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
Indeed, we show that the proposed method outperforms a sequential approach for automatic speech recognition [32] by unified approaches that have been shown to out-perform approach is shown in Fig. 1. This work is motivated in part by unified approaches 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:

\[ x_m[t] = a_m[t] * s[t] + n_m[t] , \]  

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 decouple the signal components across frequency and time. Thus, at time frame \( t \in [1, T] \) and frequency \( f \in [1, F] \), we have:

\[ x_{ft} = a_f s_{ft} + n_{ft} , \]  

where \( a_f \) explicitly in terms of the source DOA as the unit steering vector:

\[ a_f(\phi) = \frac{1}{\sqrt{M}} \exp \left[ j \frac{2\pi f_l}{c} m^\top \phi \right] , \]  

where \( \phi \in \mathbb{R}^3 \) is the source’s unit DOA vector, \( m \in \mathbb{R}^{1 \times M} \) is the matrix of \( M \) sensor positions, \( f_l \) 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:

\[ x_{ft} = A_f(\Phi) s_{ft} + n_{ft} , \]  

where \( \Phi = \{ \phi_1 : \ldots : \phi_K \} \) denotes the set of source DOAs, \( s_{ft} \in \mathbb{C}^K \) is a vector of source coefficients, and we have defined the steering matrix:

\[ A_f(\Phi) = [a_f(\phi_1) : \ldots : a_f(\phi_K)] \in \mathbb{C}^{M \times K} . \]  

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 \( w_f \in \mathbb{C}^M \) that is used to estimate a source coefficient \( s_{ft} \) via \( \hat{s}_{ft} = w_f^H x_{ft} \). The optimal \( w_f \) minimizes the expected output power of the filter:

\[ P = E [||\hat{s}_{ft}||^2] = w_f^H R_f w_f , \]  

where \( R_f \) is the channel correlation matrix:

\[ R_f = E [x_{ft} x_{ft}^H] . \]  

The solution is the well-known MVDR beamformer:
\[
\hat{w}_f^{\text{MVDR}} = \frac{\mathbf{R}_f^{-1} \mathbf{a}_f(\phi)}{\mathbf{a}_f^H(\phi) \mathbf{R}_f^{-1} \mathbf{a}_f(\phi)}.
\]

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

\[
\hat{w}_f^{\text{DS}} = \mathbf{a}_f(\phi),
\]

with corresponding empirical output power:

\[
P_{DS} = \mathbf{a}_f^H(\phi) \hat{\mathbf{R}}_f \mathbf{a}_f(\phi),
\]

where:

\[
\hat{\mathbf{R}}_f = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{ft} \mathbf{x}_{ft}^H,
\]

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:

\[
\hat{w}_f = \arg \min_{\mathbf{w}_f} \mathbf{w}_f^H \mathbf{R}_f \mathbf{w}_f, \quad \text{s.t.} \quad \mathbf{A}_f^H(\Phi) \mathbf{w}_f = \mathbf{u},
\]

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

\[
\hat{w}_f^{\text{LCMV}} = \mathbf{R}_f^{-1} \mathbf{A}_f(\Phi) \left[\mathbf{A}_f^H(\Phi) \mathbf{R}_f^{-1} \mathbf{A}_f(\Phi)\right]^{-1} \mathbf{u}.
\]

### 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:

\[
\log \mathcal{L}_{ft}(\Theta) = \log \mathcal{N}(\mathbf{x}_{ft}; \mathbf{\mu}_{ft}, \sigma_f^2 \mathbf{I}).
\]

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

\[
\mathbf{\mu}_{ft} = \mathbb{E} [\mathbf{x}_{ft}] = \mathbf{A}_f(\Theta) \mathbb{E}[s_{ft}|\mathbf{x}_{ft}],
\]

for a hypothesized DOA set \(\Theta\). We approximate the expectation with a least-squares estimate:

\[
\hat{\mathbf{s}}_{ft} = \left[\mathbf{A}_f^H(\Theta) \mathbf{A}_f(\Theta)\right]^{-1} \mathbf{A}_f^H(\Theta) \mathbf{x}_{ft},
\]

