INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multiple Sound Source Localization with SVD-PHAT

François Grondin, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
`{fgrondin,glass}@mit.edu`

Abstract

This paper introduces a modification of phase transform on singular value decomposition (SVD-PHAT) to localize multiple sound sources. This work aims to improve localization accuracy and keeps the algorithm complexity low for real-time applications. This method relies on multiple scans of the search space, with projection of each low-dimensional observation onto orthogonal subspaces. We show that this method localizes multiple sound sources more accurately than discrete SRP-PHAT, with a reduction in the Root Mean Square Error up to 0.0395 radians.

Index Terms: multiple sound source socalization, srp-phat, svd-phat, direction of arrival

1. Introduction

The cocktail party effect consists of the ability to focus on a specific conversation in a noisy environment. While humans can usually perform this task efficiently, distant speech processing remains challenging for automatic speech recognition (ASR) systems [1]. To improve ASR performances, it is common to use a beamformer with multiple microphones as a preprocessing step to enhance the corrupted speech signal [2, 3, 4]. Some beamforming methods, such as the delay and sum and the minimum variance distortionless response (MVDR) [5], require the target source direction of arrival (DOA). On the other hand, methods like geometric sound separation require both the target and interference sources direction of arrival [6, 7]. It is therefore desirable to estimate the direction of arrival of multiple sound sources.

High resolution methods such as Multiple Signal Classification (MUSIC) [8] and the Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) [9] were initially designed for narrowband signals, and subsequently adapted to broadband signals such as speech [10, 11, 12, 13, 14, 15]. However, MUSIC-based methods involve online computations of eigenvectors, which makes real-time implementation challenging on low-cost embedded hardware. On the other hand, ESPRIT-based techniques require significantly less computations, but need twice as many sensors as MUSIC to perform with similar performance, which is problematic for microphone arrays with few sensors.

Alternatively, the Steered-Response Power Phase Transform (SRP-PHAT) robustly estimates the direction of arrival of speech sources and can be computed with low-cost embedded hardware [16]. SRP-PHAT relies on the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) between each pair of microphones. The Fast Fourier Transform is often used to speed up the computation of GCC-PHAT, but this also reduces localization accuracy. This discrete SRP-PHAT approach can localize many sound sources by scanning the search space multiple times, and nulling the GCC-PHAT region related to each found DOA [17, 18, 19, 20]. Hierarchical search also reduces the number of lookups in memory [21, 22]. The discrete SRP-PHAT approach however relies on rounded TDOA values, which may reduce the localization accuracy.

Alternatively, Cai et al. propose using multiple subbands to individually localize one sound source per band [23]. Similarly, it is possible to localize multiple speech sources based on their distinct pitch values [24]. Pavlidi et al. introduce a technique to identify single-source zones in the time-frequency range and generate a histogram to count and localize multiple sound sources [25]. However, these methods rely on narrow bands to localize sound sources, which makes localization more sensitive to reverberation. On the other hand, localization can also exploit interesting properties of microphone arrays with symmetrical geometries. For instance, wavefield decomposition enables localizing multiple sound source with spherical arrays [26, 27, 28, 29, 30]. Similarly, low-complexity multiple sources localization is possible in 2-D with circular arrays [31, 32]. These methods offer interesting performance, but rely on a specific microphone array geometry, which restricts their scalability.

We recently proposed a new method called SVD-PHAT that relies on singular value decomposition to map the observations to a small subspace, and then uses a nearest neighbor search algorithm like a k-d tree to find the DOA [33]. This single source localization method is appealing as it preserves exact SRP-PHAT accuracy while greatly reducing the computational complexity, and can adapt to microphone array with arbitrary shapes. In this paper, we extend SVD-PHAT to localize multiple sound sources.

