We then computed the single- and multi-source likelihoods as well as a product of single-source likelihoods on the unit semicircle as shown in Fig. 2. Interference between the data vectors results in a spurious peak in the single-source likelihoods about halfway between the true DOAs. In contrast, the DOAs are correctly identified in the two-source likelihood up to a labeling permutation.

We can also investigate the effect of cross-talk mathematically. Consider the case of $K = 2$ sources. We can write the main term of the multi-source likelihood as:

$$\begin{aligned} &p_{MultSRP}(\mathbf{x}_{ft}) = \mathbf{x}_{ft}^H \mathbf{B} \mathbf{x}_{ft} \\ &= \frac{\|\mathbf{A}^H \mathbf{x}_{ft}\|_2^2 - 2\,\mathrm{Re}\left\{\left(\mathbf{a}_2^H \mathbf{a}_1\right)\left(\mathbf{a}_1^H \mathbf{x}_{ft}\right)\left(\mathbf{x}_{ft}^H \mathbf{a}_2\right)\right\}}{1 - |\mathbf{a}_2^H \mathbf{a}_1|^2} \quad . \end{aligned} \tag{29}$$

If we were to model the data with a sum of one-source models, we would have:

$$p_{MoSRP}(\mathbf{x}_{ft}) = \sum_{k=1}^{2} \mathbf{x}_{ft}^H \mathbf{a}_k \mathbf{a}_k^H \mathbf{x}_{ft} = \|\mathbf{A}^H \mathbf{x}_{ft}\|_2^2 \quad . \tag{30}$$

We can see that additional terms are present in the multi-source model that can change the location of the peak in the SRP function. This is observed to occur when the sources co-activate in the time-frequency plane. As an example, consider the case of two sources with equal loudness and opposite angles on the unit circle, i.e. $\mathbf{x}_{ft} = \mathbf{A}\mathbf{1} = \mathbf{a}(\boldsymbol{\theta}) + \mathbf{a}(-\boldsymbol{\theta})$. The multi-source model would exhibit a peak at the true DOA pair. Meanwhile, the single-source model would exhibit a peak at a DOA $\boldsymbol{\theta}' \neq \pm\boldsymbol{\theta}$ such that $|\mathbf{a}^H(\boldsymbol{\theta}')\, \mathbf{x}_{ft}|^2$ is maximized. A similar case is shown in Fig. 2.

V. Maximum-Likelihood Localization

We seek a maximum-likelihood estimate of the source DOAs. We will do this by gradient ascent on the SRP likelihood (27):

$$\boldsymbol{\Theta}^i \leftarrow \boldsymbol{\Theta}^{i-1} + \eta_i\, \Omega\left(\sum_{f=1}^{F} \frac{\partial \log \mathcal{L}_f(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\Big|_{\boldsymbol{\Theta}^{i-1}}\right) , \tag{31}$$

where $\Omega(-)$ performs column-wise normalization and $\eta_i = \eta_0 (I_{max} - i)/I_{max}$ is a decaying step size.

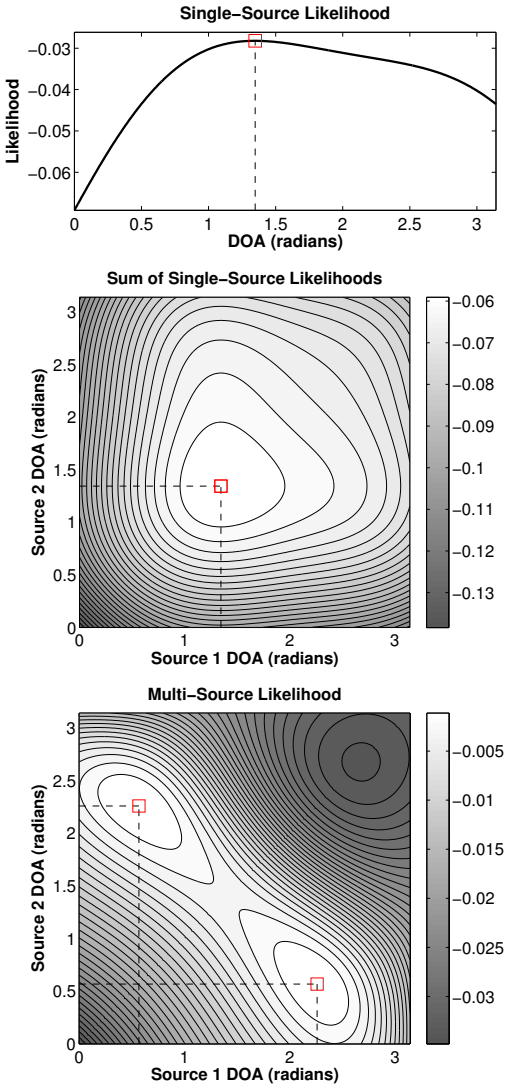

Fig. 2: SRP likelihoods for a toy data mixture of two sources with DOAs at 0.56 and 2.26 radians on the unit circle. Squares and dashed lines indicate their maxima. (Top) Single-source likelihood. (Middle) Product of single-source likelihoods. (Bottom) Multi-source likelihood.

A. One-Source Model

In the one-source case, the gradient can be written as:

$$\frac{\partial \log \mathcal{L}_f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma_f^2} \frac{2\pi l_f}{c}\, \mathbf{m}\, \mathrm{Im}\left[\mathbf{a}_f^*(\boldsymbol{\theta}) \odot (\mathbf{R}_f \mathbf{a}_f(\boldsymbol{\theta}))\right] , \tag{32}$$

where $\odot$ denotes element-wise multiplication and * denotes complex conjugation. When multiple sources are present, we model the presence of the sources at each time t with hidden variables z_{ft} that capture which source is active. We iterate between estimating the z_{ft}'s and the DOAs in a Generalized EM[1] (GEM) algorithm [36] corresponding to the complete data log likelihood:

[1]GEM is a variant of EM [35] that only requires the M step to increase the lower bound rather than explicitly maximize it.

