2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Fast and Robust 3-D Sound Source Localization with DSVD-PHAT

François Grondin and James Glass

Abstract— **This paper introduces a variant of the Singular Value Decomposition with Phase Transform (SVD-PHAT), named Difference SVD-PHAT (DSVD-PHAT), to achieve robust Sound Source Localization (SSL) in noisy conditions. Experiments are performed on a Baxter robot with a four-microphone planar array mounted on its head. Results show that this method offers similar robustness to noise as the state-of-the-art Multiple Signal Classification based on Generalized Singular Value Decomposition (GSVD-MUSIC) method, and considerably reduces the computational load by a factor of 250. This performance gain thus makes DSVD-PHAT appealing for real-time application on robots with limited on-board computing power.**

I. INTRODUCTION

Robot audition aims to provide robots with hearing capabilities to interact efficiently with people in everyday environments [1]. Sound source localization (SSL) is a typical task that consists of localizing the direction of arrival (DOA) of a target source using a microphone array. This task is challenging as the robot usually generates a significant amount of noise (fans, actuators, etc.) [2] and the target sound source is corrupted by reverberation. SSL often relies on Multiple Signal Classification (MUSIC) and Steered-Response Power Phase Transform (SRP-PHAT) methods.

MUSIC is a localization method based on Standard Eigenvalue Decomposition (SEVD-MUSIC) that was initially used for narrowband signals [3], and then adapted to broadband signals like speech [4]. However, SEVD-MUSIC assumes the speech signal is more powerful than noise at each frequency bin in the spectrogram, which is usually not the case. To cope with this limitation, Nakamura et al. introduced the MUSIC based on Generalized Eigenvalue Decomposition (GEVD-MUSIC) method [5], [6], [7]. This method solves the limitation of SEVD-MUSIC, but also introduces some localization errors because the transform provides a noise subspace with correlated bases. To deal with this issue, a variant of GEVD-MUSIC, named MUSIC based on Generalized Singular Value Decomposition (GSVD-MUSIC), enforces orthogonality between the noise subspace bases and thus improves the DOA estimation accuracy [8]. However, all MUSIC-based methods rely on online eigenvalue or singular value decompositions that are computationally expensive, and make on-board real-time processing challenging [9].

SRP-PHAT is built on the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) between each pair of microphones [10]. GCC-PHAT is often computed with the Inverse Fast Fourier Transform (IFFT) to speed up computation, at the cost of discretizing Time Difference of Arrival (TDOA) values, which reduces localization accuracy. SRP-PHAT usually scans a discretized 3-D space and returns the most likely DOA [11], [12], [13], [14], [15], [16]. This scanning process often involves a significant amount of lookups in memory, which creates a bottleneck and increases execution time. To reduce the number of lookups, a hierarchical search is proposed to speed up the space scan, but this method still relies on discrete TDOA [17]. We therefore recently proposed the Singular Value Decomposition with Phase Transform (SVD-PHAT) method, which avoids TDOA discretization, and significantly reduces computing time [18]. However, as for SRP-PHAT, SVD-PHAT remains sensitive to additive noise. To cope with this limitation, time-frequency (TF) masks can be generated to improve robustness to stationary noise [19], [20]. Stationary noise is often estimated with techniques like Minima Controlled Recursive Averaging (MCRA) [21] and Histogram-based Recursive Level Estimation (HRLE) [22], or recorded offline prior to test if the robot's environment is static. Pertilä et al. also propose a method that generates TF masks using convolutional neural networks for non-stationary noise sources [23]. However, these TF masks ignore noise spatial coherence, which carries useful insights for robust localization, and is in fact exploited by GSVD-MUSIC.

In this paper, we propose a variant of the SVD-PHAT method, called Difference SVD-PHAT (DSVD-PHAT), that performs correlation matrix subtraction, which considers noise spatial coherence, while preserving the low complexity of the original SVD-PHAT. Section II reviews the state of the art GSVD-MUSIC method, and section III introduces the proposed DSVD-PHAT method. Section IV describes the experimental setup on a Baxter robot, and then section V compares results from GSVD-MUSIC and the proposed DSVD-PHAT approach.