2. SRP-PHAT

We first introduce SRP-PHAT with rounded TDOA that allows efficient localization of multiple sound sources with arbitrary array shapes. Let $X^l_m[k] \in \mathbb{C}$ be the Short Time Fourier Transform (STFT) coefficients, where $N \in \mathbb{N}$ and $\Delta N \in \mathbb{N}$ stand for the frame and hop sizes in samples, respectively, and $k \in \{0, 1, \dots, N/2\}$, $m \in \mathcal{M} = \{1, 2, \dots, M\}$ and $l \in \mathbb{N}$ stand for the frequency bin, microphone and frame indexes, respectively. The cross-correlation $X^l_{i,j}[k]$ for each microphone pair $(i, j) \in \mathcal{P} = \{(x, y) \in \mathcal{M}^2 : x < y\}$ is obtained with the following recursive estimation with $\alpha \in [0, 1]$:

$$X^l_{i,j}[k] = (1-\alpha)X^{l-1}_{i,j}[k] + \alpha X^l_i[k](X^l_j[k])^* \qquad (1)$$

where $\{\dots\}^*$ stands for the complex conjugate. For clarity, the frame index l is omitted in this paper without loss of generality. The phase transformed spectrum $\hat{X}_{i,j}[k] \in \mathbb{C}$ for each microphone pair is then obtained in (2), where $|\dots|$ stands for the absolute value.

http://dx.doi.org/10.21437/Interspeech.2019-2653

$$\hat{X}_{i,j}[k] = X_{i,j}[k]/|X_{i,j}[k]| \tag{2}$$

The generalized cross correlation with phase transform (GCC-PHAT) for each pair of microphone and TDOA $\tau \in \mathbb{R}$ is given in (3), where $W[k,\tau] = \exp\left(2\pi\sqrt{-1}k\tau/N\right)$.

$$x_{i,j}[\tau] = \sum_{k=0}^{N/2} \hat{X}_{i,j}[k]W[k,\tau] \tag{3}$$

The TDOA $\tau_{i,j,q} \in \mathbb{R}$ (in samples) corresponds to the difference between the direction of arrival (DOA) from a source $\mathbf{s}_q \in \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 = 1\}$ to microphone i at position $\mathbf{r}_i \in \mathbb{R}^3$, and the DOA between the same source and another microphone j at position $\mathbf{r}_j \in \mathbb{R}^3$, scaled with the speed of sound in air $c \in \mathbb{R}^+$ (in m/sec) and the sample rate ($f_S \in \mathbb{N}$):

$$\tau_{i,j,q} = \frac{f_S}{c}(\mathbf{r}_j - \mathbf{r}_i) \cdot \mathbf{s}_q \tag{4}$$

where $\{\cdot\}$ stands for the dot product.

It is common to discretize $\tau_{i,j,q}$ by rounding to the closest integer (denoted as $\lfloor \tau_{i,j,q} \rceil \in \mathbb{Z}$), and compute the GCC-PHAT in (3) using an Inverse Fourier Transform (IFFT) for all $\tau = n \in \mathcal{N} = \{0, 1, \dots, N-1\}$. The expression $Y_q \in \mathbb{R}$ is then obtained as follows for every possible DOA $\mathbf{d}_q$, where $q \in \mathcal{Q} = \{1, 2, \dots, Q\}$::

$$Y_q = \sum_{(i,j)\in\mathcal{P}} x_{i,j}[\lfloor \tau_{i,j,q} \rceil \pmod N] \tag{5}$$

The index of the most likely DOA then corresponds to:

$$q^* = \arg\max_{q\in\mathcal{Q}} \{Y_q\} \tag{6}$$

Once the DOA at index q^* is found, we can remove its contribution from the current observations and perform a new scan to detect other active sound sources. A naive approach consists in nulling the expression Y_{q^*} and some of its closest neighbors, and then scan for a new maximum value. However, this approach ignores the possible side lobes generated by the found source, and these may lead to false positives in the next scan iteration. To address this issue, a popular solution consists in nulling some regions in the GCC-PHAT results instead, recompute $Y_q \; \forall \; q \in \mathcal{Q}$ with (5), and then find a new maximum as in (6). For each DOA index q, we define a subset of DOA indexes $\mathcal{Q}_q = \{x \in \{1, 2, \dots, Q\} : \arccos\left(\mathbf{s}_x \cdot \mathbf{s}_q\right) \leq \Delta\theta\}$ that gathers DOAs close in space to the DOA $\mathbf{s}_q$, where $\Delta\theta$ is a user-defined parameter that stands for the maximum angle difference. We then define the set $\mathcal{T}_{i,j,q} = \{\tau^{min}_{i,j,q}, \tau^{min}_{i,j,q}+1, \dots, \tau^{max}_{i,j,q}-1, \tau^{max}_{i,j,q}\}$ that contains all TDOAs related to the DOA $\mathbf{s}_q$ and its closest neighbors, for the microphone pair (i,j), where:

$$\tau^{min}_{i,j,q} = \left\lfloor \min_{p\in\mathcal{Q}_q} \{\tau_{i,j,p}\} \right\rfloor \text{ and } \tau^{max}_{i,j,q} = \left\lceil \max_{p\in\mathcal{Q}_q} \{\tau_{i,j,p}\} \right\rceil \tag{7}$$

and the GCC-PHAT values in range of $\mathcal{T}_{i,j,q^*}$ are then set to zero for all pairs.

Algorithm 1 summarizes how SRP-PHAT can be adapted to localize R multiple sources. At each scan $r \in \mathcal{R} = \{1, 2, \dots, R\}$, the GCC-PHAT values are updated, and the following scans are thus performed without the contribution of the source recently found. The expressions $\mathbf{d}_r$ and e_r stand for the DOA and energy level found at scan r.

Algorithm 1 SRP-PHAT for multiple sources

```
Offline:
Generate $\tau_{i,j,q}, \mathcal{T}_{i,j,q} \; \forall \; (i,j) \in \mathcal{P}, \forall \; q \in \mathcal{Q}$.
Online:
Compute $x_{i,j}[n] \; \forall \; (i,j) \in \mathcal{P}, \forall \; n \in \mathcal{N}$.
for $r \in \mathcal{R}$ do
    Compute $Y_q \; \forall \; q \in \mathcal{Q}$.
    Find $q^*$ using linear search.
    for $(i,j) \in \mathcal{P}$ do
        $x_{i,j}[\tau] \leftarrow 0, \forall \; \tau \in \mathcal{T}_{i,j,q^*}$
    end for
    $\mathbf{d}_r \leftarrow \mathbf{s}_{q^*}$, $e_r \leftarrow Y_{q^*}$
end for
```

Although appealing as it relies on an efficient implementation of GCC-PHAT with IFFTs, this approach relies on discrete cross-correlation results, which reduces the accuracy. We therefore propose to adapt SVD-PHAT to estimate the direction of arrival (DOA) of multiple sound sources with more accuracy.

3. SVD-PHAT

To define the SVD-PHAT method, it is convenient to start from SRP-PHAT in matrix form. Let us define the vector $\mathbf{X}_{i,j} \in \mathbb{C}^{(N/2+1)\times 1}$ for the microphone pair $(i,j) \in \mathcal{P}$ that holds the phase normalized cross-correlation coefficients for all bins $k \in \{0, 1, \dots, N/2\}$:

$$\mathbf{X}_{i,j} = \begin{bmatrix} \hat{X}_{i,j}[0] & \hat{X}_{i,j}[1] & \cdots & \hat{X}_{i,j}[N/2] \end{bmatrix}^T \tag{8}$$

where $\{\dots\}^T$ stands for the transpose operator.