Algorithm 1 Robust MoSRP Localization

Preprocessing

Compute STFT matrices $\mathbf{X}^{(k)} \in \mathbb{C}^{F \times T}$

Apply PHAT weighting: $\tilde{\mathbf{X}} \longleftarrow \mathbf{X} \,./\, |\mathbf{X}|$

Initialize DOAs: $\boldsymbol{\theta}^{(k)} \longleftarrow \boldsymbol{\theta}_0^{(k)}$

Optimization

for $i = 0 : I_{max} - 1$ **do**

Compute steering matrices: $\mathbf{A}_f = \frac{1}{\sqrt{M}} e^{j \frac{2\pi l_f}{c}} \mathbf{m}^\top \boldsymbol{\Theta}$

Compute projectors: $\mathbf{B}_f = \mathbf{A}_f \left[\mathbf{A}_f^H \mathbf{A}_f \right]^{-1} \mathbf{A}_f^H$

if *Wiener_mask* **then**

Estimate TF weights: $\widehat{\gamma}_{ft} = \frac{\|\mathbf{B}_f \mathbf{x}_{ft}\|_2^2}{\|\mathbf{x}_{ft}\|_2^2}$

else

No weighting: $\widehat{\gamma}_{ft} = 1$

end if

Compute posteriors: $\log w_{ft}^{(k)} \propto \frac{|\mathbf{a}_f^{(k)H} \tilde{\mathbf{x}}_{ft}|^2}{2\sigma_f^2}$

Estimate correlation matrices:

$$\widehat{\mathbf{R}}_f^{(k)} = \sum_{t=1}^{T} \widehat{\gamma}_{ft}\, w_{ft}^{(k)} \tilde{\mathbf{x}}_{ft} \tilde{\mathbf{x}}_{ft}^H$$

Compute gradients: $\mathbf{g}^{(k)} = \sum_{f=1}^{F} l_f \cdots$

$$\mathbf{m}\,\mathrm{Im}\left[\mathbf{a}_f^{(k)^*} \odot \left(\widehat{\mathbf{R}}_f^{(k)} \mathbf{a}_f^{(k)} \right) \right]$$

Normalize gradients: $\mathbf{g}^{(k)} \longleftarrow \mathbf{g}^{(k)} \,/\, \|\mathbf{g}^{(k)}\|_2$

Compute step size: $\eta_i = \eta_0 \left(I_{max} - i\right) / I_{max}$

Update DOA vectors: $\boldsymbol{\theta}^{(k)} \longleftarrow \boldsymbol{\theta}^{(k)} + \eta_i\, \mathbf{g}^{(k)}$

Project DOAs to hemisphere: $\boldsymbol{\theta}^{(k)} \longleftarrow \boldsymbol{\theta}^{(k)} / \|\boldsymbol{\theta}^{(k)}\|_2$

end for

Algorithm 2 Robust MultSRP Localization

Preprocessing

Compute STFT matrices $\mathbf{X}^{(k)} \in \mathbb{C}^{F \times T}$

Apply PHAT weighting: $\tilde{\mathbf{X}} \longleftarrow \mathbf{X} \,./\, |\mathbf{X}|$

Estimate correlation matrices: $\widehat{\mathbf{R}}_f = \frac{1}{N} \sum_{t=1}^{T} \tilde{\mathbf{x}}_{ft} \tilde{\mathbf{x}}_{ft}^H$

Initialize DOAs: $\boldsymbol{\Theta} \longleftarrow \boldsymbol{\Theta}_0$

Optimization

for $i = 0 : I_{max} - 1$ **do**

Compute steering matrices: $\mathbf{A}_f = \frac{1}{\sqrt{M}} e^{j \frac{2\pi l_f}{c}} \mathbf{m}^\top \boldsymbol{\Theta}$

Compute projectors: $\mathbf{B}_f = \mathbf{A}_f \left[\mathbf{A}_f^H \mathbf{A}_f \right]^{-1} \mathbf{A}_f^H$

if *Wiener_mask* **then**

Estimate TF weights: $\widehat{\gamma}_{ft} = \frac{\|\mathbf{B}_f \mathbf{x}_{ft}\|_2^2}{\|\mathbf{x}_{ft}\|_2^2}$

Estimate correlation matrices:

$$\widehat{\mathbf{R}}_f = \sum_{t=1}^{T} \widehat{\gamma}_{ft} \tilde{\mathbf{x}}_{ft} \tilde{\mathbf{x}}_{ft}^H$$

end if

Compute gradient: $\mathbf{G} = \sum_{f=1}^{F} l_f \cdots$

$$\mathbf{m}\,\mathrm{Im}\left[\mathbf{A}_f^* \odot \left(\left(\mathbf{I} - \mathbf{B}_f\right) \widehat{\mathbf{R}}_f \mathbf{A}_f \left[\mathbf{A}_f^H \mathbf{A}_f \right]^{-1} \right) \right]$$

Normalize source gradients: $\mathbf{g}^{(k)} \longleftarrow \mathbf{g}^{(k)} \,/\, \|\mathbf{g}^{(k)}\|_2$

Compute step size: $\eta_i = \eta_0 \left(I_{max} - i\right) / I_{max}$

Update DOA matrix: $\boldsymbol{\Theta} \longleftarrow \boldsymbol{\Theta} + \eta_i\, \mathbf{G}$

Project DOAs to hemisphere: $\boldsymbol{\theta}^{(k)} \longleftarrow \boldsymbol{\theta}^{(k)} / \|\boldsymbol{\theta}^{(k)}\|_2$

end for

$$\log\, p\left(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\Theta}\right) \propto \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{1}{2\sigma_f^2} \left[z_{ft} = k\right] \left|\mathbf{a}_f^H\left(\boldsymbol{\theta}_k\right) \mathbf{x}_{ft}\right|^2 . \tag{33}$$

where $[-]$ is the indicator function. The EM lower bound can be written as:

$$\mathcal{Q} \propto \sum_{f=1}^{F} \sum_{k=1}^{K} \frac{1}{2\sigma_f^2}\, \mathbf{a}_f^H\left(\boldsymbol{\theta}_k\right) \widehat{\mathbf{R}}_f^{(k)}\, \mathbf{a}_f\left(\boldsymbol{\theta}_k\right) \quad , \tag{34}$$

where:

$$\widehat{\mathbf{R}}_f^{(k)} = \frac{\sum_{t=1}^{T} w_{ft}^{(k)}\, \mathbf{x}_{ft}\, \mathbf{x}_{ft}^H}{\sum_{t=1}^{T} w_{ft}^{(k)}} \quad , \tag{35}$$

are source-specific correlation matrices defined in terms of the posterior probabilities of the z_{ft}'s:

$$\log w_{ft}^{(k)} = \log p\left(z_{ft} = k \,|\, \mathbf{x}_{ft}\right) \propto \frac{\left|\mathbf{a}_f^H\left(\boldsymbol{\theta}_k\right) \mathbf{x}_{ft}\right|^2}{2\sigma_f^2} \quad . \tag{36}$$

In the E step, we compute soft TF weights and correlation matrices with (35)-(36), and in the M step, we optimize each source's DOA with (31)-(32). Thus, the EM procedure alternates between estimating localization (DOA) and separation (TF mask) parameters to fit a Mixture of SRP functions (MoSRP). For simplicity, we assume that all the mixing weights are equal and use a constant variance of $\sigma_f^2 = 10^{-2}$ when calculating posterior probabilities.

