MULTIDIMENSIONAL LOCALIZATION OF MULTIPLE SOUND SOURCES USING FREQUENCY DOMAIN ICA AND AN EXTENDED STATE COHERENCE TRANSFORM

Benedikt Loesch, Stefan Uhlich and Bin Yang

Chair of System Theory and Signal Processing, University of Stuttgart
Email: {benedikt.loesch, stefan.uhlich, bin.yang}@LSS.uni-stuttgart.de

ABSTRACT

Recently, direction-of-arrival (DOA) and position estimation for acoustic signals has been studied intensively and many different algorithms have been proposed. Among different solutions for multiple sources, blind source separation (BSS) methods have drawn much attention due to their good performance. In this paper, we present a localization algorithm using the results from a frequency domain independent component analysis (ICA) algorithm combined with an extended version of the state coherence transform (SCT). We motivate the SCT as an approximated maximum likelihood (ML) approach and compare our localization algorithm with the steered-response power with phase transform (SRP-PHAT) and the averaged directivity pattern (BSS-ADP) algorithm. 2D localization results show superior performance of our algorithm.

Index Terms— direction-of-arrival estimation, source localization, blind source separation, state coherence transform

1. INTRODUCTION

The task of acoustic source localization is to estimate the position of one or multiple sound sources by using an array of microphones. There are non-BSS based approaches such as SRP-PHAT [1], DATEMM [2], and BSS based approaches such as [3], [4], [5]. In this paper, we use BSS to estimate the propagation model first. Using this propagation model, we then extract location information such as 1D/2D DOA or 2D/3D position.

In contrast to [3], our algorithm does not suffer from spatial aliasing. Different from [4], our system can perform localization directly without an intermediate time-difference of arrival (TDOA) estimation and without a spatial ambiguity resolver. [5] proposed to use the averaged directivity pattern (ADP) of the BSS solution to estimate 2D DOA. In this way, there is no need for a spatial ambiguity resolver. Our approach is similar since the SCT compares an assumed propagation model against the estimated one. However, in contrast to [5], our algorithm works in the frequency domain and can handle not only DOA but also source position estimation. It shows improved localization performance.

2. FREQUENCY DOMAIN ICA

The goal of blind source separation is to separate M convolutive mixtures $x_m[i], m = 1, \ldots, M$ into N statistically independent source signals. Mathematically, we write the sensor signals $x_m[i]$ as a sum of convolved source signals

$$x_m[i] = \sum\nolimits_{n=1}^{N} h_{mn}[i] * s_n[i], \quad m = 1 \ldots M. \tag{1}$$

Our goal is to find signals $y_n[i], n = 1 \ldots N$ such that, after solving the permutation ambiguity, $y_n[i] \approx g_n[i] * s_n[i]$. We focus on frequency domain ICA because of the lower computational complexity and better convergence properties than time domain approaches. The convolutive mixture in the time domain can be approximated as an instantaneous mixture in the frequency domain:

$$\mathbf{X}[k, l] \approx \mathbf{H}[k]\mathbf{S}[k, l] \tag{2}$$

The authors would like to thank Francesco Nesta for fruitful discussions.

$1 \leq k \leq K$ is the frequency bin index and l is the time frame index. For the purpose of ICA, we assume an identical number of sources and sensors $M = N$. $\mathbf{H}[k]$ is a square mixing matrix. We can then apply any frequency domain ICA algorithm such as the scaled infomax algorithm from [6] to separate the signals in the time-frequency domain:

$$\mathbf{Y}[k, l] = \mathbf{W}[k]\mathbf{X}[k, l] \tag{3}$$

$\mathbf{W}[k]$ is a square demixing matrix. The scaled infomax algorithm shows a fast convergence for a wide range of step sizes and regardless of the scaling of the source signals. The separation is done by maximizing the entropy of the output signals $\mathbf{Y}[k, l]$ or equivalently minimizing their mutual information. Each frequency bin is treated independently from the others and hence a permutation and scaling problem occurs at each frequency bin. To achieve a good separation, both problems need to be solved. Many different approaches have been proposed, but most of them do not work reliably under adverse conditions such as low signal-to-noise ratio (SNR), high amount of reverberation and small sample size. The approach in [7] solves the permutation problem by using a recursive initialization of $\mathbf{W}[k]$ with a smoothed version of the demixing matrices at previous frequencies and a permutation correction step with the SCT. It jointly considers all frequency bins and hence allows wide microphone spacings. We briefly summarize the SCT in the next section before we provide an ML interpretation of SCT, extend its idea to multiple microphone pairs, and apply it to the problem of localization.