A single vector $\mathbf{X} \in \mathbb{C}^{P(N/2+1)\times 1}$ then holds all these vectors:

$$\mathbf{X} = \begin{bmatrix} (\mathbf{X}_{1,2})^T & (\mathbf{X}_{1,3})^T & \cdots & (\mathbf{X}_{M-1,M})^T \end{bmatrix}^T \tag{9}$$

For each pair of microphones (i,j), all coefficients $W[k,\tau] \in \mathbb{C}$ are concatenated in a matrix $\mathbf{W}_{i,j} \in \mathbb{C}^{Q\times(N/2+1)}$:

$$\mathbf{W}_{i,j} = \begin{bmatrix} W[0,\tau_{1,i,j}] & \cdots & W[N/2,\tau_{1,i,j}] \\ W[0,\tau_{2,i,j}] & \cdots & W[N/2,\tau_{2,i,j}] \\ \vdots & \ddots & \vdots \\ W[0,\tau_{Q,i,j}] & \cdots & W[N/2,\tau_{Q,i,j}] \end{bmatrix} \tag{10}$$

The supermatrix $\mathbf{W} \in \mathbb{C}^{Q\times P(N/2+1)}$ then holds all the matrices $\mathbf{W}_{i,j} \; \forall \; (i,j) \in \mathcal{P}$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{1,2} & \mathbf{W}_{1,3} & \cdots & \mathbf{W}_{M-1,M} \end{bmatrix} \tag{11}$$

Finally, the vector $\mathbf{Y} \in \mathbb{R}^{Q\times 1}$ holds the results $Y_q \; \forall \; q \in \{1, 2, \dots, Q\}$, where $\Re\{\dots\}$ returns the real part:

$$\mathbf{Y} = \begin{bmatrix} Y_1 & \dots & Y_Q \end{bmatrix}^T = \Re\{\mathbf{W}\mathbf{X}\} \tag{12}$$

The supermatrix $\mathbf{W}$ can be estimated with SVD of rank $K \in \{1, 2, \dots, K_{max}\}$, with $K_{max} = \max\{Q, P(N/2+1)\}$ and where $\mathbf{U} \in \mathbb{C}^{Q\times K}$, $\mathbf{S} \in \mathbb{C}^{K\times K}$ and $\mathbf{V} \in \mathbb{C}^{P(N/2+1)\times K}$:

$$\mathbf{W} \approx \mathbf{U}\mathbf{S}\mathbf{V}^H \tag{13}$$

The rank K corresponds to the minimum value for which the following condition holds, where $\delta \in (0,1)$ is a user-defined parameter that stands for the reconstruction tolerable error:

$$\mathrm{Tr}\left\{\mathbf{S}\mathbf{S}^T\right\} \geq (1-\delta)\,\mathrm{Tr}\left\{\mathbf{W}\mathbf{W}^H\right\} \tag{14}$$

where $\mathrm{Tr}\{\dots\}$ stands for the trace of the matrix.

The vector $\mathbf{Z} \in \mathbb{C}^{K\times 1}$ then results from the projection of the observations $\mathbf{X}$ in the K-dimensions subpace:

$$\mathbf{Z} = \mathbf{V}^H\mathbf{X} \tag{15}$$

Similarly, we define the dictionary $\mathbf{D} \in \mathbb{C}^{Q\times K}$, made of the vectors $\mathbf{D}_q \in \mathbb{C}^{1\times K}\ \forall\, q \in \{1,2,\dots,Q\}$:

$$\mathbf{D} = \mathbf{U}\mathbf{S} = \left[\begin{array}{cccc} (\mathbf{D}_1)^T & (\mathbf{D}_2)^T & \dots & (\mathbf{D}_Q)^T \end{array}\right]^T \tag{16}$$

As explained in [33], the DOA index then corresponds to q^*, obtained as follows:

$$q^* = \underset{q\in\mathcal{Q}}{\arg\max}\left\{\Re\{\mathbf{D}_q \cdot \mathbf{Z}^H\}\right\} \tag{17}$$

which can be converted into the following nearest neighbor problem with an algorithm such as k-d tree:

$$q^* = \underset{q\in\mathcal{Q}}{\arg\min}\left\{\|\hat{\mathbf{D}}_q - \hat{\mathbf{Z}}^H\|_2^2\right\} \tag{18}$$

where $\hat{\mathbf{D}}_q = \mathbf{D}_q/\|\mathbf{D}_q\|_2$ and $\hat{\mathbf{Z}} = \mathbf{Z}/\|\mathbf{Z}\|_2$.