B. Multiple-Source Model

The gradient for multiple sources is (see Appendix B):

$$\frac{\partial \log \mathcal{L}_f\left(\boldsymbol{\Theta}\right)}{\partial\, \boldsymbol{\Theta}} = \frac{1}{\sigma_f^2} \frac{2\pi l_f}{c} \cdots$$
$$\mathbf{m}\,\mathrm{Im}\left[\mathbf{A}_f^* \odot \left(\mathbf{P}_f \mathbf{R}_f \mathbf{A}_f \left[\mathbf{A}_f^H \mathbf{A}_f \right]^{-1} \right) \right] , \tag{37}$$

Gradient ascent with this model avoids the complexity of the EM algorithm while explicitly accounting for cross-talk. An efficient implementation for $K = 2$ sources is derived by expanding (37) in terms of the source-specific steering vectors (see Appendix C).

C. Robustness to Non-Directional Interference

When non-directional interference $\mathbf{e}_{ft}$ is present, we have:

$$\mathbf{x}_{ft} = \mathbf{A}_f\left(\boldsymbol{\Phi}\right) \mathbf{s}_{ft} + \mathbf{n}_{ft} + \mathbf{e}_{ft} = \mathbf{b}_{ft} + \mathbf{c}_{ft} \quad , \tag{38}$$

where $\mathbf{b}_{ft} = \mathbf{A}_f\left(\boldsymbol{\Phi}\right) \mathbf{s}_{ft}$ and $\mathbf{c}_{ft} = \mathbf{n}_{ft} + \mathbf{e}_{ft}$. The Wiener mask [25] gives the MMSE-optimal weighting to recover $\mathbf{b}_{ft}$:

$$\gamma_{ft}^{opt} = \frac{\|\mathbf{b}_{ft}\|_2^2}{\|\mathbf{b}_{ft}\|_2^2 + \|\mathbf{c}_{ft}\|_2^2} \quad . \tag{39}$$

Thus, a robust estimate of the correlation matrices is:

$$\widehat{\mathbf{R}}_f = \frac{\sum_{t=1}^{T} \gamma_{ft}^{opt} \mathbf{x}_{ft} \mathbf{x}_{ft}^H}{\sum_{t=1}^{T} \gamma_{ft}^{opt}} \quad . \tag{40}$$

In practice, we approximate the weights using:

$$\widehat{\mathbf{b}}_{ft} = \mathbf{B}_f(\boldsymbol{\Theta})\, \mathbf{x}_{ft} \quad , \tag{41}$$

$$\widehat{\mathbf{c}}_{ft} = \mathbf{x}_{ft} - \widehat{\mathbf{b}}_{ft} \quad , \tag{42}$$

which gives:

$$\widehat{\gamma}_{ft} = \frac{\|\widehat{\mathbf{b}}_{ft}\|_2^2}{\|\widehat{\mathbf{b}}_{ft}\|_2^2 + \|\widehat{\mathbf{c}}_{ft}\|_2^2} = \frac{\|\widehat{\mathbf{b}}_{ft}\|_2^2}{\|\mathbf{x}_{ft}\|_2^2} \quad . \tag{43}$$

Interleaving this step with DOA optimization improves localization accuracy in the presence of ambient noise. For a mixture of one-source models, we estimate the correlation matrices as in (35) by multiplying the posteriors with the Wiener filter weights. Since the gradients are accumulated over all frequency bands, we weight the contributions from each band by the sum of all the corresponding TF weights. Due to the form of the gradient, this is equivalent to omitting the denominator in (40).

The overall iterative scheme is shown in Fig. 1. Pseudocode for the Mixture of SRPs (MoSRP) and Multi-source SRP (MultSRP) localizers are given in Algorithms 1 and 2.

VI. Source Separation

Once the gradient ascent procedure has converged, any number of methods can be used to enhance/separate the directional signals, if necessary. For example, we can isolate each source by estimating time-frequency (TF) masks and applying them to $\mathbf{X}$. The mask weights are calculated as:

$$\log w_{ft}^{(k)} \propto \frac{1}{2\nu^2} |\, \widehat{s}_{ft}^{(k)} \,|^2 \quad , \tag{44}$$

using estimates of the source coefficients provided by K data-independent LCMV beamformers, each designed to isolate a single source while blocking out the others. This can be implemented as:

$$\widehat{\mathbf{s}}_{ft} = \left[\mathbf{A}_f^H \mathbf{A}_f\right]^{-1} \mathbf{A}_f^H \mathbf{x}_{ft} \quad . \tag{45}$$

The variance $\nu \in (0, \infty)$ controls the hardness of the mask such that as $\nu \to 0$, the mask becomes binary, assigning each TF bin entirely to a single source. In our experiments, we used $\nu = 0.1$.

VII. Experiments

A. Speaker Localization in Noisy Conditions

We ran speaker localization experiments in a room simulator of size $5 \times 5 \times 5$ meters with an 8-channel, 10×10 centimeter square array placed in the center of the room. The microphones were equally spaced around the perimeter of the square (including the corners). Between 2 and 4 sources were placed on the unit hemisphere above the array at random, ensuring sufficient separation: $\forall\ j \neq k \quad \cos^{-1}\left(\boldsymbol{\phi}_j^\top \boldsymbol{\phi}_k\right) \geq \frac{2\pi}{K+2}$. For each speaker, we concatenated random sentences from the TSP database [37] to form a 10-second-long source signal. Reverberation with a T_{60} time of 260 milliseconds was simulated with the image method [38]. Varying amounts of either of two types of ambient noise were added instantaneously to the reverberant mixtures: (1) white Gaussian noise and (2) crowd noises from the BBC Sounds Effects Library. In each simulation with crowd noise, one single-channel noise signal was replicated across the channels and each replicate was randomly circularly shifted in time to ensure spatial incoherence. All STFTs were computed with window and hop sizes of 1024 and 256 using a Hann window.