II. GSVD-MUSIC

GSVD-MUSIC relies on the Time Difference of Arrival (TDOA) between each microphone and a reference in space. The TDOA (in sec) stands for the propagation delay for the signal emitted by the sound source DOA $\mathbf{s}_q \in \{\mathbb{R}^3 : \|\mathbf{s}_q\|_2 = 1\}$ (where $\|\ldots\|_2$ stands for the l_2-norm) to reach microphone $\mathbf{r}_m \in \mathbb{R}^3$ with respect to the origin. For discrete-time signals, the TDOA is usually expressed in terms of samples, as shown in (1), where $c \in \mathbb{R}^+$ stands for the speed of sound in air (in m/sec), and $f_S \in \mathbb{R}^+$ is the sample rate

This work was supported in part by the Toyota Research Institute and by the Fonds de recherche du Québec – Nature et technologies.

F. Grondin and J. Glass are with the Computer Science & Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA 02139, USA {fgrondin,glass}@mit.edu

(in samples/sec). The operator $\cdot$ stands for the dot product.

$$\tau_{q,m} = \left(\frac{f_S}{c}\right) \mathbf{r}_m \cdot \mathbf{s}_q \tag{1}$$

The expression $X_m^l[k] \in \mathbb{C}$ stands for the Short Time Fourier Transform coefficient of microphone $m \in \{1, \dots, M\}$, at frequency bin $k \in \{0, \dots, N/2\}$ and frame $l \in \mathbb{N}$, where $N \in \mathbb{N}$ and $\Delta N \in \mathbb{N}$ stand for the frame and hop sizes in samples, respectively. The STFT values are concatenated in the vector $\mathbf{x}^l[k] \in \mathbb{C}^{M\times 1}$, as shown in (2).

$$\mathbf{x}^l[k] = \begin{bmatrix} X_1^l[k] & X_2^l[k] & \cdots & X_M^l[k] \end{bmatrix}^T \tag{2}$$

GSVD-MUSIC uses a steering vector $\mathbf{A}_q[k] \in \mathbb{C}^{M\times 1}$ for each potential DOA $\mathbf{s}_q$:

$$\mathbf{A}_q[k] = \begin{bmatrix} A_{q,1}[k] & \cdots & A_{q,M}[k] \end{bmatrix}^T \tag{3}$$

where $A_{q,m}[k] = \exp\left(-2\pi\sqrt{-1}k\tau_{q,m}/N\right)$.

The $\mathbb{C}^{M\times M}$ correlation matrix of the vector $\mathbf{x}^l[k]$ at each frequency bin k can be estimated at each frame l using the following recursive approximation, where the parameter $\alpha \in (0, 1)$ is the adaptive rate:

$$\mathbf{R}_{xx}^l[k] = (1-\alpha)\mathbf{R}_{xx}^{l-1}[k] + \alpha \mathbf{x}^l[k](\mathbf{x}^l[k])^H \tag{4}$$

where $\{\dots\}^H$ stands for the Hermitian operator.

The GSVD-MUSIC method performs a generalized singular value decomposition with respect to the noise correlation matrix $\mathbf{R}_{nn}^l[k]$ (which can be estimated as in (4) during silence periods or precomputed offline if the test environment is known):

$$(\mathbf{R}_{nn}^l[k])^{-1}\mathbf{R}_{xx}^l[k] = \mathbf{E}^l[k]\mathbf{\Lambda}^l[k](\mathbf{F}^l[k])^H \tag{5}$$

where the diagonal matrix $\mathbf{\Lambda}^l[k] \in (\mathbb{R}^+)^{M\times M}$ holds the singular values in descending order ($\lambda_1^l[k] > \lambda_2^l[k] > \cdots > \lambda_M^l[k]$), and $\mathbf{E}^l[k] \in \mathbb{C}^{M\times M}$ and $\mathbf{F}^l[k] \in \mathbb{C}^{M\times M}$ are the left and right singular vectors $\mathbf{e}_1^l[k], \dots, \mathbf{e}_M^l[k] \in \mathbb{C}^{M\times 1}$ and $\mathbf{f}_1^l[k], \dots, \mathbf{f}_M^l[k] \in \mathbb{C}^{M\times 1}$, respectively:

$$\mathbf{\Lambda}^l[k] = \begin{bmatrix} \lambda_1^l[k] & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_M^l[k] \end{bmatrix} \tag{6}$$

$$\mathbf{E}^l[k] = \begin{bmatrix} \mathbf{e}_1^l[k], & \dots, & \mathbf{e}_M^l[k] \end{bmatrix} \tag{7}$$

$$\mathbf{F}^l[k] = \begin{bmatrix} \mathbf{f}_1^l[k], & \dots, & \mathbf{f}_M^l[k] \end{bmatrix} \tag{8}$$