Intuitively, we would like to remove the component in $\mathbf{Z}$ that spans the space spanned by $(\mathbf{D}_{q^*})^*$, and then perform a new scan to find another source. We thus define the vector $\mathbf{v}_r \in \mathbb{C}^{1\times K}$ as follows:

$$\mathbf{v}_r = (\mathbf{D}_{q^*})^* \tag{19}$$

The Gram-Schmidt process then makes the current vector $\mathbf{v}_r$ at scan r orthogonal to all the vectors previously found ($\hat{\mathbf{u}}_n\ \forall\, n \in \{1,2,\dots,r-1\}$), and generates $\mathbf{u}_r$:

$$\mathbf{u}_r = \mathbf{v}_r - \sum_{n=1}^{r-1}(\hat{\mathbf{u}}_n \cdot \mathbf{v}_r)\hat{\mathbf{u}}_n \tag{20}$$

which is then normalize to have a unit norm:

$$\hat{\mathbf{u}}_r = \mathbf{u}_r/\|\mathbf{u}_r\|_2 \tag{21}$$

Finally, the current observation $\mathbf{Z}$ is projected in the subspace orthogonal to $\hat{\mathbf{u}}_r$ to remove the current contribution of the source previously found:

$$\mathbf{Z}' = \mathbf{Z} - (\hat{\mathbf{u}}_r \cdot \mathbf{Z})\hat{\mathbf{u}}_r \tag{22}$$

Algorithm 2 summarizes these steps for SVD-PHAT. This approach is appealing as it involves R k-d tree search instead of computing R times $Y_q\ \forall\, q \in \mathcal{Q}$ as in (5), which reduces the algorithm complexity.

4. Results

We investigate three different microphone array geometries: a 1-D linear array, a 2-D planar array and a 3-D array. The microphones xyz-positions with respect to the center of the array are given in cm in Table 1.

Simulations are conducted to measure the accuracy of the proposed method and compare it to the SRP-PHAT approach discretized with IFFTs. The microphone array is positioned and rotated randomly in a 10m × 10m × 3m rectangular room, with a minimum distance of 0.5m from the walls, ceiling and floor. The target sources are also positioned randomly in the room, and the random setup ensures a minimum angle difference of 30° between each source, a distance of at least 0.5m between each source and the center of the microphone array, and a distance of at least 0.5m between each source and the walls, ceiling and floor. For each configuration, the room reverberation is modeled with Room Impulse Responses (RIRs) generated with the image method [34], where the reverberation time (RT60) is sampled randomly in the uniform interval between 200 and 500 msecs, which corresponds to the levels previously used in [33]. Sound segments selected randomly from the TIMIT dataset [35] are normalized to have the same energy levels, and are convolved with the generated RIRs. For each type of array (1-D, 2-D and 3-D) and number of active sources (1, 2 and 3), we perform 1000 simulations.

Table 2 introduces the parameters used with SRP-PHAT and SVD-PHAT. The sample rate f_S captures all the frequency content of speech (including wideband fricatives that contain relevant localization information), and the speed of sound c matches typical indoor conditions at room temperature. The frame size N analyze speech segments of 32 msecs, and the hop size provides a 50% overlap. The DOAs are scanned on a unit sphere generated recursively from a tetrahedron, for a total of 2562 points, as in [21]. Moreover, the cross-correlation adaptation rate α estimates the sound statistic over the past 400 msecs. In the case of SRP-PHAT, the maximum angle difference $\Delta\theta$ corresponds to 0.1745 radians to null the current source within a region of 10°. In the specific case of SVD-PHAT, the user-defined parameter is set to $\delta = 10^{-5}$ as in [33].