3. STATE COHERENCE TRANSFORM

The main idea of the state coherence transform is to compare the "state" $e^{-j\omega\tau}$ from the propagation model against its estimate from the result of ICA. We first define the state

$$r_{ab}[k, \mathbf{p}] = e^{-j\omega_k \tau_{ab}(\mathbf{p})} \tag{4}$$

where ω_k is the angular frequency in frequeny bin k and $\tau_{ab}(\mathbf{p})$ is the TDOA of a source at position $\mathbf{p}$ observed at the microphone pair (a, b) located at $\mathbf{d}_a$ and $\mathbf{d}_b$:

$$\begin{aligned}
\text{near-field (2D/3D position): } \mathbf{p} &= [x, y]^T \quad \text{or} \quad \mathbf{p} = [x, y, z]^T, \\
\tau_{ab}(\mathbf{p}) &= c^{-1}(\|\mathbf{d}_a - \mathbf{p}\| - \|\mathbf{d}_b - \mathbf{p}\|) \\
\text{far-field (1D/2D angle): } \mathbf{p} &= [\sin\theta, \cos\theta]^T, \quad \text{or} \\
\mathbf{p} &= [\sin\theta\cos\phi, \cos\theta\cos\phi, \sin\phi]^T \\
\tau_{ab}(\mathbf{p}) &= c^{-1}(\mathbf{d}_a - \mathbf{d}_b)^T\mathbf{p}
\end{aligned} \tag{5}$$

c is the sound propagation speed.

If we assume that the microphone pair (a, b) has a small enough distance, the impulse responses from the source to both microphones will look similar up to a delay τ_{ab} and an amplitude scaling. In terms of the frequency response $H_i[k](i = a, b)$ from the source to both microphones, we have

$$\frac{H_a[k]}{H_b[k]} = \frac{|H_a[k]|}{|H_b[k]|} e^{-j\omega_k \tau_{ab}(\mathbf{p})}. \tag{6}$$

Note that the impulse responses can be arbitrary as long as they are

similar up to a delay and a scaling. Hence our signal model is much more general than the typical assumption of a dominant path. Generally, we could also use the amplitude ratio $\frac{|H_a[k]|}{|H_b[k]|}$ for localization, but it is much less reliable than the TDOA, especially in reverberant environments and when the sources and microphones have directivity patterns. A comparison with (4) shows

$$r_{ab}[k,\mathbf{p}] = \exp\left\{j \arg\left(\frac{H_a[k]}{H_b[k]}\right)\right\} \tag{7}$$

Now we estimate this state from the results of ICA

$$\hat{r}_{ab,n}[k] = \exp\left\{j \arg \frac{\left[\mathbf{W}^{-1}[k]\right]_{an}}{\left[\mathbf{W}^{-1}[k]\right]_{bn}}\right\} \tag{8}$$

The notation $[\mathbf{A}]_{ij}$ denotes the (i,j)-th element of the matrix $\mathbf{A}$. Assuming a quite successfull blind source separation, we obtain

$$\mathbf{W}[k] \approx \mathbf{\Pi}[k]\mathbf{D}[k]\mathbf{H}^{-1}[k] \text{ or } \mathbf{H}[k] \approx \mathbf{W}^{-1}[k]\mathbf{\Pi}[k]\mathbf{D}[k]. \tag{9}$$

