Speaker Localization by Humanoid Robots in Reverberant Environments

Vladimir Tourbabin and Boaz Rafaely

Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
{tourbabv,br}@ee.bgu.ac.il

***Abstract*—One of the important tasks of a humanoid-robot auditory system is speaker localization. It is used for the construction of the surrounding acoustic scene and as an input for additional processing methods. Localization is usually required to operate indoors under high reverberation levels. Recently, an algorithm for speaker localization under these conditions was proposed. The algorithm uses a spherical microphone array and the processing is performed in the spherical harmonics domain, requiring a relatively large number of microphones to efficiently cover the entire frequency range of speech. However, the number of microphones in the auditory system of a humanoid robot is usually limited. The current paper proposes an improvement of the previously published algorithm. The improvement aims to overcome the frequency limitations imposed by the insufficient number of microphones. The improvement is achieved by using a novel space-domain distance algorithm that does not requires the transformation to the spherical harmonics domain, thereby avoiding the frequency range limitations. A numerical study shows two important results. The first is that, using the improved algorithm, the operation frequency range can be significantly extended. The second important result is related to the fact that higher frequencies contain more detailed information about the surrounding sound field. Hence, the additional higher frequencies lead to improved localization accuracy.**

I. Introduction

Auditory systems of humanoid robots have gained increased attention in recent years. These systems typically acquire the surrounding sound field by means of an array of microphones distributed on the surface of the robot's head. One of the fundamental processing objectives of these arrays is estimation of the direction of arrival (DOA) for speech sources, which is used in various tasks including sound source localization [1], spatial filtering [2], and automatic speech recognition [3]. DOA estimation in humanoid robots is usually performed indoors under potentially high reverberation levels, i.e. in a highly coherent environment. Recently, a method was developed enabling robust speaker localization in such highly coherent environments [4]. The method uses a spherical microphone array and the processing is performed in the spherical harmonics (SH) domain. It utilizes the direct path dominance (DPD) test in order to select the time-frequency (TF) bins that contain a major contribution from the direct sound, therefore significantly reducing the detrimental effect of the reverberation. In the final stage, the MUltiple SIgnal Classification (MUSIC) algorithm in the SH domain [5] is used to construct the spatial spectrum and obtain the DOA estimates. This method is suitable for humanoid robots because the head, on which the array is positioned, usually has a sphere-like shape [6]. However, efficient wide-band processing in the SH domain requires a relatively large number of microphones. In humanoid robots this number is usually limited, typically ranging from 2 [7] to, at most, 16 [8]. This reduces the frequency range that can be used and therefore limits the estimation performance.

The current paper proposes an improvement to the method described in [4]. The improvement enables the use of an extended frequency range when applied to arrays with a limited number of microphones. The improvement focuses on the final stages of the algorithm in [4]. First, we show that the DPD test can operate under the condition of spatial aliasing. Then, instead of MUSIC in the SH domain, a novel Space-Domain Distance (SDD) algorithm is utilized for the construction of the spatial spectrum. In this way, the spatial aliasing inherent in the transformation to the SH domain is avoided, removing the frequency range limitations. Increasing the frequency range offers two potential improvements. The first is enabling the use of higher frequency components, which is particularly helpful for signals with high frequency content. The second is that the higher frequency components of the field contain more detailed spatial information. Hence, utilizing higher frequencies is expected to increase the overall estimation accuracy.

The remainder of the paper is organized as follows. Section II briefly outlines the method described in [4]. Then, in Sections III and IV the proposed modification is described and its performance is evaluated numerically. Finally, Section V concludes the paper.

II. Background

The current section briefly summarizes the state-of-the-art method for DOA estimation introduced in [4]. The frequency limitations related to small numbers of microphones are discussed at the end of the section.

A. DOA estimation with spherical arrays in highly coherent environments

Consider a sound field produced by a broad-band acoustic source inside an enclosure. Suppose that the sound field is sampled by an array of M microphones distributed on the head surface of a humanoid robot. Denote the time-sampled outputs from all the microphones by $\mathbf{p}(t) = [p_1(t)\, p_2(t)\, \cdots\, p_M(t)]^T$, where $(\cdot)^T$ stands for the transpose operator. The microphone outputs are fed into the algorithm depicted in Fig. 1. The first

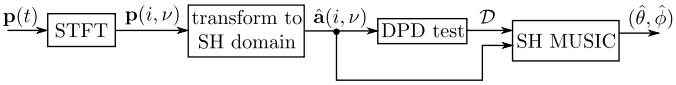

Fig. 1. Flowchart of the algorithm introduced in [4].

step is the short-time Fourier transform (STFT) of the outputs, which is defined as:

$$\mathbf{p}(i,\nu) = \sum_{t=0}^{T-1} \mathbf{p}(t+iD)w(t)e^{-j\frac{2\pi}{T}t\nu}, \quad i = 0,1,...,I-1,\ \nu = 0,1,...,T-1, \tag{1}$$

where i and ν denote the time and frequency indices, respectively, T is the STFT frame length, D is the offset between adjacent frames, $w(t)$ is an optional window function, and $j = \sqrt{-1}$.