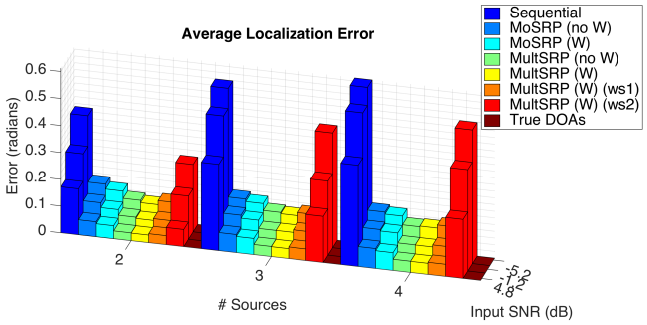

Fig. 3: Localization errors for white Gaussian noise and an ideal initialization. All gradient methods perform well with the multi-source models performing slightly better.

We compared three methods: mixture of one-source SRPs (MoSRP), multi-source SRP (MultSRP), and traditional SRP localization ("Sequential"). The first two were tested with and without the Wiener filtering stage described in Section V-C. In the figures, algorithms using Wiener filtering are labeled with "(W)". In addition, we ran the MultSRP method initialized with (i.e. with a warm start from) either the MoSRP method or the sequential method (MultSRP (ws1) and MultSRP (ws2), respectively). The sequential method was implemented with a grid of 709 DOAs uniformly spread over the hemisphere.

We found that the initialization method is important for the SRP models. To evaluate this, we ran experiments both with and without an ideal initialization. The ideal initialization involves placing the initial DOA vector estimates at the true source DOAs. Our practical initialization method is based on fitting wrapped lines to Inter-channel Phase Difference (IPD) features with the Random Sample Consensus (RANSAC) algorithm [39]. A similar procedure was used in [23].

Localization errors averaged over 100 trials are shown in Figs. 3-6. The error is averaged over the sources and captures angular deviation from the true DOAs:

$$e = \min_{\mathcal{P}} \frac{1}{K} \sum_{k=1}^{K} \cos^{-1}\left(\boldsymbol{\phi}_k^\top \widehat{\boldsymbol{\theta}}_{\mathcal{P}(k)}\right) \tag{46}$$

where $\mathcal{P}$ is a permutation mapping. Minimizing over permu-

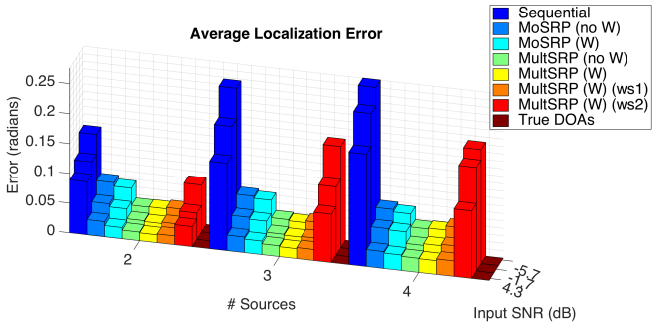

Fig. 4: Localization errors for crowd noise and an ideal initialization. All gradient methods perform well with the multi-source models performing slightly better.

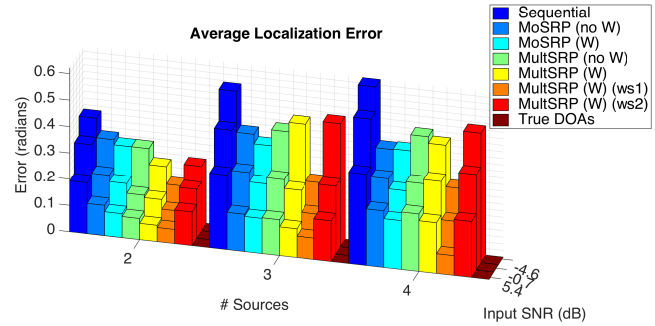

Fig. 5: Localization errors for white Gaussian noise and a practical initialization. The MultSRP model performs best when initialized with the MoSRP result.

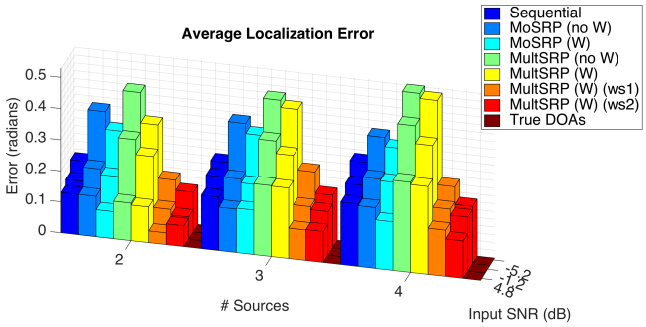

Fig. 6: Localization errors for crowd noise and a practical initialization. The MultSRP model with a warm start performs best in all cases.

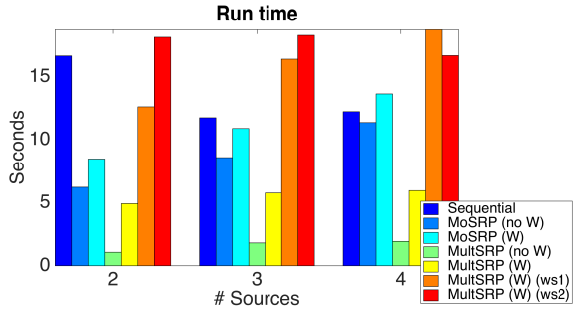

Fig. 7: Average run times for noisy, 10-second speech mixtures. All experiments were conducted on an iMac with 16 GB of RAM and a 3.4 GHz processor.

tations deals with the source labeling ambiguity.

We see that with an ideal initialization and either noise type, the MultSRP algorithms consistently out-perform the others. This is probably due its robustness to noise once the correct DOAs are identified. We can understand this by observing that the projection matrices $\mathbf{B}_f$ behave like a series of LCMV beamformers steered towards the sources (see Algorithm 2). In contrast, the MoSRP method effectively steers delay-and-sum beamformers towards the sources, making it more sensitive to noise and cross-talk.