This method projects the steering vector $\mathbf{A}_q[k]$ in the noise subspace, spanned by the singular vectors $\mathbf{e}_m^l[k]\ \forall\ m \in \{2, 3, \dots, M\}$ (when there is only one target source). The inverse of the projections for each frequency bin k is summed over the full spectrum (which may also be restricted to a more specific range of frequency bins [8]):

$$P_q^l = \sum_{k=0}^{N/2} \left(\sum_{m=2}^{M} \|(\mathbf{A}_q[k])^H \mathbf{e}_m^l[k]\|_2 \right)^{-1} \tag{9}$$

The sound source DOA then corresponds to $\mathbf{s}_{\bar{q}_l}$, where:

$$\bar{q}_l = \arg\max_q \{P_q^l\} \tag{10}$$

GSVD-MUSIC involves $(N/2+1)$ singular value decompositions of $M \times M$ matrices per frame, as shown in (5), which is challenging from a computing point of view for real-time applications. Moreover, it also involves computing (9) for Q potential sources, which also implies a significant amount of computations. The proposed DSVD-PHAT aims to reduce the amount of computations, while preserving a similar robustness to noise.

III. DSVD-PHAT

DSVD-PHAT relies on the TDOA between each pair of microphones i and j (as opposed to (1), where the TDOA is between a microphone and to the origin), which leads to the following expression, for a total of $P = M(M-1)/2$ pairs:

$$\tau_{q,i,j} = \frac{f_S}{c}(\mathbf{r}_j - \mathbf{r}_i) \cdot \mathbf{s}_q \tag{11}$$

Since noise and speech sources are independent, it is reasonable to assume that the clean speech correlation matrix $\mathbf{R}_{ss}^l$ can be estimated from the difference between the noisy speech and the noise correlation matrices at each frame l, as proposed in [24]:

$$\mathbf{R}_{ss}^l[k] \approx \mathbf{R}_{xx}^l[k] - \mathbf{R}_{nn}^l[k] \tag{12}$$

The normalized cross-spectra in DSVD-PHAT at each frequency bin k are thus obtained as follows, where $(\dots)_{i,j}$ refers to the element in the ith row and jth column:

$$X_{i,j}^l[k] = \frac{(\mathbf{R}_{ss}^l[k])_{i,j}}{\|(\mathbf{R}_{ss}^l[k])_{i,j}\|_2} \tag{13}$$

Note how DSVD-PHAT differs from the original SVD-PHAT, as the latter uses directly the noisy correlation matrix (e.g. $\mathbf{R}_{xx}^l[k]$ replaces $\mathbf{R}_{ss}^l[k]$ in (12)).

We then define the vector $\mathbf{X} \in \mathbb{C}^{P(N/2+1)\times 1}$ to concatenate all normalized cross-spectra introduced in (13):

$$\mathbf{X}^l = \begin{bmatrix} X_{1,2}^l[0] & X_{1,2}^l[1] & \cdots & X_{M-1,M}^l[N/2] \end{bmatrix}^T \tag{14}$$

The matrix $\mathbf{W} \in \mathbb{C}^{Q\times P(N/2+1)}$ holds all the SRP-PHAT coefficients $W_{q,i,j}[k] = \exp\left(2\pi\sqrt{-1}k\tau_{q,i,j}/N\right)$:

$$\mathbf{W} = \begin{bmatrix} W_{1,1,2}[0] & W_{1,1,2}[1] & \cdots & W_{1,M-1,M}[N/2] \\ \vdots & \vdots & \ddots & \vdots \\ W_{Q,1,2}[0] & W_{Q,1,2}[1] & \cdots & W_{Q,M-1,M}[N/2] \end{bmatrix} \tag{15}$$

The vector $\mathbf{Y}^l \in \mathbb{R}^{Q\times 1}$ stores the SRP-PHAT energy for all Q potential DOAs, where $\Re\{\dots\}$ extracts the real part of the expression:

$$\mathbf{Y}^l = \begin{bmatrix} Y_1^l & \dots & Y_Q^l \end{bmatrix}^T = \Re\{\mathbf{W}\mathbf{X}^l\} \tag{16}$$