For a 1-D array with all microphones on the x-axis, the spatial resolution is limited to an arc that goes from 0° to 180° in

Algorithm 2 SVD-PHAT for multiple sources

```
Offline:
1: Generate $\mathbf{D}$, $\mathbf{V}$, and $\bar{\mathbf{V}}_q\ \forall\, q \in \{1,2,\dots,Q\}$.
Online:
1: Compute $\mathbf{Z}$ from $\mathbf{V}$ and observations $\mathbf{X}$.
2: for $r \in \{1,2,\dots,R\}$ do
3:     Find $q^*$ using a k-d tree to minimize $\|\hat{\mathbf{D}}_q - \hat{\mathbf{Z}}^H\|_2^2$.
4:     Compute $Y_{q^*}$, $\mathbf{v}_r$, $\hat{\mathbf{u}}_r$ and $\mathbf{Z}'$.
5:     $\mathbf{Z} \leftarrow \mathbf{Z}'$, $\mathbf{d}_r \leftarrow \mathbf{s}_{q^*}$, $e_r \leftarrow Y_{q^*}$
6: end for
```

Table 1: *Positions (x,y,z) of the microphones in cm*

Mic	1-D	2-D	3-D
1	$(-5.0,0,0)$	$(0,0,0)$	$(0,0,0)$
2	$(-3.3,0,0)$	$(5,0,0)$	$(-5,0,0)$
3	$(-1.7,0,0)$	$(2.5,4.3,0)$	$(5,0,0)$
4	$(0,0,0)$	$(-2.5,4.3,0)$	$(0,-5,0)$
5	$(1.7,0,0)$	$(-5.0,0,0)$	$(0,5,0)$
6	$(3.3,0,0)$	$(-2.5,-4.3,0)$	$(0,0,-5)$
7	$(5.0,0,0)$	$(2.5,-4.3,0)$	$(0,0,5)$

Table 2: *Parameters for SRP-PHAT and SVD-PHAT*

f_S	c	N	ΔN	Q	α
16000	340.0	512	128	2562	0.1

the xy-plane. The 3-D DOA is therefore projected to this subspace as follows:

$$f_1(\mathbf{x}) = [\cos(g(\mathbf{x})), \sin(g(\mathbf{x})), 0] \tag{23}$$

where:

$$g(\mathbf{x}) = \text{atan2}\left\{(\mathbf{x})_x, \sqrt{(\mathbf{x})_y^2 + (\mathbf{x})_z^2}\right\} \tag{24}$$

Similarly, a 2-D array that spans the xy-plane allows DOA estimation on a half hemisphere oriented in the z-axis, and therefore all DOAs are projected to the positive z-axis:

$$f_2(\mathbf{x}) = [(\mathbf{x})_x, (\mathbf{x})_y, |(\mathbf{x})_z|] \tag{25}$$

Finally, for a 3-D array, the DOA can span the full space:

$$f_3(\mathbf{x}) = \mathbf{x} \tag{26}$$

For each speech source $t \, \forall \, \{1, 2, \dots, T\}$, the minimum angle difference (in radians) between the theoretical DOA and all estimated DOAs at frame $l \in \{1, 2, \dots, L\}$ is given as follows:

$$\phi_t^l = \min_{r \in \mathcal{R}} \left\{\arccos\left(f_\beta(\mathbf{d}_r^l) \cdot f_\beta(\mathbf{c}_t)\right)\right\} \tag{27}$$

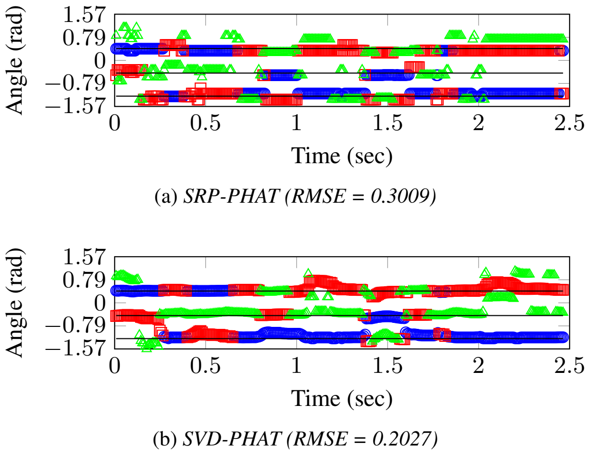
where $\beta \in \{1, 2, 3\}$ matches the array geometry. The goal is therefore to have at least one DOA estimation that matches each speech source true DOA.