When the practical initialization is used, the results differ significantly. For white Gaussian noise, the MultSRP algorithm with Wiener filtering and a warm start from MoSRP gives the best performance. Wiener filtering generally appears to help both SRP methods, especially in the low-SNR regime. The most accurate from among the SRP methods almost always out-performs traditional SRP localization with the exception being in the case of crowd noise. This, of course, is mostly an issue of initialization in the SRP methods. Once initialized with the result of the sequential method, MultSRP with a warm start performs significantly better.

Run times averaged over 50 trials using the practical initialization are shown in Fig. 7. The MultSRP method is by far the most efficient. This is mainly due to its simplicity and the fact that computations account for all the sources at once. The MoSRP method varies linearly with the number of sources because it computes source-specific gradients. We note that these algorithms can be parallelized over sources (where applicable) and frequencies. We also note that the efficient form of the 2-source MultSRP gradient (see Appendix C) reduces its computation time by about a factor of 1.5.

Figure 8 shows the source separation results corresponding to the experiments of Figure 5. Separation performance was measured via the Signal-to-Interference Ratio (SIR) metric with the BSS Eval toolbox [40]. These results qualitatively mirror the localization results. We would expect this to be the case since we are better able to isolate the energy corresponding to each speaker with more accurate estimates of their DOAs.

B. *Early Reflections*

Figure 9 shows the result of a synthetic experiment with early reflections. An 8-channel, square array of side-length 20 centimeters was positioned 10 centimeters from the middle of a wall in a 3-dimensional room of size $5\times5\times5$ meters. The image method [38] was applied to simulate reverberation for two speakers using random sentences from the TSP database [37]. In order to capture the effect of the early reflections, we set the number of estimated sources to $K = 4$ (twice the true number). We can observe that the mixture of SRPs model (MoSRP) fails to find the source images while the multi-source SRP model (MultSRP) localizes them correctly. The MoSRP and MultSRP models achieve localization errors of $e = 0.71$

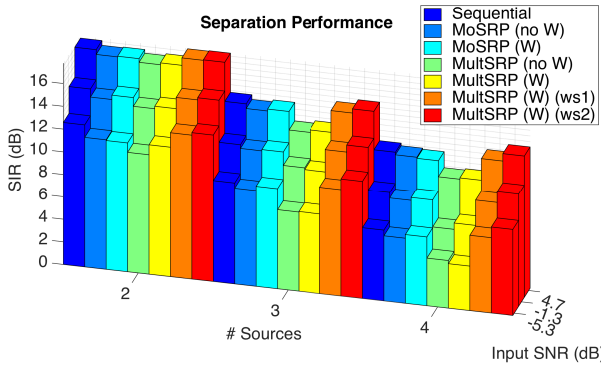

Fig. 8: Source separation results for 10-second speech mixtures in white Gaussian noise. The MultSRP method (with a warm start) increasingly performs better than the others for more sources, especially for a low SNR.

Fig. 9: Localization results with early reflections off of a wall. In this figure, the 3-dimensional room is viewed from above.

and $e = 0.18$ averaged over 10 trials with a similar setup (the source positions were perturbed relative to the positions shown in Fig. 9). In this experiment, we used a simple initialization strategy that places the initial DOA vectors near the top of the hemisphere (i.e. the center of Fig. 9).

C. Other SRP objective functions

We applied the single- and multi-source optimization approaches described in Section V to the empirical output power functions of the MVDR and data-independent and -dependent LCMV beamformers. However, we found that this does not perform nearly as well in practice as optimizing the proposed SRP likelihoods. This may be due to sensitivity to the inversion of $\mathbf{R}_f$, which must be estimated from data, and to other real-world factors such as reverberation and overlap in the TF plane. In addition, the EM algorithm with a mixture of MVDR objectives is quite slow since KF matrices $\widehat{R}_f^{(k)}$ must be calculated *and inverted* in every iteration.

VIII. Conclusions

We used a probabilistic interpretation of the Steered Response Power (SRP) function to describe two algorithms for the simultaneous source localization and separation problem. The first was based on a mixture-of-SRPs (MoSRP) model and lead to a GEM algorithm for optimizing the sources' Directions-Of-Arrival (DOA). The second was based on a multi-source SRP (MultSRP) function that was robust to source overlap in the Time-Frequency (TF) domain and lead to a simple gradient ascent scheme. Unlike in traditional sequential methods, Wiener filtering was tied into the DOA optimization procedure to make it robust to non-directional interference. Experiments showed that the proposed unified approach outperforms a sequential one. We also showed that the MultSRP model can localize highly coherent sources with significant overlap in the TF plane, unlike the MoSRP model.

Appendix A
Derivation of SRP likelihood

We derive a generalized form for the SRP likelihood of Section IV-A. We assume a Gaussian distribution for each data vector $\mathbf{x}_{ft}$ (dependence on f and t are omitted for clarity):

$$\mathcal{L} = -\frac{1}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^H\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) \quad , \tag{47}$$

To find estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ specific to $\mathbf{x}$, we consider the following generative model:

$$\mathbf{s} \sim \mathcal{N}\left(\boldsymbol{\mu}_\mathbf{s}, \boldsymbol{\Sigma}_\mathbf{s}\right) \quad , \quad \mathbf{x}|\mathbf{s} \sim \mathcal{N}\left(\mathbf{A}\,\mathbf{s}, \boldsymbol{\Sigma}_\mathbf{n}\right) \quad . \tag{48}$$

The general form of the posterior distribution [41] is:

$$\mathbf{s}|\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{x}}\right) \quad , \tag{49}$$

$$\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{x}}\left(\mathbf{A}^H\boldsymbol{\Sigma}_\mathbf{n}^{-1}\mathbf{x} + \boldsymbol{\Sigma}_\mathbf{s}^{-1}\boldsymbol{\mu}_\mathbf{s}\right) \quad , \tag{50}$$

$$\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{x}}^{-1} = \mathbf{A}^H\boldsymbol{\Sigma}_\mathbf{n}^{-1}\mathbf{A} + \boldsymbol{\Sigma}_\mathbf{s}^{-1} \quad . \tag{51}$$