$\mathbf{D}[k]$ is a diagonal complex-valued scaling matrix and $\mathbf{\Pi}[k]$ is a permutation matrix. Since $\mathbf{D}[k]$ is diagonal, the ratio of elements in (8) is invariant with respect to $\mathbf{D}[k]$:

$$\frac{\left[\mathbf{W}^{-1}[k]\right]_{a\Pi(n)}}{\left[\mathbf{W}^{-1}[k]\right]_{b\Pi(n)}} \approx \frac{[\mathbf{H}[k]]_{an}}{[\mathbf{H}[k]]_{bn}}. \tag{10}$$

This implies

$$\hat{r}_{ab,\Pi(n)}[k] \approx r_{ab}[k,\mathbf{p}_n], \tag{11}$$

where $\mathbf{p}_n$ is the position of the source associated with column n of $\mathbf{H}[k]$. $\Pi(n)$ describes the permutation, i.e. the column n in $\mathbf{H}[k]$ corresponds to the column $1 \le \Pi(n) \le N$ in $\mathbf{W}^{-1}[k]$[1]. This means $[\mathbf{\Pi}[k]]_{n\Pi(n)} = 1$.

[7] proposed to use the states $\hat{r}_{ab,n}[k]$ from all frequency bins and all columns of $\mathbf{W}^{-1}[k]$ to solve the permutation problem in frequency domain ICA. It defines a so called state coherence transform (SCT) for sensor pair (a,b)

$$\mathrm{SCT}(\tau) = \sum_{n=1}^{N}\sum_{k=1}^{K} \rho(|\hat{r}_{ab,n}[k] - r_{ab}[k,\tau]|) \tag{12}$$

with $\rho(t) = 1 - \tanh(\alpha t/2)$. $\mathrm{SCT}(\tau)$ has maxima for $\tau = \tau_{ab}(\mathbf{p}_n)$ if we choose a large enough value of $\alpha > 0$. By looking for maxima in $\mathrm{SCT}(\tau)$, we can estimate the TDOA of the sources. In [7], (12) was derived heuristically and uses only one sensor pair. Below we will provide an ML motivation of SCT and apply it to localization using multiple sensor pairs.

4. OUR LOCALIZATION ALGORITHM BSS-SCT

Taking more than just one sensor pair into account, we define a state column vector and a corresponding estimate for certain sensor pairs $(a,b) \in \mathcal{I} \subseteq \{(a,b)|1 \le a < b \le M\}$:

$$\begin{aligned}\mathbf{r}[k,\mathbf{p}] &= [r_{ab}[k,\mathbf{p}]]_{(a,b)\in\mathcal{I}} \\ \hat{\mathbf{r}}_n[k] &= [\hat{r}_{ab,n}[k]]_{(a,b)\in\mathcal{I}}\end{aligned} \tag{13}$$

We introduce the following model for $\hat{\mathbf{r}}_n[k]$:

$$\hat{\mathbf{r}}_{\Pi(n)}[k] = \mathbf{r}[k,\mathbf{p}_n] + \mathbf{v}_n[k] \tag{14}$$

$\mathbf{v}_n[k]$ is the noise with an unknown probability density function (pdf). Considering all N columns of $\mathbf{W}^{-1}[k]$, we obtain:

$$\begin{aligned}&\hat{\mathbf{R}}[k]\mathbf{\Pi}[k] = \mathbf{R}[k,\mathbf{P}] + \mathbf{V}[k] \qquad \text{with} \\ &\hat{\mathbf{R}}[k] = [\hat{\mathbf{r}}_1[k],\cdots,\hat{\mathbf{r}}_n[k]],\ \mathbf{R}[k] = [\mathbf{r}[k,\mathbf{p}_1],\cdots,\mathbf{r}[k,\mathbf{p}_n]], \\ &\mathbf{V}[k] = [\mathbf{v}_1[k],\cdots,\mathbf{v}_N[k]],\ \mathbf{P} = [\mathbf{p}_1,\cdots,\mathbf{p}_N].\end{aligned} \tag{15}$$

$\mathbf{\Pi}[k]$ is a permutation matrix at frequency bin k.

[1] If the number of sources N is smaller than the number of sensors M, some of the estimated states $\hat{r}_{ab,n}[k]$ will not correspond to a true source. However, this is not a problem since these states will not be coherent across the frequency and hence will be suppressed in the SCT.