In the proposed experiments, the number of sources varies with $T = \{1, 2, 3\}$, and the number of scans R matches this number. The root mean square error (RMSE) in rad for a simulation therefore corresponds to:

$$RMSE = \sqrt{\frac{1}{LT}\sum_{l=1}^{L}\sum_{t=1}^{T}(\phi_t^l)^2} \tag{28}$$

Figure 1 shows the estimated DOAs obtained with SRP-PHAT and SVD-PHAT for a 1-D array with three speech sources located at -1.2192 rad, -0.4335 rad and 0.4015 rad, and a reverberation time (RT60) of 238 msecs. In this example, the SRP-PHAT method fails to detect the source at -0.4335 rad at different times, whereas SVD-PHAT detects this source most of the time. The RMSEs of SRP-PHAT and SVD-PHAT correspond to 0.3009 rad and 0.2027 rad, respectively, which indicates that SVD-PHAT outperforms SRP-PHAT in this specific example.

The RMSE gap between both SRP-PHAT and SVD-PHAT is however usually smaller than the one shown in Figure 1. To better compare both methods, Table 3 shows the mean of all RMSEs for the 1000 simulations with each configuration. In all cases, the proposed multiple source SVD-PHAT reduces the RMSE compared to the discrete SRP-PHAT, but with a smaller gap that oscillates between 0.0244 rad and 0.0395 rad. It is interesting to note that for multiple sound sources ($T > 1$), the RMSE increases rapidly. This is expected as multiple active sources partially overlap each other in the time-frequency domain, which makes localization more challenging. The best improvement for multiple sources occurs for the 3-D array when two sources are active, with a reduction in the RMSE of 0.0395 rad.



(a) *SRP-PHAT (RMSE = 0.3009)*

(b) *SVD-PHAT (RMSE = 0.2027)*

Figure 1: *Azimuth angle (in rad, obtained with $g(\mathbf{x})$ in (24)) of potential sources found with SRP-PHAT and SVD-PHAT, for $r = 1$ (blue circles), $r = 2$ (red squares) and $r = 3$ (green triangles). The theoretical DOAs are -1.2192, -0.4335 and 0.4015, and are plotted with solid black lines. For this simulation, the reverberation time (RT60) corresponds to 238 msecs.*

Table 3: *Root Mean Square Error (RMSE) – less is better*

Geometry	Nb. Sources	SRP-PHAT	SVD-PHAT
1-D	1	0.0884	**0.0509**
	2	0.2656	**0.2274**
	3	0.2763	**0.2519**
2-D	1	0.1356	**0.0820**
	2	0.4516	**0.4200**
	3	0.4201	**0.3828**
3-D	1	0.0708	**0.0296**
	2	0.4550	**0.4155**
	3	0.5445	**0.5189**

5. Conclusion

This paper extends SVD-PHAT for multiple sound source localization. This technique outperforms the discrete SRP-PHAT approach in terms of accuracy, while preserving the low complexity of the original SVD-PHAT. On average, the reduction in the RMSE varies between 0.0244 and 0.0395 radians, and the best improvement is observed for an array that spans 3-D space with two simultaneous speech sources.

In future work, we will investigate alternatives to k-d tree search to address the curse of dimensionality during the nearest neighbor search [36]. The method could also be extended to deal with speed of sound mismatch and the near-field effect. Microphone directivity could also be combined with SVD-PHAT to make the propagation model more realistic [37]. The sound source tracking method proposed in [21] could also be combined to SVD-PHAT to estimate the number of sound sources and track their positions over time. Finally, it would be interesting to implement SVD-PHAT in C code for easy deployment on real-time embedded systems.

6. Acknowledgements

This work was supported in part by the Toyoto Research Institute and Signify.