Assuming a vague prior (i.e. $\boldsymbol{\Sigma}_\mathbf{s} \to \infty\,\mathbf{I}$), we have:

$$\widehat{\boldsymbol{\mu}} = \mathrm{E}\left[\mathbf{x}\right] = \mathrm{E}\left[\mathbf{A}\mathbf{s}+\mathbf{n}\right] = \mathbf{A}\,\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}} = \mathbf{B}\,\mathbf{x} \quad , \tag{52}$$

$$\widehat{\boldsymbol{\Sigma}} = \mathrm{E}\left[\left(\mathbf{x}-\widehat{\boldsymbol{\mu}}\right)\left(\mathbf{x}-\widehat{\boldsymbol{\mu}}\right)^H\right] \tag{53}$$

$$= \mathrm{E}\left[\left(\mathbf{A}\,\mathbf{s}+\mathbf{n}-\mathbf{A}\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}}\right)\left(\mathbf{A}\,\mathbf{s}+\mathbf{n}-\mathbf{A}\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}}\right)^H\right] \tag{54}$$

$$= \mathrm{E}\left[\left(\mathbf{A}\left(\mathbf{s}-\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}}\right)+\mathbf{n}\right)\left(\mathbf{A}\left(\mathbf{s}-\boldsymbol{\mu}_{\mathbf{s}|\mathbf{x}}\right)+\mathbf{n}\right)^H\right] \tag{55}$$

$$= \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{x}}\mathbf{A}^H + \boldsymbol{\Sigma}_\mathbf{n} \tag{56}$$

$$= \left(\mathbf{B}+\mathbf{I}\right)\boldsymbol{\Sigma}_\mathbf{n} \quad , \tag{57}$$

where $\mathbf{B} = \mathbf{A}\left[\mathbf{A}^H\boldsymbol{\Sigma}_\mathbf{n}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}^H\boldsymbol{\Sigma}_\mathbf{n}^{-1}$. We substitute $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ into (47), evaluate $\left|\widehat{\boldsymbol{\Sigma}}\right|$ with Sylvester's determinant theorem and $\widehat{\boldsymbol{\Sigma}}^{-1}$ with the Woodbury identity, and simplify using the fact that $\mathbf{B}$ is idempotent (i.e. $\mathbf{B}^2 = \mathbf{B}$). Thus, we have:

$$\mathcal{L} = -\frac{1}{2}\log\left|2\pi\,\boldsymbol{\Sigma}_\mathbf{n}\right| - \frac{K}{2}\log 2 - \frac{1}{2}\mathbf{x}^H\mathbf{P}^H\boldsymbol{\Sigma}_\mathbf{n}^{-1}\mathbf{P}\mathbf{x} \quad , \tag{58}$$

where $\mathbf{P} = \mathbf{I} - \mathbf{B}$. Finally, if $\boldsymbol{\Sigma}_\mathbf{n} = \sigma^2\,\mathbf{I}$, we have:

$$\mathcal{L} \propto -\frac{1}{2\sigma^2}\mathbf{x}^H\mathbf{P}\mathbf{x} \quad . \tag{59}$$

APPENDIX B
GRADIENT OF SRP LIKELIHOOD

We derive the gradient of the multiple-source SRP likelihood with matrix calculus [42]. The DOA-dependent contribution to the SRP likelihood (27) from a single observed vector $\mathbf{x}$ is:

$$\mathcal{L} \propto \frac{1}{2\sigma^2}\,\mathbf{x}^H\,\mathbf{B}\,\mathbf{x}\ \ , \tag{60}$$

where $\mathbf{B} = \mathbf{A}\left[\mathbf{A}^H\mathbf{A}\right]^{-1}\mathbf{A}^H$ and:

$$A_{iv} = \frac{1}{\sqrt{M}} e^{j\frac{2\pi l_f}{c}\sum\limits_u m_{ui}\Theta_{uv}}\ \ . \tag{61}$$

Applying the product rule, we identify three terms in the gradient corresponding to the gradients of:

$$\mathcal{P}^{(1)} = \mathbf{x}^H\mathbf{A}\mathbf{y}\ \ , \tag{62}$$

$$\mathcal{P}^{(2)} = \mathbf{z}^H\left[\mathbf{A}^H\mathbf{A}\right]^{-1}\mathbf{z}\ \ , \tag{63}$$

$$\mathcal{P}^{(3)} = \mathbf{y}^H\mathbf{A}^H\mathbf{x}\ \ , \tag{64}$$

where $\mathbf{y} = \left[\mathbf{A}^H\mathbf{A}\right]^{-1}\mathbf{z}$ and $\mathbf{z} = \mathbf{A}^H\mathbf{x}$ are treated as constants. Based on the result:

$$\frac{\partial\,\mathcal{P}^{(1)}}{\partial\,\Theta_{uv}} = j\frac{2\pi l_f}{c}\sum_i m_{ui}x_i^*A_{iv}y_v\ \ , \tag{65}$$

we can write:

$$\frac{\partial\,\mathcal{P}^{(1)}}{\partial\,\mathbf{\Theta}} = j\,\frac{2\pi l_f}{c}\mathbf{m}\,\mathrm{diag}\left(\mathbf{x}^*\right)\mathbf{A}\,\mathrm{diag}\left(\mathbf{y}\right)\ \ , \tag{66}$$

$$\frac{\partial\,\mathcal{P}^{(3)}}{\partial\,\mathbf{\Theta}} = -j\,\frac{2\pi l_f}{c}\mathbf{m}\,\mathrm{diag}\left(\mathbf{x}\right)\mathbf{A}^*\,\mathrm{diag}\left(\mathbf{y}^*\right)\ \ . \tag{67}$$