Combining this model for all frequencies, we get

$$\begin{bmatrix}\hat{\mathbf{R}}[1]\\ \vdots \\ \hat{\mathbf{R}}[k]\end{bmatrix} \cdot \begin{bmatrix}\mathbf{\Pi}[1] & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Pi}[K]\end{bmatrix} = \begin{bmatrix}\mathbf{R}[1]\\ \vdots \\ \mathbf{R}[K]\end{bmatrix} + \begin{bmatrix}\mathbf{V}[1]\\ \vdots \\ \mathbf{V}[K]\end{bmatrix},$$

$$\tilde{\hat{\mathbf{R}}} \cdot \tilde{\mathbf{\Pi}} = \tilde{\mathbf{R}}(\mathbf{P}) + \tilde{\mathbf{V}} \tag{16}$$

Let $f_{\tilde{\mathbf{V}}}(\cdot)$ be the pdf of $\tilde{\mathbf{V}}$. The likelihood function of $\tilde{\hat{\mathbf{R}}}$ is then $f_{\tilde{\mathbf{V}}}(\tilde{\hat{\mathbf{R}}}\tilde{\mathbf{\Pi}} - \tilde{\mathbf{R}}(\mathbf{P}))$. Hence, the ML estimate of $\mathbf{P}$ and $\tilde{\mathbf{\Pi}}$ is

$$\arg\max_{\mathbf{P},\tilde{\mathbf{\Pi}}} f_{\tilde{\mathbf{V}}}(\tilde{\hat{\mathbf{R}}}\tilde{\mathbf{\Pi}} - \tilde{\mathbf{R}}(\mathbf{P})) \tag{17}$$

To facilitate a simpler estimation, we assume independence of $\hat{\mathbf{r}}_n[k]$ among different sources n and frequency bins k. Then we can write

$$f_{\tilde{\mathbf{V}}} = \prod_{n=1}^{N}\prod_{k=1}^{K} f_{\mathbf{v}_n[k]}(\hat{\mathbf{R}}[k]\mathbf{\Pi}_n[k] - \mathbf{r}[k,\mathbf{p}_n]) \tag{18}$$

where $\mathbf{\Pi}_n[k]$ is the n-th column of $\mathbf{\Pi}[k]$. Equivalently we maximize

$$\ln f_{\tilde{\mathbf{V}}} = \sum_{n=1}^{N}\sum_{k=1}^{K} \ln f_{\mathbf{v}_n[k]}(\hat{\mathbf{R}}[k]\mathbf{\Pi}_n[k] - \mathbf{r}[k,\mathbf{p}_n]). \tag{19}$$

However, this joint multiple-source localization would be very challenging due to the dimensionality of the problem and the fact that we need an extensive search for the discrete valued permutation matrices $\mathbf{\Pi}[k]$. Furthermore, the pdf of the noise $\mathbf{v}_n[k]$ is unknown and we cannot derive the ML estimator of the source parameters $\mathbf{p}_n$ analytically.

An approximation of the ML solution is to assume that $\mathbf{v}_n[k]$ has a spherical pdf. Then $f_{\mathbf{v}_n[k]}(\mathbf{v})$ is only a function of $\|\mathbf{v}\|$. By assuming identical pdf of $\mathbf{v}_n[k]$ for different frequency bins and different sources, we can approximate $\ln f_{\tilde{\mathbf{V}}}$

$$\ln f_{\tilde{\mathbf{V}}} \approx \mathcal{H}(\mathbf{P}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \rho(\|\hat{\mathbf{R}}[k]\mathbf{\Pi}_n[k] - \mathbf{r}[k,\mathbf{p}_n]\|) \tag{20}$$

where $\rho(t)$ is a so called Kernel function which is monotonically decreasing for $t \ge 0$.