The sound source DOA corresponds to $\mathbf{s}_{\bar{q}_l}$, where:

$$\bar{q}_l = \arg\max_q \{Y_q^l\} \tag{17}$$

Computing Y_q^l for all values of q is expensive, and therefore SVD-PHAT provides a more efficient way of finding $\bar{q}_l$. The Singular Value Decomposition is first performed

on the $\mathbf{W}$ matrix, where $\mathbf{U} \in \mathbb{C}^{Q\times K}$, $\mathbf{S} \in \mathbb{C}^{K\times K}$ and $\mathbf{V} \in \mathbb{C}^{P(N/2+1)\times K}$:

$$\mathbf{W} \approx \mathbf{U}\mathbf{S}\mathbf{V}^H \tag{18}$$

The parameter $K \in \{1, 2, \ldots, K_{max}\}$ (where $K_{max} = \max\{Q, P(N/2+1)\}$) satisfies the condition in (19), which ensures accurate reconstruction of $\mathbf{W}$, where $\delta \in (0, 1)$ is a user-defined small value that stands for the tolerable reconstruction error. The operator $\mathrm{Tr}\{\ldots\}$ represents the trace of the matrix.

$$\mathrm{Tr}\left\{\mathbf{S}\mathbf{S}^T\right\} \geq (1-\delta)\,\mathrm{Tr}\left\{\mathbf{W}\mathbf{W}^H\right\} \tag{19}$$

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

$$\mathbf{Z}^l = \mathbf{V}^H\mathbf{X}^l \tag{20}$$

Similarly, the matrix $\mathbf{D} \in \mathbb{C}^{Q\times K}$ holds a set of Q vectors $\mathbf{D}_q \in \mathbb{C}^{1\times K}$:

$$\mathbf{D} = \mathbf{U}\mathbf{S} = \begin{bmatrix} \mathbf{D}_1^T & \mathbf{D}_2^T & \ldots & \mathbf{D}_Q^T \end{bmatrix}^T \tag{21}$$

The optimization in (17) can then be converted to a nearest neighbor problem:

$$\bar{q}_l = \arg\min_q \{\|\hat{\mathbf{D}}_q - (\hat{\mathbf{Z}}^l)^H\|_2^2\} \tag{22}$$

where $\hat{\mathbf{D}}_q = \mathbf{D}_q/\|\mathbf{D}\|_2$ and $\hat{\mathbf{Z}}_q = \mathbf{Z}_q/\|\mathbf{Z}\|_2$. A k-d tree then solves efficiently this nearest neighbor search problem. The corresponding amplitude for the optimal DOA at index $\bar{q}_l$ corresponds to:

$$Y_{\bar{q}_l}^l = \mathbf{W}_{\bar{q}_l}\mathbf{X}^l \tag{23}$$

where $\mathbf{W}_{\bar{q}_l}$ stands for the $\bar{q}_l$-th row of $\mathbf{W}$.

Both GSVD-MUSIC and DSVD-PHAT rely on SVD decompositions, but DSVD-PHAT computes them offline. The online processing only involves the projection in (20) and the k-d tree search, which is appealing for real-time processing.

IV. EXPERIMENTAL SETUP

The GSVD-MUSIC and DSVD-PHAT methods are evaluated for a Baxter robot setup, equipped with a 4-microphone ReSpeaker[1] array mounted on its head, as shown in Fig. 1.

To compare both methods with a wide range of conditions, we perform simulations to evaluate numerous room configurations and signal-to-noise ratios (SNRs). Noise from Baxter's fans is therefore recorded and then mixed with male and female speech utterances from the TIMIT dataset [25], convolved with simulated Room Impulse Responses (RIRs) and amplified with various gains. The room impulse response (RIR) corresponds to the impulse response obtained with the image method [26] between the microphone array and the target sound sources, both positioned randomly in a 10m x 10m x 3m room. For each pair of SNR and room reverberation time RT60, we generate 100 RIRs and use the same number of speech sources picked randomly from the TIMIT dataset.