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

\[
\hat{\mathbf{\mu}}_{ft} = \mathbf{A}_f(\Theta) \hat{\mathbf{s}}_{ft} = \mathbf{B}_f(\Theta) \mathbf{x}_{ft},
\]

where:

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

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

\[
\log \mathcal{L}_{ft}(\Theta) \propto -\frac{1}{2\sigma_f^2} \mathbf{x}_{ft}^H \mathbf{P}_f(\Theta) \mathbf{x}_{ft},
\]

in terms of a precision matrix:

\[
\mathbf{P}_f(\Theta) = \mathbf{I} - \mathbf{B}_f(\Theta).
\]

Aggregating over all \(t\) and expanding, we have:

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

which, in the one-source case, simplifies to:
\[
\log \mathcal{L}_f (\theta) \propto - \frac{1}{2\sigma_f^2} \sum_{t=1}^{T} ||x_{ft}||^2_2 - |a^H_f (\theta) x_{ft}|^2 .
\]  
(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:

\[
p_{\text{MultiSRP}} (x_{ft}) = x^{H}_{ft} \mathbf{B} x_{ft} \\
= \frac{||A^H x_{ft}||^2_2 - 2 \text{Re} \left\{ (a^H_f a_1) (a^H_f x_{ft}) (x^H_{ft} a_2) \right\}}{1 - |a^H_2 a_1|^2} .
\]

(29)

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

\[
p_{\text{MoSRP}} (x_{ft}) = \sum_{k=1}^{2} x^{H}_{ft} a_k a^H_k x_{ft} = ||A^H x_{ft}||^2_2 .
\]

(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. \( x_{ft} = A1 = a (\theta) + a (-\theta) \). 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 \( \theta' \neq \pm \theta \) such that \( |a^H (\theta') x_{ft}|^2 \) 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):

\[
\Theta^i \leftarrow \Theta^{i-1} + \eta_i \Omega \left( \sum_{f=1}^{F} \frac{\partial \log \mathcal{L}_f (\Theta)}{\partial \Theta} \bigg|_{\Theta^{i-1}} \right),
\]

(31)

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

A. One-Source Model

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

\[
\frac{\partial \log \mathcal{L}_f (\Theta)}{\partial \Theta} = \frac{2 \pi f}{\sigma_f^2} c \text{ Im} \left[ a^* (\theta) \odot (R_f a_f (\theta)) \right],
\]

(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:

\( ^1\text{GEM is a variant of EM [35] that only requires the M step to increase the lower bound rather than explicitly maximize it.} \)

---

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 as well as a product of 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. (Bottom) Multi-source likelihood.
Algorithm 1 Robust MoSRP Localization

Preprocessing

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

Apply PHAT weighting: $X \leftarrow X \cdot |X|$

Initialize DOAs: $\theta^{(k)} \leftarrow \theta_0^{(k)}$

Optimization

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

Compute steering matrices: $A_f = \frac{1}{\sqrt{M}} e^{2\pi i f \pi \mathbf{m}^T \Phi}$

Compute projectors: $B_f = A_f A_f^H \frac{1}{A_f^H A_f}$

if Wiener mask then

Estimate TF weights: $\tilde{\gamma}_{ft} = \frac{\|B_f x_{ft}\|^2}{\|x_{ft}\|^2}$

else

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

end if

Compute posteriors: $\log u_{ft}^{(k)} \propto \frac{|\mathbf{a}_f^H (\theta_k) x_{ft}|^2}{2\sigma_f^2}$

Estimate correlation matrices:

$\tilde{R}_{f}^{(k)} = \sum_{t=1}^{T} \tilde{\gamma}_{ft} u_{ft}^{(k)} x_{ft} x_{ft}^H$