If we knew the permutation $\mathbf{\Pi}[k]$ or equivalently if there was no permutation with $\mathbf{\Pi}[k] = \mathbf{I}$, the multiple-source localization $\max_{\mathbf{P}} \mathcal{H}(\mathbf{P})$ could be simplified to N independent single-source localization problems. The ML estimates would then be given by

$$\hat{\mathbf{p}}_n = \arg\max_{\mathbf{p}_n} \sum_{k=1}^{K} \rho(\|\hat{\mathbf{r}}_n[k] - \mathbf{r}[k,\mathbf{p}_n]\|). \quad (1 \le n \le N) \tag{21}$$

However, due to the unknown permutation $\mathbf{\Pi}[k]$, we do not know which state estimate $\hat{\mathbf{r}}_n[k]$ belongs to which source and hence cannot estimate each source position $\mathbf{p}_n$ individually. A suboptimal but easily feasible solution of this problem is to exploit the fact that estimated states $\hat{\mathbf{r}}[k]$ that belong to source n will lie in proximity to the ideal state $\mathbf{r}[k,\mathbf{p}_n]$ while states $\hat{\mathbf{r}}[k]$ belonging to a different source will lie further away from $\mathbf{r}[k,\mathbf{p}_n]$.

By using a locally-confined kernel function $\rho(t)$ that puts more weight on $t \approx 0$, we can implicitly resolve the permutation. In this case, we replace (20) for multiple sources by

$$\mathcal{H}_{\text{BSS-SCT}}(\mathbf{p}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \rho(\|\hat{\mathbf{r}}_n[k] - \mathbf{r}[k,\mathbf{p}]\|) \tag{22}$$

for a single source at $\mathbf{p}$. This new cost function still has maxima at the correct locations $\mathbf{p} = \mathbf{p}_n$ if we select the width of the kernel function $\rho(t)$ narrow enough. Note that (22) contains two approximations of the original joint multiple-source localization cost function (19):

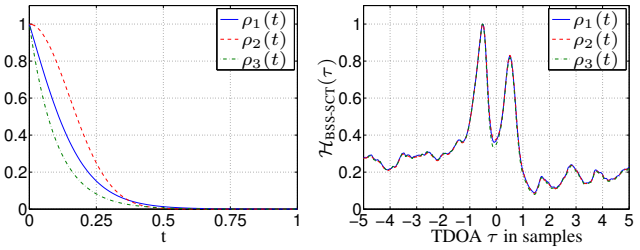
- Approximate $\ln f_{\mathbf{v}_n[k]}(\mathbf{v})$ by a locally-confined kernel function $\rho(\|\mathbf{v}\|)$.

- Replace the joint multiple-source search over $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_N]$ by a sequential single-source search over $\mathbf{p}$ in order to avoid the resolution of permutation.

Examples of kernel functions are

$$\rho_1(t) = 1 - \tanh(\alpha t/2), \quad \rho_2(t) = e^{-\frac{t^2}{2\sigma^2}},$$
$$\rho_3(t) = e^{-\frac{t}{\beta}}, \qquad (t \geq 0). \tag{23}$$

It turns out that the exact shape of the kernel function is not important, as long as it is locally confined. Fig. 1(a) shows different kernel functions from (23) with $\alpha = 10$, $\sigma = 0.15, \beta = 0.1$. Fig. 1(b) shows the values of $\mathcal{H}_{\text{BSS-SCT}}(\tau)$ in (22) using these kernel functions. We have used the TDOA τ as the parameter of the SCT as in [7]. We considered a scenario with $N = M = 2$ and $T_{60} = 700\,\text{ms}$. The results look almost identical for different kernel functions.



(a) Kernel functions $\rho(t)$ (b) SCT

Fig. 1: Effect of different kernel functions

5. COMPARISON WITH OTHER METHODS

5.1. Comparison with SRP-PHAT

In this section, we want to compare our proposed algorithm BSS-SCT with the well-known SRP-PHAT method [1]. The generalized cross correlation with phase transform (GCC-PHAT) for each microphone pair (a, b) is defined as

$$c_{ab}[\tau] = \text{IFFT}\left\{\frac{X_a[k]X_b^*[k]}{|X_a[k]X_b[k]|}\right\} \tag{24}$$

with $X_a[k] = \text{FFT}\{x_a[i]\}$ and $X_b[k] = \text{FFT}\{x_b[i]\}$. The SRP-PHAT method combines the GCC-PHAT of all $|\mathcal{I}|$ microphone pairs in $\mathcal{I}$ evaluated at the theoretical TDOA $\tau_{ab}(\mathbf{p})$:

$$\mathcal{H}_{\text{SRP-PHAT}}(\mathbf{p}) = \sum_{(a,b)\in\mathcal{I}} |c_{ab}[\tau_{ab}(\mathbf{p})]| \tag{25}$$