[1] http://seeedstudio.io

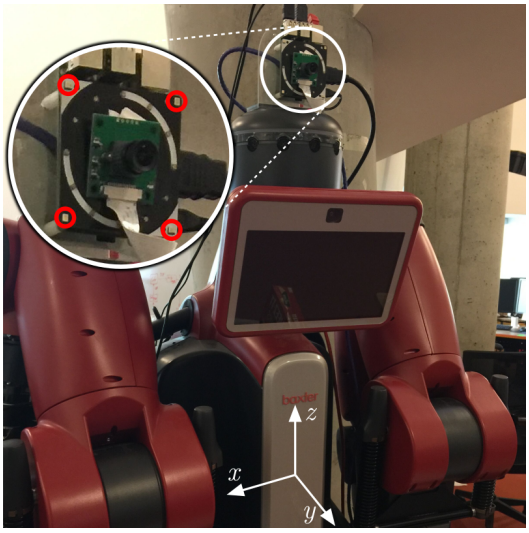


Fig. 1. Baxter robot equipped with a 4-microphone ReSpeaker array mounted on its head (microphones are circled in red)

TABLE I
GSVD-MUSIC AND GSVD-PHAT PARAMETERS

f_S	c	M	N	ΔN	Q	α	δ
16000	343.0	4	256	128	1282	0.05	10^{-5}

The parameters for the experiments are summarized in Table I. The sample rate f_S captures all the frequency content of speech, and the speed of sound c corresponds to typical indoor conditions. The frame size N analyzes segments of 16 msecs, and the hop size ΔN provides a 50% overlap. The potential DOAs are represented by equidistant points on a unit halfsphere generated recursively from a tetrahedron, for a total of 1282 points, as in [17]. The smoothing parameter α provides a context of roughly 800 msecs to estimate the correlation matrices, which captures multiple phonemes. The parameter δ is set to the value found in [18], which ensures a good accuracy. For this array configuration, the dimensionality of the subspace corresponds to $K = 23$ with $\delta = 10^{-5}$.

Table II lists the positions of the ReSpeaker array microphones (in cm) w.r.t. to the center of the array.

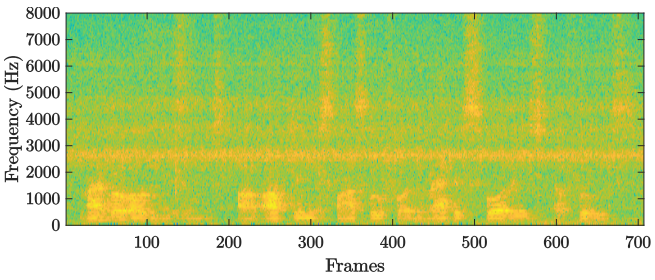
TABLE II
POSITIONS (X,Y,Z) OF THE MICROPHONES IN CM

m	x	y	z
1	+2.9	0.0	+2.9
2	+2.9	0.0	−2.9
3	−2.9	0.0	+2.9
4	−2.9	0.0	−2.9

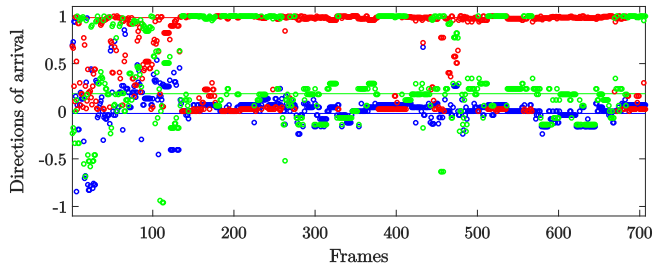
In all experiments, the noise correlation matrix comes from the offline recording of the robot's fans. This ensures we compare both methods independently of the performance of the online background noise estimation method.

V. RESULTS

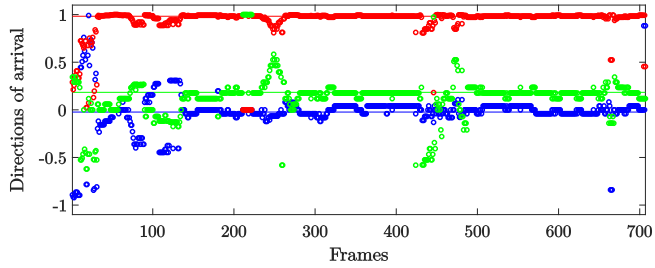
To get some intuition about the SSL with GSVD-MUSIC and DSVD-PHAT, we first analyze an example of a speech utterance with a SNR of 5 dB and a reverberation level of RT60 = 400 msecs, shown in Fig. 2. The spectrogram in Fig. 2a displays the speech signal, corrupted by some stationary noise between 2500Hz and 5000Hz. Fig. 2b shows the DOAs obtained from GSVD-MUSIC, with the true DOA represented by straigh lines. This example demonstrates that, in this specific case, GSVD-MUSIC estimates many DOAs that differ from the theoretical DOA. Similarly, Fig. 2c displays the DOAs obtained from DSVD-PHAT for the same noisy signal. Here the estimated DOAs are closer to the theoretical DOA.