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

$m \text{Im} \left[ A_f^\dagger \odot \left( I - B_f \right) \tilde{R}_f A_f \left[ A_f^H A_f \right]^{-1} \right]$ Normalize source gradients: $g^{(k)} \leftarrow g^{(k)} / \|g^{(k)}\|_2$

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

Update DOA vectors: $\theta^{(k)} \leftarrow \theta^{(k)} + \eta_i g^{(k)}$

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

end for

Algorithm 2 Robust MultSRP Localization

Preprocessing

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

Apply PHAT weighting: $X \leftarrow X \cdot |X|$

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

Initialize DOAs: $\Theta \leftarrow \Theta_0$

Optimization

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

Compute steering matrices: $A_f = \frac{1}{\sqrt{M}} e^{2\pi i f \pi \mathbf{m}^T \Phi}$

Compute projectors: $B_f = A_f A_f^H \frac{1}{A_f^H A_f}$

if Wiener mask then

Estimate TF weights: $\tilde{\gamma}_{ft} = \frac{\|B_f x_{ft}\|^2}{\|x_{ft}\|^2}$

Estimate correlation matrices:

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

end if

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

$m \text{Im} \left[ A_f^\dagger \odot \left( I - B_f \right) \tilde{R}_f A_f \left[ A_f^H A_f \right]^{-1} \right]$ Normalize source gradients: $g^{(k)} \leftarrow g^{(k)} / \|g^{(k)}\|_2$

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

Update DOA matrix: $\Theta \leftarrow \Theta + \eta_i G$

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

end for

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

$$Q \propto \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{1}{2\sigma_f^2} \left[ z_{ft} = k \right] |\mathbf{a}_f^H (\theta_k) x_{ft}|^2,$$

(33)

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

$$\log p(\mathbf{X}, \mathbf{Z} | \Theta) \propto \sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{1}{2\sigma_f^2} \left[ z_{ft} = k \right] |\mathbf{a}_f^H (\theta_k) x_{ft}|^2,$$

(34)

$$\tilde{R}_{f}^{(k)} = \frac{\sum_{t=1}^{T} u_{ft}^{(k)} x_{ft} x_{ft}^H}{\sum_{t=1}^{T} u_{ft}^{(k)}},$$

(35)

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):

$$\frac{\partial}{\partial \Theta} \log \mathcal{L}_f (\Theta) = \frac{1}{\sigma_f^2} \sum_{f=1}^{P} l_f \cdots$$

$$m \text{Im} \left[ A_f^\dagger \odot \left( I - B_f \right) \tilde{R}_f A_f \left[ A_f^H A_f \right]^{-1} \right],$$

(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 $e_{ft}$ is present, we have:

$$x_{ft} = A_f (\Phi) s_{ft} + n_{ft} + e_{ft} = b_{ft} + c_{ft},$$

(38)

where $b_{ft} = A_f (\Phi) s_{ft}$ and $c_{ft} = n_{ft} + e_{ft}$. The Wiener mask [25] gives the MMSE-optimal weighting to recover $b_{ft}$:
Thus, a robust estimate of the correlation matrices is:

$$\hat{R}_f = \frac{\sum_{i=1}^{T} \gamma_{ft} \mathbf{x}_f \mathbf{x}_f^H}{\sum_{i=1}^{T} \gamma_{ft}} .$$

(40)

In practice, we approximate the weights using:

$$\mathbf{\hat{b}}_{ft} = \mathbf{B}_f (\Theta) \mathbf{x}_f ,$$

(41)

$$\mathbf{\hat{c}}_{ft} = \mathbf{x}_f - \mathbf{\hat{b}}_{ft} ,$$

(42)

which gives:

$$\gamma_{ft} = \frac{\|\mathbf{\hat{b}}_{ft}\|^2_2}{\|\mathbf{\hat{b}}_{ft}\|^2_2 + \|\mathbf{\hat{c}}_{ft}\|^2_2} = \frac{\|\mathbf{\hat{b}}_{ft}\|^2_2}{\|\mathbf{x}_f\|^2_2} .$$

(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:

$$\log w^{(k)}_{ft} \propto \frac{1}{2\nu^2} | \tilde{s}^{(k)}_{ft} |^2 ,$$

(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:

$$\tilde{s}_{ft} = \left[ \mathbf{A}^H_f \mathbf{A}_f \right]^{-1} \mathbf{A}^H_f \mathbf{x}_f .$$

(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, $10 \times 10$ 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: $\forall j \neq k \cos^{-1} \left( \phi_j^T \phi_k \right) \geq \frac{2\pi}{K+2}$. 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.