If there is only one source, we can estimate its position by maximizing $\mathcal{H}_{\text{SRP-PHAT}}(\mathbf{p})$ over all potential source positions $\mathbf{p}$. In this case, the height of the peak is independent of the signal amplitude, since in (24) $X_a[k]X_b^*[k]$ is normalized by $|X_a[k]X_b[k]|$. However, if we have more than one source, $\mathcal{H}_{\text{SRP-PHAT}}(\mathbf{p})$ will show multiple peaks with different heights. Since $X_a[k]$ in (24) does not represent the individual source spectrum, but rather the mixture spectrum, the height of the peaks of $\mathcal{H}_{\text{SRP-PHAT}}(\mathbf{p})$ will depend on the power of each source.

5.2. Comparison with BSS-ADP

[5] proposed the averaged directivity patterns (ADP) for 2D DOA estimation. This principle is based on the obvservation that BSS forms spatial nulls to the position of the unwanted sources in order to suppress them. In the context of localization, the spatial nulls are interpreted to point to the source positions. Strictly speaking, this holds only in anechoic environments, but we can also apply it in reverberant rooms when the direct path is dominant. We calculate the directivity pattern for each source n at position $\mathbf{p}$ by calculating the squared amplitude response of the demixing filter $w_{ni}[k]$:

$$B_n[k, \mathbf{p}] = \left|\sum_{i=1}^{N} w_{ni}[k] e^{-j\omega_k \tau_i(\mathbf{p})}\right|^2 \tag{26}$$

Each directivity pattern has $N-1$ spatial nulls or minima and hence would allow to localize $N-1$ sources. If we average the directivity patterns for all outputs $n = 1 \ldots N$, except for that directivity pattern with the highest amplitude [5]

$$n^*[k, \mathbf{p}] = \arg\max_n B_n[k, \mathbf{p}]$$
$$B(\mathbf{p}) = \sum_{k=1}^{K} \sum_{n \neq n^*[k,\mathbf{p}]} B_n[k, \mathbf{p}], \tag{27}$$

we get the BSS-ADP which has minima at the positions of the N sources. In order to compare this approach with SRP-PHAT and BSS-SCT, we define the cost function to be maximized as

$$\mathcal{H}_{\text{BSS-ADP}}(\mathbf{p}) = 1 - \frac{B(\mathbf{p})}{\max_{\mathbf{p}} B(\mathbf{p})} \tag{28}$$

Our BSS-SCT uses a similar idea by comparing the propagation model against its estimate from BSS. However, the SCT operates on the inverse demixing matrix $\mathbf{W}^{-1}[k]$ instead of the demixing matrix $\mathbf{W}[k]$.

6. EXPERIMENT

We used a regular office room of size $4.9\,\text{m} \times 3.5\,\text{m} \times 3\,\text{m}$ with $T_{60} = 450\,\text{ms}$ for 2D localization. The average SNR for all experiments was 15 to 20 dB. Fig. 2 shows the room layout with the

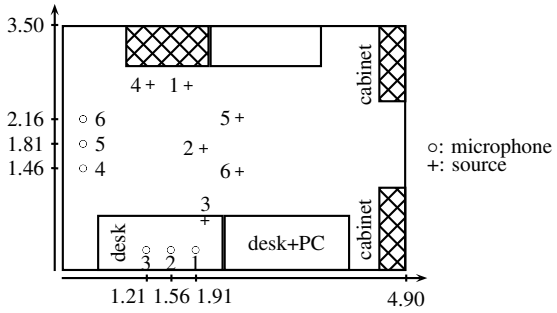


Fig. 2: Room layout and experimental setup

experimental setup. We played back 3 s of speech on loudspeakers with membrane diameter of 8 cm.We compare our proposed algorithm BSS-SCT with SRP-PHAT and BSS-ADP. We used a sampling frequency of $f_s = 48\,\text{kHz}$ for SRP-PHAT and $f_s = 16\,\text{kHz}$ for the two BSS based approaches. The search grid step for all three methods is 1 cm.