With some abuse of notation, the second gradient term is:

$$\frac{\partial\,\mathcal{P}^{(2)}}{\partial\,\mathbf{\Theta}} = \mathbf{z}^H\left(\frac{\partial\left[\mathbf{A}^H\mathbf{A}\right]^{-1}}{\partial\,\mathbf{\Theta}}\right)\mathbf{z} \tag{68}$$

$$= -\mathbf{z}^H\left[\mathbf{A}^H\mathbf{A}\right]^{-1}\left(\frac{\partial\,\mathbf{A}^H\mathbf{A}}{\partial\,\mathbf{\Theta}}\right)\left[\mathbf{A}^H\mathbf{A}\right]^{-1}\mathbf{z} \tag{69}$$

$$= -\mathbf{y}^H\left[\left(\frac{\partial\,\mathbf{A}}{\partial\,\mathbf{\Theta}}\right)^H\mathbf{A} + \mathbf{A}^H\left(\frac{\partial\,\mathbf{A}}{\partial\,\mathbf{\Theta}}\right)\right]\mathbf{y} \tag{70}$$

$$= -\left[\frac{\partial\,\mathbf{y}^H\mathbf{A}^H\mathbf{q}}{\partial\,\mathbf{\Theta}} + \frac{\partial\,\mathbf{q}^H\mathbf{A}\mathbf{y}}{\partial\,\mathbf{\Theta}}\right] \tag{71}$$

$$= -j\frac{2\pi l_f}{c}\mathbf{m}\big[-\mathrm{diag}\left(\mathbf{q}\right)\mathbf{A}^*\,\mathrm{diag}\left(\mathbf{y}^*\right)\ \cdots + \mathrm{diag}\left(\mathbf{q}^*\right)\mathbf{A}\,\mathrm{diag}\left(\mathbf{y}\right)\big]\ , \tag{72}$$

where $\mathbf{q} = \mathbf{A}\,\mathbf{y}$ is treated as a constant and we have applied the results in (66)-(67) to (71). The temporary abuse of notation is tolerable once we observe that the non-zero terms in the tensor-valued gradient expressions form a matrix.

It follows that:

$$\frac{\partial\,\mathcal{P}}{\partial\,\mathbf{\Theta}} = \sum_{i=1}^{3}\frac{\partial\,\mathcal{P}^{(i)}}{\partial\,\mathbf{\Theta}} \tag{73}$$

$$= \frac{4\pi l_f}{c}\mathbf{m}\big(-\mathrm{Im}\left[\mathrm{diag}\left(\mathbf{x}^*\right)\mathbf{A}\,\mathrm{diag}\left(\mathbf{y}\right)\right]\ \cdots + \mathrm{Im}\left[\mathrm{diag}\left(\mathbf{q}^*\right)\mathbf{A}\,\mathrm{diag}\left(\mathbf{y}\right)\right]\big) \tag{74}$$

$$= \frac{4\pi l_f}{c}\mathbf{m}\left(-\mathrm{Im}\left[\mathbf{A}\odot\left(\mathbf{x}^*\mathbf{y}^\top\right)\right] + \mathrm{Im}\left[\mathbf{A}\odot\left(\mathbf{q}^*\mathbf{y}^\top\right)\right]\right) \tag{75}$$

$$= \frac{4\pi l_f}{c}\mathbf{m}\left(-\mathrm{Im}\left[\mathbf{A}\odot\left(\mathbf{y}\,\mathbf{x}^H\right)^\top\right] + \mathrm{Im}\left[\mathbf{A}\odot\left(\mathbf{y}\,\mathbf{q}^H\right)^\top\right]\right)\ . \tag{76}$$

Aggregating over time and consolidating terms, we have:

$$\frac{\partial\,\mathcal{P}}{\partial\,\mathbf{\Theta}} = \frac{4\pi l_f}{c}\mathbf{m}\,\mathrm{Im}\left[\mathbf{A}^*\odot\left(\mathbf{P}\mathbf{C}^H\right)\right]\ \ , \tag{77}$$

where $\mathbf{C} = \left[\mathbf{A}^H\mathbf{A}\right]^{-1}\mathbf{A}^H\mathbf{R}$ and $\mathbf{P} = \mathbf{I} - \mathbf{B}$ is the projection matrix defined in (26). The final result for the gradient is:

$$\frac{\partial\,\mathcal{L}}{\partial\,\mathbf{\Theta}} = \frac{1}{2\sigma^2}\frac{\partial\,\mathcal{P}}{\partial\,\mathbf{\Theta}} = \frac{1}{\sigma^2}\frac{2\pi l_f}{c}\mathbf{m}\,\mathrm{Im}\left[\mathbf{A}^*\odot\left(\mathbf{P}\mathbf{C}^H\right)\right]\ \ . \tag{78}$$

APPENDIX C
EFFICIENT COMPUTATION OF GRADIENT AND LIKELIHOOD FOR 2 SOURCES

The gradient derived in Appendix B can be implemented more efficiently in the case of $K = 2$ sources. We start by explicitly evaluating the matrix inverse:

$$\left[\mathbf{A}^H\mathbf{A}\right]^{-1} = \frac{1}{d}\begin{bmatrix}1 & -\mathbf{a}_1^H\mathbf{a}_2\\ -\mathbf{a}_2^H\mathbf{a}_1 & 1\end{bmatrix}\ \ , \tag{79}$$

where $d = 1 - |\mathbf{a}_1^H\mathbf{a}_2|^2$. Substituting this into (78), expanding, and re-arranging, we have:

$$\frac{\partial\,\mathcal{L}}{\partial\,\mathbf{\Theta}} = \frac{1}{\sigma^2}\frac{2\pi l_f}{c\,d}\mathbf{m}\,\mathrm{Im}\left[\mathbf{z}_{11} - v_{21}\,\mathbf{z}_{12} - \mathbf{y}\ ,\quad \mathbf{z}_{22} - v_{12}\,\mathbf{z}_{21} + \mathbf{y}\right]\ \ , \tag{80}$$

where:

$$\mathbf{y} = \frac{r_{12}v_{21}^2 - \left(r_{11} + r_{22}\right)v_{21} + r_{21}}{d}\mathbf{a}_1^*\odot\mathbf{a}_2\ \ , \tag{81}$$

$$\mathbf{q}_i = \mathbf{R}\,\mathbf{a}_i\ \ , \tag{82}$$

$$\mathbf{z}_{ij} = \mathbf{a}_i^*\odot\mathbf{q}_j\ \ , \tag{83}$$

$$v_{ij} = \mathbf{a}_i^H\mathbf{a}_j\ \ , \tag{84}$$

$$r_{ij} = \mathbf{a}_i^H\mathbf{R}\,\mathbf{a}_j = \mathbf{a}_i^H\mathbf{q}_j\ \ . \tag{85}$$

The likelihood can be computed efficiently as:

$$\mathcal{L} = -\frac{1}{2\sigma^2}\left(-\|\mathbf{X}\|_F^2 + \frac{r_{11} + r_{22} - 2\,\mathrm{Re}\left[v_{21}r_{12}\right]}{d}\right)\ \ . \tag{86}$$

REFERENCES

[1] H. K. van Trees, *Detection, Estimation, and Modulation Theory: Optimum Array Processing (Part IV)*, Wiley, 2002.