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:

$$e = \min_{\mathcal{P}} \frac{1}{K} \sum_{k=1}^{K} \cos^{-1} \left( \phi_k^T \hat{\theta}_{\mathcal{P}(k)} \right)$$

(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.

Source separations 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 $B_f$ 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 $5 \times 5 \times 5$ 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 $e = 0.71$.
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 $x_{ft}$ (dependence on $f$ and $t$ are omitted for clarity):

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

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

$$s \sim \mathcal{N} \left( \mu_s, \Sigma_s \right) , \quad x|s \sim \mathcal{N} \left( As, \Sigma_n \right) . \quad (48)$$

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

$$s|x \sim \mathcal{N} \left( \mu_{s|x}, \Sigma_{s|x} \right) , \quad (49)$$

$$\mu_{s|x} = \Sigma_{s|x} \left( A^H \Sigma_n^{-1} A + \Sigma_s^{-1} \right)^{-1} A^H \Sigma_n^{-1} A + \Sigma_s^{-1} . \quad (50)$$

Assuming a vague prior (i.e. $\Sigma_s \rightarrow \infty I$), we have:

$$\hat{\mu} = E[x] = E[As + n] = A \mu_s + B x , \quad (52)$$

$$\hat{\Sigma} = E \left[ (x - \hat{\mu})(x - \hat{\mu})^H \right] \quad (53)$$

$$= E \left[ \left( A s + n - A \mu_s \right) \left( A s + n - A \mu_s \right)^H \right] \quad (54)$$

$$= E \left[ (A (s - \mu_s) + n) (A (s - \mu_s) + n)^H \right] \quad (55)$$

$$= A \Sigma_{s|\mu} A^H + \Sigma_n \quad (56)$$

$$= (B + I) \Sigma_n , \quad (57)$$

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

$$\mathcal{L} = -\frac{1}{2} \log \left| 2\pi \Sigma_n \right| - \frac{K}{2} \log 2 - \frac{1}{2} x^H P^H \Sigma_n^{-1} P x \quad , \quad (58)$$

where $P = I - B$. Finally, if $\Sigma_n = \sigma^2 I$, we have:

$$\mathcal{L} \propto -\frac{1}{2\sigma^2} x^H P x , \quad (59)$$
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 \(x\) is:

\[
\mathcal{L} \propto \frac{1}{2\sigma^2} x^H B x ,
\]

(60)

where \(B = AA^H\) is invertible and:

\[
A_{iv} = \frac{1}{\sqrt{M}} e^{j\pi f_i} \sum_{m} m_{mv} \Theta_{uv} .
\]

(61)

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

\[
P^{(1)} = x^H Ay ,
\]

(62)

\[
P^{(2)} = z^H [A^H A]^{-1} z ,
\]

(63)

\[
P^{(3)} = y^H A^H x ,
\]

(64)

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

\[
\frac{\partial P^{(1)}}{\partial \Theta_{uv}} = \frac{1}{c} \sum_{i} m_{ai} x_i^* A_{iv} y_v ,
\]

(65)

we can write:

\[
\frac{\partial P^{(1)}}{\partial \Theta} = j \frac{2\pi f_i}{c} \text{Im} \left( A^* \right) \text{Im} \left( y^* \right) ,
\]

(66)

\[
\frac{\partial P^{(2)}}{\partial \Theta} = -j \frac{2\pi f_i}{c} \text{Im} \left( A^* \right) \text{Im} \left( x^* \right) ,
\]