To evaluate the localization performance, we normalized each cost function $\mathcal{H}(\mathbf{p})$ to the range of $[0, 1]$ and performed a 2D peak search. Fig. 3 compares the normalized cost functions $\mathcal{H}_{\text{SRP-PHAT}}(\mathbf{p})$, $\mathcal{H}_{\text{BSS-ADP}}(\mathbf{p})$, $\mathcal{H}_{\text{BSS-SCT}}(\mathbf{p})$ for three cases of N sources and M microphones where both BSS-ADP and BSS-SCT work well. The dark regions represent hyperbola of possible source locations for a microphone pair. We used microphones 1, 2, 5, 6 for $M = 4$ and all 6 microphones for $M = 6$. For BSS-SCT, we used the microphone pairs $(1,2), (5,6), (1,5), (1,6)$ for $M = 4$ and $(1,2), (1,3), (4,5), (4,6), (1,4), (1,6)$ for $M = 6$. Some microphone pairs contain rather closely spaced microphones (35 cm) for a unique localization and some other pairs contain widely spaced microphones ($> 100\,\text{cm}$) for an accurate localization. For SRP-PHAT we used all available sensor pairs. Since all three cost functions exhibit many local maxima for a fine search grid, we perform the peak search in the following way: In each plot of Fig. 3, we marked the true source positions with □ and the 10 highest peaks with ×. Each peak is enumerated according to its height in descending order. If there are multiple peaks closer than 20 cm, we keep the strongest and discard the others. In the top right plot in Fig. 3, for example, peak 2 is close to peak 1, peaks 4 to 8 are close to peak 1 or 3, and peak 10 is close to peak 1 or 3 or 9.