The second step is based on the following model that relates the surrounding sound field to the STFTs of the microphone outputs [6]:

$$\mathbf{p}(i,\nu) = \mathbf{V}(\nu)\mathbf{a}(i,\nu) + \mathbf{n}(i,\nu), \tag{2}$$

where $\mathbf{a}(i,\nu) = [a_{0,0}(i,\nu)\, a_{1,-1}(i,\nu)\, \cdots\, a_{N,N}(i,\nu)]^T$ represents the surrounding sound field and holds the spherical Fourier transform (SFT) [9] coefficients of the plane-wave density (PWD) function of the field. In (2), the PWD function of the sound field is assumed to have a limited SH order N. Assuming that the robot head has a sphere-like shape, the effective SH order can be expressed as [10] $N = \lceil kr \rceil$, where $\lceil\cdot\rceil$ denotes the ceiling operator, r is the average head radius, and $k = \frac{2\pi f_s}{Tc}\nu$ is the wave number, with f_s and c denoting the sampling frequency and the speed of sound, respectively. Vector $\mathbf{n}(i,\nu)$ holds the STFT of the additive noise component present in the measurements. Matrix

$$\mathbf{V}(\nu) = [\mathbf{v}^*_{0,0}(\nu)\, \mathbf{v}^*_{1,-1}(\nu)\, \cdots\, \mathbf{v}^*_{N,N}(\nu)] \in \mathbb{C}^{M\times(N+1)^2}, \tag{3}$$

where $(\cdot)^*$ denotes the complex-conjugate operator and $\mathbf{v}_{nm}(\nu)$ holds the SFT coefficient of the conjugate of the array steering vector of order n and degree m. In practice, these can be obtained by measurements [8] or by numerical simulations [11]. Using the model in (2), the SFT coefficients of the PWD function are estimated by:

$$\hat{\mathbf{a}}(i,\nu) = \mathbf{V}^\dagger(\nu)\mathbf{p}(i,\nu), \tag{4}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudo-inverse operator.

The third step is the DPD test. This test operates in the SH domain and enables the selection of the TF bins, (i,ν), in which the PWD function contains only the direct sound. The test is based on the decoupling of the frequency and space dependencies that is an inherent feature of the SH domain. In order to see this, consider a sound field produced by a single source inside an enclosure. Assume that the field can be approximated by S plane waves travelling in directions related to the direction of the source and the dominant reflections, while the remaining reflections can be treated as a part of the additive noise. In this case, the PWD function estimated using (4) is related to the arrival directions via [4]:

$$\hat{\mathbf{a}}(i,\nu) = \mathbf{Y}\mathbf{s}(i,\nu) + \hat{\mathbf{n}}(i,\nu), \tag{5}$$

where $\mathbf{s}(i,\nu) = [s_1(i,\nu)\, \cdots\, s_S(i,\nu)]^T$ contains the amplitudes of the plane waves at the time frame indexed by i and the frequency indexed by ν. Vector $\hat{\mathbf{n}}(i,\nu) = \mathbf{V}^\dagger(\nu)\mathbf{n}(i,\nu)$ is the additive noise in the SH domain. Matrix $\mathbf{Y} = [\mathbf{y}(\theta_0,\phi_0)\, \cdots\, \mathbf{y}(\theta_S,\phi_S)]$ has the columns given by $\mathbf{y}(\theta_k,\phi_k) = [Y_0^0(\theta_k,\phi_k)\, Y_1^{-1}(\theta_k,\phi_k)\, \cdots\, Y_N^N(\theta_k,\phi_k)]^H$, where $\{(\theta_k,\phi_k)\}_{k=1}^S$ describe the elevation θ and the azimuth ϕ of the plane-wave arrival directions in the standard spherical coordinate system [9], $Y_n^m(\cdot,\cdot)$ is the spherical harmonic function, and $(\cdot)^H$ is the conjugate-transpose operator. The selection criterion for the direct-path TF bins utilizes the model in (5). The criterion is based on the individual covariance matrix $\hat{\mathbf{Q}}(i,\nu)$ of each TF bin, which is estimated using the adjacent bins, as follows:

$$\hat{\mathbf{Q}}(i,\nu) = \frac{1}{|\mathcal{A}_{i,\nu}|}\sum_{(k,l)\in\mathcal{A}_{i,\nu}} \hat{\mathbf{a}}(k,l)\hat{\mathbf{a}}^H(k,l), \tag{6}$$