(a) Spectrogram of the signal captured at microphone 1.



(b) Circles represent the $\mathbf{s}_{\bar{q}_l}$ found with GSVD-MUSIC, and lines stand for the theoretical DOA. The x-, y-, z-coordinates are represented by blue, red and green colors, respectively.



(c) Circles represent the $\mathbf{s}_{\bar{q}_l}$ found with DSVD-PHAT, and lines stand for the theoretical DOA. The x-, y-, z-coordinates are represented by blue, red and green colors, respectively.

Fig. 2. SSL with GSVD-MUSIC and DSVD-PHAT when RT60 = 400 msecs and SNR = 5 dB.

It is also convenient to define the expression $\theta_l \in [0, \pi/2]$ to denote the angle difference between the estimated DOA $\mathbf{s}_{q_l}$ at frame l (obtained using GSVD-MUSIC or DSVD-PHAT), and the theoretical DOA $\mathbf{s}_{true}$ extracted from the simulated room parameters:

$$\theta_l = \arccos\{\mathbf{s}_{q_l} \cdot \mathbf{s}_{true}\} \tag{24}$$

Let us define the margin $\Delta\theta \in [0, \pi/2]$, that corresponds to the DOA error tolerance for a localized source to be considered as a valid DOA. In this section, we arbitrary define the tolerance to $\Delta\theta = 0.2$ radians, which corresponds to 11.5°. Expression Θ_l takes a value of 1 when the localized sound source is within the range, or 0 otherwise:

$$\Theta_l = \begin{cases} 1 & \theta_l \leq \Delta\theta \\ 0 & \theta_l > \Delta\theta \end{cases} \tag{25}$$

Similarly, the expression e_l corresponds to the observation amplitude ($e_l = P^l_{q_l}$ for GSVD-MUSIC from (9), and $e_l = Y^l_{q_l}$ from (23) for DSVD-PHAT). This metric is relevant as it is often assumed that the confidence in the DOA $\mathbf{s}_{\bar{q}_l}$ depends on the associated amplitude of e_l [16], [17]. Therefore, a DOA is considered as a positive when the amplitude e_l equals or exceeds the fixed threshold T_{min}, and as a negative otherwise:

$$E_l = \begin{cases} 1 & e_l \geq T_{min} \\ 0 & e_l < T_{min} \end{cases} \tag{26}$$

Fig. 3 illustrates the angle difference of the DOAs estimated previously with both methods, and also displays the associated amplitudes. Note that for DSVD-PHAT in particular, the amplitude goes down when the value of θ gets outside the acceptable range, which suggests that a well-tuned T_{min} could discriminate between accurate and inaccurate estimated DOAs.

To measure the performance of both methods, we vary the value of T_{min} and compute the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). A TP occurs when the amplitude is greater or equal to the threshold, and the measured DOA falls within the acceptable range of the theoretical DOA:

$$TP = \sum_{l=0}^{L} \Theta_l e_l \tag{27}$$

Similarly, a TN happens when a DOA out of the acceptable range is rejected as its associated amplitude is below the fixed threshold:

$$TN = \sum_{l=0}^{L} (1 - \Theta_l)(1 - e_l) \tag{28}$$

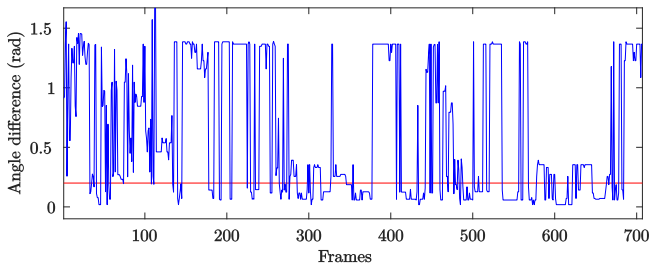
Finally, FP and FN occur when an erroneous DOA is picked and when a valid DOA is rejected, respectively:

$$FP = \sum_{l=0}^{L} (1 - \Theta_l) e_l \tag{29}$$

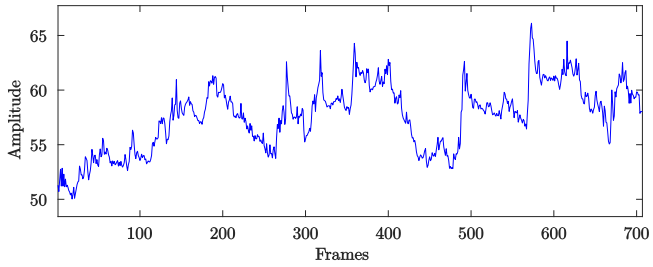
$$FN = \sum_{l=0}^{L} \Theta_l (1 - e_l) \tag{30}$$

The True Positive Rate (TPR) and False Positive Rate (FPR) then correspond to (31) and (32), respectively, and are used to build the ROC curve.

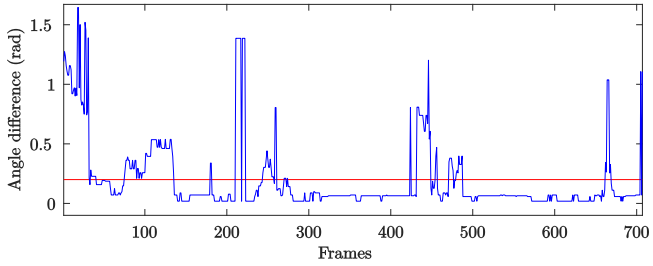
$$TPR = \frac{TP}{TP + FN} \tag{31}$$



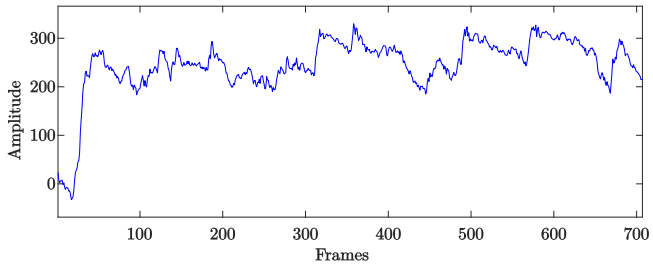
(a) Angle difference θ_l for GSVD-MUSIC (blue) and $\Delta\theta$ threshold (red).



(b) Amplitude $e_l = P^l_{\bar{q}_l}$ for GSVD-MUSIC.



(c) Angle difference θ_l for DSVD-PHAT (blue) and $\Delta\theta$ threshold (red).



(d) Amplitude $e_l = Y^l_{\bar{q}_l}$ for DSVD-PHAT.

Fig. 3. Comparisons between GSVD-MUSIC and DSVD-PHAT methods.

$$FPR = \frac{FP}{FP + TN} \tag{32}$$

Fig. 4 shows both ROC curves with GSVD-MUSIC and DSVD-PHAT for the previous example. In this case, the DSVD-PHAT surpasses the GSVD-MUSIC results as the Area Under the Curve (AUC) is clearly closer to 1.

Table III shows the AUC results for SNRs $\in \{-10, -5, \ldots, 20\}$ dB and RT60 $\in \{200, 400, 600, 800\}$ msecs. In general, GSVD-MUSIC generates higher AUC values for cases when the SNR is below 0dB. However, the DSVD-PHAT still provides AUC values close to GSVD-MUSIC, which demonstrates that the proposed method also allows accurate DOA estimation under reverberant and noisy conditions. Moreover, the proposed DSVD-PHAT approach

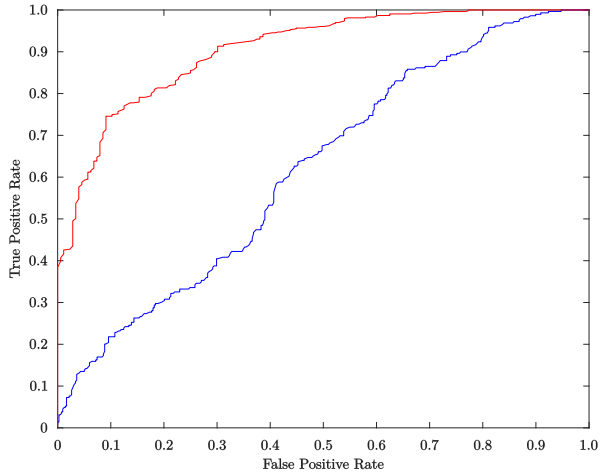


Fig. 4. ROC curves for GSVD-MUSIC (blue) and DSVD-PHAT (red).

provides better results for all scenarios where the SNR is greater or equal to 5 dB, at all reverberation levels.