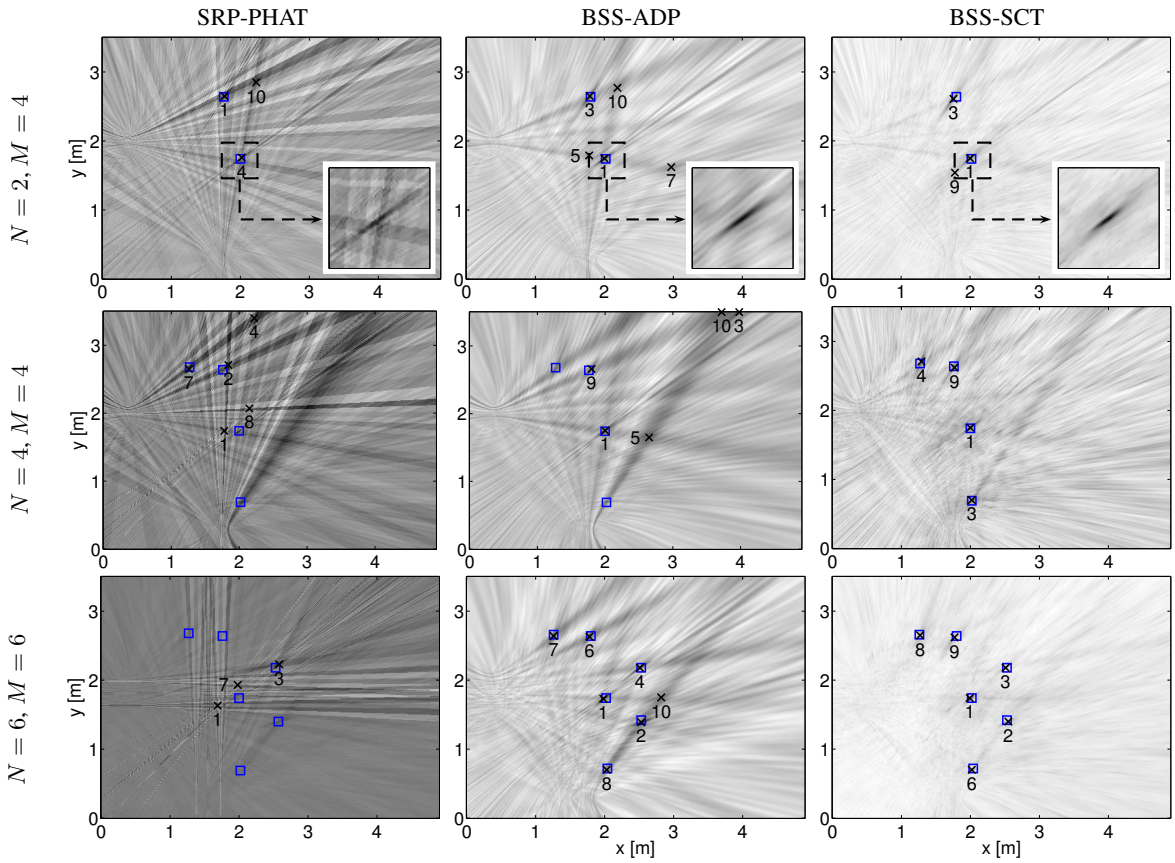


Fig. 3: Comparison of 2D position estimates with 3 s speech (□: true position, ×: estimated position)

	grid size	$N = 2$	$N = 3$	$N = 4$
BSS-ADP	5 cm	4.1 (73%)	4.0 (50%)	3.6 (7%)
	1 cm	2.9 (80%)	2.7 (55%)	2.7 (40%)
BSS-SCT	5 cm	4.4 (100%)	4.1 (100%)	4.7 (80%)
	1 cm	3.1 (100%)	3.1 (100%)	3.2 (87%)

Table 1: 2D localization errors in cm and detection rate (x%)

We see that all methods work well for $N = 2$ sources (top row) since the first 2 distinct peaks match the true source positions. However, when we increase the number of sources to 4 (middle row) or 6 (bottom row), SRP-PHAT and BSS-ADP yields peaks at erroneous locations, while BSS-SCT yield the first N distinct peaks at the correct locations. Comparing BSS-ADP and BSS-SCT, we observe that BSS-ADP shows much higher "sidelobes", which is harmful when the number of sources is unknown and results in a lower detection rate. Table 1 summarizes the mean localization errors and the detection rate for $M = 4$ for all possible combinations of $N = 2, 3, 4$ sources located at the six positions denoted in Fig. 2. Localization is performed by finding the first N distinct peaks of each cost function in the same way as in Fig. 3. A source is correctly detected if its position estimate is within 20 cm of the true source position. Localization errors are only calculated for correctly detected sources. As we can see, localization accuracy of BSS-ADP and BSS-SCT is comparable. However, BSS-ADP has a much lower detection rate and is less robust to a coarser search grid or smoothing operations which both are methods to overcome the problem of many local maxima.

7. CONCLUSIONS

In this paper, we have proposed a localization algorithm for multiple sources in reverberant environments. It relies on BSS-SCT and allows 1D/2D DOA or 2D/3D position estimation by comparing the signal propagation model against its estimate from BSS. Since our algorithm explicitly takes multiple sources and arbitrary room impulse responses into account, it shows a superior performance in comparison to SRP-PHAT and BSS-ADP.

8. REFERENCES

[1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein, Ed. Springer Verlag, 2001.

[2] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Multiple source localization using independent component analysis," *Proc. AP-S International Symposium and USNC/URSI National Radio Science Meeting (AP-S/URSI)*, 2005.

[4] A. Lombard, H. Buchner, and W. Kellermann, "A real-time demonstrator for the 2D localization of two sound sources using blind adaptive MIMO system identification," *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.

[5] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," *Proc. ICASSP*, 2009.

[6] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," *Proc. ICASSP*, 2007.

[7] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," *Proc. International Workshop for Acoustic Echo and Noise Control (IWAENC)*, 2008.