[2] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3 edition, 1995.

[3] J. Benesty, J. Chen, and Y. Huang, *Topics in Signal Processing: Microphone Array Signal Processing*, vol. 1, Springer, 2008.

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

[5] Y. Mahieux C. Marro and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, 1998.

[6] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *IEE Proceedings*, vol. 140, pp. 362–370, 1993.

[7] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatio-temporal information," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–17, 2006.

[8] W. Herbordt, *Sound Capture for Human-Machine Interfaces*, Springer, 2005.

[9] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, 2009.

[10] S. Gazor, S. Affes, and Y. Grenier, "Wideband multi-source beamforming with adaptive array location calibration and direction finding," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1904–1907, 1995.

[11] 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, 2007.

[12] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian acoustic localization," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 1793–1796, 2002.

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

[14] S. T. Birchfield and D. K. Gillmor, "Acoustic source direction by hemisphere sampling," *IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3053–3056, 2001.

[15] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[16] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 19, no. 1, pp. 134–139, 1983.

[17] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[18] M. R. Azimi-Sadjadi, A. Pezeshki, and L. L. Scharf, "Wideband DOA estimation algorithms for multiple target detection and tracking using unattended acoustic sensors," in *Proceedings of the Society of Photographic Instrumentation Engineers*, 2004, vol. 5417, pp. 1–11.

[19] K. H. Knuth, "Bayesian source separation and localization," *Proceedings of SPIE: Bayesian Inference for Inverse Problems*, vol. 3459, pp. 147–158, 1998.

[20] B. Lee, *Robust speech recognition in a car using a microphone array*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2006.

[21] K. Harmanci, J. Tabrikian, and J.L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Transactions on Signal Processing*, vol. 48, no. 1, 2000.

[22] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 528–548, 2008.

[23] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[24] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1327–1339, 2007.

[25] Y. Li and D Wang, "On the optimality of ideal binary time-frequency masks," *Elsevier - Speech Communication*, vol. 51, pp. 230–239, 2009.

[26] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," *IEEE Conference on Acoustics, Speech, and Signal Processing (CASSP)*, vol. 1, pp. 529–532, 2002.

[27] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[28] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833 – 1847, 2007.

[29] S. M. Naqvi, W. Wang, M. S. Khan, M. Barnard, and J. A. Chambers, "Multimodal (audiovisual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking," *Signal Processing, IET*, vol. 6, no. 5, pp. 466–477, 2012.

[30] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic source tracking applying Rao-Blackwellized Monte Carlo data association," *IEEE 15th Workshop on Statistical Signal Processing*, pp. 253–256, 2009.

[31] M. Wax and A. Leshem, "Joint estimation of time delays and directions of arrival of multiple reflections of a known signal," *IEEE Transactions on Signal Processing*, vol. 45, no. 10, pp. 2477–2484, 1997.

[32] M. L. Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," *IEEE Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 104–107, 2008.

[33] X. Zhao, J. Tang, L. Zhou, and Z. Wu, "A fast search method of steered response power with small-aperture microphone array for sound source localization," *Journal of Electronics (China)*, vol. 30, no. 5, pp. 483–490, 2013.

[34] Hoang Do, H.F. Silverman, and Ying Yu, "A real-time srp-phat source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *IEEE Internatinal Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 1.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[36] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. 1998, pp. 355 – 368, Kluwer Academic Publishers.

[37] P. Kabal, "TSP speech database," 2002, Telecommunications and Signal Processing Lab, McGill University.

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

[39] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with ransac and directional statistics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, pp. 2233–2243, 2014.

[40] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, 2006.

[41] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[42] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2012.

Johannes Traa (traa2@illinois.edu) received the B.S. degree in electrical engineering from Northwestern University, Evanston, IL, in 2011 and the M.S. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, in 2013. He is currently pursuing the Ph.D. degree in electrical and computer engineering at UIUC.

He has been a Research Assistant with Paris Smaragdis at UIUC since 2011. In the summers of 2012-2014, he interned with the theory group at Lyric Labs, Analog Devices in Boston, MA. His research interests include audio source separation and localization, sound mixture analysis with additive modeling techniques like non-negative matrix factorization, and applications of various areas of statistics (e.g. compositional, directional) to audio problems.

David Wingate (wingated@cs.byu.edu) received his BS and MS degrees in computer science from Brigham Young University in 2002 and 2004, and a PhD in computer science from the University of Michigan in 2008. He was a postdoctoral fellow at MIT from 2008-2010 with a joint appointment in the Computer Science and Artificial Intelligence Laboratory and the Computational Cognitive Science group in the Brain and Cognitive Science Department. From 2010-2012 he was a research scientist at MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 he was a research scientist at Analog Devices, Inc. in their machine learning group, where he worked on hardware accelerated Bayesian inference and audio processing. Since 2015 he is an assistant professor at Brigham Young University. His research interests include probabilistic programming, reinforcement learning, deep learning, and the intersection of hardware and machine learning.

Noah D. Stein (noah.stein@analog.com) is a senior research scientist at the Lyric Labs research unit of Analog Devices, Inc in Cambridge, MA. He received the B.S. degree in electrical and computer engineering from Cornell University in 2005 and the Ph.D. in electrical engineering from MITs Laboratory for Information and Decision Systems in 2011.

Noahs research at Lyric Labs focuses on algorithms for audio source separation with an emphasis on practical machine learning approaches for exploiting directional information as alternatives to beamforming. His doctoral research applied semidefinite programming and other algebraic methods in optimization to theoretical questions in game theory.

Paris Smaragdis (paris@illinois.edu) is an assistant professor in the Computer Science Department and the Electrical and Computer Science Department at the University of Illinois at Urbana-Champaign, as well as a senior research scientist at Adobe Research. Prior to that he was a research scientist at Mitsubishi Electric Research Labs, during which time he was selected by MIT Technology Review as one of the top 35 young innovators of 2006. His research interests lie in the intersection of machine learning and signal processing. He is a Senior Member of the IEEE.