TABLE III
AUC OF THE ROC CURVES

SNR (dB)	RT60 (msec)	GSVD-MUSIC	DSVD-PHAT
−10	200	**0.68**	0.64
	400	**0.55**	0.49
	600	**0.52**	0.47
	800	**0.51**	0.45
−5	200	**0.77**	0.75
	400	**0.66**	0.62
	600	**0.59**	0.55
	800	0.52	**0.53**
0	200	**0.84**	**0.84**
	400	0.70	**0.71**
	600	**0.66**	0.65
	800	0.63	**0.64**
5	200	0.87	**0.91**
	400	0.73	**0.76**
	600	0.71	**0.74**
	800	**0.64**	**0.64**
10	200	0.93	**0.95**
	400	0.76	**0.84**
	600	0.71	**0.77**
	800	0.64	**0.69**
15	200	0.93	**0.98**
	400	0.80	**0.86**
	600	0.73	**0.80**
	800	0.66	**0.69**
20	200	0.95	**0.99**
	400	0.76	**0.84**
	600	0.70	**0.71**
	800	0.64	**0.65**

Both methods are compared in terms of the execution times per frame. These methods run in the MATLAB environment, and their implementation relies mostly on vectorization to speed up processing. The hardware used

consists of an Intel Xeon CPU E5-1620 clocked at 3.70GHz. Table IV shows the average execution time per frame. This demonstrates the significant efficiency gain with DSVD-PHAT that avoids the expensive online SVD computations, as it runs approximately 250 times faster than GSVD-MUSIC. In this experiment, with $\Delta N/f_S = 8$ msecs between each frame, GSVD-MUSIC requires roughly 300% of the actual computing resources to achieve real-time, whereas DSVD-PHAT easily meets real-time requirements by using only 1% of the computing power.

TABLE IV
EXECUTION TIME PER FRAME

Method	GSVD-MUSIC	DSVD-PHAT
Time (msecs)	23.3	0.093

VI. CONCLUSION

This paper introduces a variant of the SVD-PHAT method to improve noise robustness. Results demonstrate that the proposed method performs similarly to the state of the art GSVD-MUSIC technique, but runs approximately 250 times faster. This makes DSVD-PHAT appealing for localization on robots with limited on-board computing power.

In future work, we will investigate multiple sound source localization with the proposed DSVD-PHAT method. Moreover, DSVD-PHAT could be incorporated to existing SSL frameworks such as HARK[2] [27] and ODAS[3] [17].

REFERENCES

[1] H. G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational auditory scene analysis and its application to robot audition," in *Proceedings of the International Conference on Informatics Research for Development of Knowledge Society Infrastructure*. IEEE, 2004, pp. 73–80.

[2] G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, and J.-I. Imura, "Robust ego noise suppression of a robot," in *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2010, pp. 62–71.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] 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 *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 2027–2032.

[5] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proceedings of the IEEE/RSJ international conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 664–669.

[6] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent sound source localization and its application to multimodal human tracking," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 143–148.

[7] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *Proceedings of the IEEE International Conference on Signal Processing, Communication and Computing*. IEEE, 2012, pp. 125–130.

[8] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 694–699.

[9] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1902–1907.

[10] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1997, pp. 375–378.

[11] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2. IEEE, 2003, pp. 1228–1233.

[12] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 1. IEEE, 2004, pp. 1033–1038.

[13] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 4. IEEE, 2006, pp. 841–844.

[14] ——, "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.

[15] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 2033–2038.

[16] 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.

[17] 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.

[18] F. Grondin and J. Glass, "SVD-PHAT: A fast sound source localization method," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signals Processing*, 2019.

[19] F. Grondin and F. Michaud, "Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2015, pp. 6149–6154.

[20] ——, "Noise mask for tdoa sound source localization of speech on mobile robots in noisy environments," in *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 2016, pp. 4530–4535.

[21] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, vol. 9, no. 1, pp. 12–15, 2002.

[22] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 958–963.

[23] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 6125–6129.

[24] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5210–5214.

[25] 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.

[26] J. B. Allen and D. A. 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.

[27] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system'hark'open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.

[2] http://hark.jp
[3] http://odas.io