7. References

[1] H. Tang, W.-N. Hsu, F. Grondin, and J. Glass, "A study of enhancement, augmentation, and autoencoder methods for domain adaptation in distant speech recognition," in *Proc. INTERSPEECH*, 2018, pp. 2928–2932.

[2] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE ASRU*, 2015, pp. 444–451.

[3] B.-K. Lee and J. Jeong, "Deep Neural Network-based Speech Separation Combining with MVDR Beamformer for Automatic Speech Recognition System," in *Proc. IEEE ICCE*, 2019, pp. 1–4.

[4] X. Sun, Z. Wang, R. Xia, J. Li, and Y. Yan, "Effect of steering vector estimation on MVDR beamformer for noisy speech recognition," in *Proc. IEEE DSP*, 2018, pp. 1–5.

[5] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, p. 158, 2010.

[6] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[7] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. IEEE/RSJ IROS*, vol. 3, 2004, pp. 2123–2128.

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

[9] R. Roy, A. Paulraj, and T. Kailath, "Estimation of signal parameters via rotational invariance techniques - ESPRIT," in *Proc. IEEE MILCOM*, 1986.

[10] C. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. IEEE/RSJ IROS*, 2009, pp. 2027–2032.

[11] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent sound source localization and its application to multimodal human tracking," in *Proc. IEEE/RSJ IROS*, 2011, pp. 143–148.

[12] K. Nakamura, K. Nakadai, and H. Okuno, "A real-time super resolution robot audition system that improves the robustness of simultaneous speech recognition," *Advanced Robotics*, vol. 27, no. 12, pp. 933–945, 2013.

[13] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of mulitple wideband acoustic sources using eigen-beams," in *Proc. ICASSP*, 2005, pp. 89–92.

[14] S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proc. IEEE/RSJ IROS*, 2007, pp. 2009–2014.

[15] P. Danès and J. Bonnal, "Information-theoretic detection of broadband sources in a coherent beamspace MUSIC scheme," in *Proc. IEEE/RSJ IROS*, 2010, pp. 1976–1981.

[16] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001, pp. 157–180.

[17] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, "The ManyEars open framework," *Autonomous Robots*, vol. 34, no. 3, pp. 217–232, 2013.

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

[19] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound source for mobile robot using a frequency-domain steered beamformer approach," in *Proc. IEEE ICRA*, 2004, pp. 1033–1038.

[20] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization nad tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE ICASSP*, 2006, pp. 841–844.

[21] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robotics and Autonomous Systems*, vol. 113, pp. 63–80, 2019.

[22] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 20–26, 2016.

[23] W. Cai, X. Zhao, and Z. Wu, "Localization of multiple speech sources based on sub-band steered response power," in *Proc. IEEE ICECE*, 2010, pp. 1246–1249.

[24] M. Kepesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. IEEE HSCMA*, 2008, pp. 85–88.

[25] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3D localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. IEEE EUSIPCO*, 2015, pp. 1556–1560.

[26] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *Proc. IEEE ICASSP*, 2008, pp. 5276–5279.

[27] C. Evers, A. Moore, and P. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. IWAENC*, 2014, pp. 258–262.

[28] S. Hafezi, A. Moore, and P. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. IEEE EUSIPCO*, 2016, pp. 602–606.

[29] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *Proc. IEEE ICASSP*, 2011, pp. 117–120.

[30] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[31] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[32] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2724–2736, 2006.

[33] F. Grondin and J. Glass, "SVD-PHAT: A fast sound source localization method," in *Proc. IEEE ICASSP*, 2019.

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

[35] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[36] S. Berchtold, C. Böhm, and H.-P. Kriegal, "The pyramid-technique: towards breaking the curse of dimensionality," in *Proc. ACM SIGMOD Record*, vol. 27, no. 2, 1998, pp. 142–153.

[37] M. Thomas, J. Ahrens, and I. Tashev, "Optimal 3D beamforming using measured microphone directivity patterns," in *Proc. IWAENC*, 2012, pp. 1–4.