(67)

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

\[
\frac{\partial P^{(2)}}{\partial \Theta} = z^H \left( \frac{\partial [A^H A]^{-1}}{\partial \Theta} \right) z
\]

(68)

\[
= -z^H [A^H A]^{-1} \left( \frac{\partial A^H A}{\partial \Theta} \right) [A^H A]^{-1} z
\]

(69)

\[
= -y^H \left[ \left( \frac{\partial A}{\partial \Theta} \right)^H A + A^H \left( \frac{\partial A}{\partial \Theta} \right) \right] y
\]

(70)

\[
= - \frac{\partial y^H A^H q}{\partial \Theta} + \frac{\partial q^H A y}{\partial \Theta}
\]

(71)

\[
= -j \frac{2\pi f_i}{c} \text{Im} \left( A^* \right) \text{Im} \left( y^* \right) ...
\]

(72)

\[
= \frac{\partial P}{\partial \Theta} = \sum_{i=1}^{3} \frac{\partial P^{(i)}}{\partial \Theta}
\]

(73)

\[
= 4\pi f_i \frac{c}{m} \left( \text{Im} \left[ \text{diag} \left( x^* \right) A \text{diag} \left( y \right) \right] \cdots
\]

(74)

\[
+ \text{Im} \left[ \text{diag} \left( q^* \right) A \text{diag} \left( y \right) \right] \right)
\]

\[
= 4\pi f_i \frac{c}{m} \left( -\text{Im} \left[ A \circ \left( x^* y^T \right) \right] + \text{Im} \left[ A \circ \left( q^* y^T \right) \right] \right)
\]

(75)

\[
= 4\pi f_i \frac{c}{m} \left( -\text{Im} \left[ A \circ \left( y x^H \right)^T \right] + \text{Im} \left[ A \circ \left( y q^H \right)^T \right] \right)
\]

(76)

Aggregating over time and consolidating terms, we have:

\[
\frac{\partial P}{\partial \Theta} = 4\pi f_i \frac{c}{m} \text{Im} \left[ A^* \circ \left( PC^H \right) \right]
\]

(77)

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

\[
\frac{\partial L}{\partial \Theta} = \frac{1}{\sigma^2} \frac{\partial P}{\partial \Theta} = \frac{1}{\sigma^2} 2\pi f_i \frac{c}{m} \text{Im} \left[ A^* \circ \left( PC^H \right) \right]
\]

(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:

\[
[A^H A]^{-1} = \frac{1}{d} \begin{bmatrix}
-1 & -a_1^H a_2^T \\
-1 & 1
\end{bmatrix}
\]

(79)

where \(d = 1 - |a_2^T a_2|^2\). Substituting this into (78), expanding, and re-arranging, we have:

\[
\frac{\partial L}{\partial \Theta} = \frac{1}{\sigma^2} \frac{2\pi f_i}{c} m \text{Im} \left[ z_{11} - v_{21} z_{12} + y , \ z_{22} - v_{12} z_{21} + y \right]
\]

(80)

where:

\[
y = \frac{r_{12} v_{21} - (r_{11} + r_{22}) v_{21} + r_{21} a_i^T \circ a_2}{d}
\]

(81)

\[
q_i = R a_i
\]

(82)

\[
z_{ij} = a_i^T \circ q_j
\]

(83)

\[
v_{ij} = a_i^T a_j
\]

(84)

\[
r_{ij} = a_i^T R a_j = a_i^T q_j
\]

(85)

The likelihood can be computed efficiently as:

\[
\mathcal{L} = -\frac{1}{\sigma^2} \left( -||X||_F^2 + \frac{r_{11} + r_{22} - 2 \text{Re} \left[ v_{21} v_{12}^* \right]}{d} \right)
\]

(86)

REFERENCES


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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MIT with a joint appointment in BCS and the Laboratory for Information and Decision Systems. From 2012-2015 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 MI...