where $\mathcal{A}_{i,\nu} = \{(k,l)\,|\,k = i-K,...,i,\ l = \nu-L,...,\nu\}$ denotes a small neighbourhood around the TF bin (i,ν) and $|\mathcal{A}_{i,\nu}| = (K+1)(L+1)$ is the cardinality of $\mathcal{A}_{i,\nu}$. Note that the SH-domain steering matrix $\mathbf{Y}$ defined in (5) does not depend on frequency. This implies that if all the TF bins in the neighbourhood of (i,ν) correspond to a plane wave travelling in the same direction and if the effect of the additive noise can be ignored, the rank of $\hat{\mathbf{Q}}(i,\nu)$ is expected to be unity. Furthermore, the direction of the plane wave is expected to correspond to the direct sound, because the direct sound is expected to arrive before the reflections. Based on these observations, the direct-path TF bins are determined by selecting the pairs of (i,ν) with the associated covariance matrices that have an effective rank of unity. In particular, the set of direct-path TF bins is determined as:

$$\mathcal{D} = \{(i,\nu)\,|\,\sigma_1(i,\nu)/\sigma_2(i,\nu) \geq THR\}, \tag{7}$$

where $\sigma_1(i,\nu)$ and $\sigma_2(i,\nu)$ are the largest and second largest singular values of $\hat{\mathbf{Q}}(i,\nu)$, respectively, and THR is a threshold value that determines whether the effective rank of a matrix is unity. For further details on the DPD test the reader is referred to [4].

In the final step, the DOAs are estimated in the SH domain

by applying the MUSIC algorithm to the set of the selected TF bins $\mathcal{D}$. Again, this algorithm utilizes the fact that in the SH domain the steering matrix $\mathbf{Y}$ does not depend on frequency, which implies that the covariance matrix of $\hat{\mathbf{a}}(i,\nu)$ can be estimated by averaging over all the bins in $\mathcal{D}$:

$$\hat{\mathbf{Q}} = \frac{1}{|\mathcal{D}|} \sum_{(i,\nu)\in\mathcal{D}} \hat{\mathbf{a}}(i,\nu)\hat{\mathbf{a}}^H(i,\nu). \tag{8}$$

The noise subspace is constructed by removing the eigenvector related to the largest eigenvalue of $\hat{\mathbf{Q}}$. This is based on the assumption that the direct-path component has the highest energy. Next, the MUSIC spectrum $P(\theta,\phi)$ is constructed [12] and the DOA of the source is estimated by selecting the values of (θ,ϕ) corresponding to the peak of the spectrum.

The algorithm described above enables DOA estimation in a reverberant environment. Its frequency-range limitations related to the limited number of microphones are discussed in the following subsection.

B. Frequency-range limitations

The final stage of the algorithm described above requires reliable estimates of the PWD function $\mathbf{a}(i,\nu)$. A major factor that can potentially degrade the quality of these estimates is spatial aliasing. In order to see this, recall that the ability to describe a sound field from its samples depends on the complexity of the sound field and on the spatial sampling scheme, which is the number and the positioning of the microphones. The complexity of the sound field on the surface of a robot's head can be interpreted as the SH order of the field. As was mentioned above, the effective SH order of the field is approximately $N = \lceil kr \rceil$. In order to capture a sound field of order N using the relatively efficient nearly-uniform sampling scheme [13], one would practically require at least $1.5(N+1)^2$ microphones [10]. Hence, for example, a medium-size array of 6 microphones and radius 7 cm would enable the estimation of $\mathbf{a}(i,\nu)$ only up to the frequency of 800 Hz. This frequency range may be further reduced because the microphones will not always be distributed in a way that minimizes the spatial aliasing [11]. This is in contrast to the available frequency content, which can extend up to 5 kHz or more, when considering a speech source. This discussion illustrates the limitation on the frequency range imposed by an insufficient number of microphones. The following section presents an approach for overcoming this limitation.

III. The Proposed Approach

In this section, a modification of the above-described algorithm that overcomes the frequency-range limitations is proposed. The modified algorithm is summarized in Fig. 2, where the modification is highlighted using the bold blue lines. The dashed blue line around the DPD test signifies the fact that a new analysis of this part is presented in the current paper. The modification focuses on the final stage, where the set $\mathcal{D}$ and the STFTs $\mathbf{p}(i,\nu)$ are fed into the SDD algorithm to produce the DOA estimate. The following subsection outlines the SDD algorithm. The second subsection discusses the effect of spatial aliasing on the proposed algorithm.

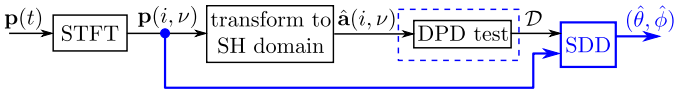

Fig. 2. Flowchart of the proposed algorithm.

A. SDD algorithm

The SDD algorithm is based on the following measure of distance between vectors:

$$\begin{aligned} d(\mathbf{v}_1,\mathbf{v}_2) &= \frac{\min_{\alpha} \|\mathbf{v}_1 - \alpha\mathbf{v}_2\|^2}{\|\mathbf{v}_1\|^2} \\ &= \frac{1}{\|\mathbf{v}_1\|^2} \left\| \left(\mathbf{I} - \frac{\mathbf{v}_2\mathbf{v}_2^H}{\|\mathbf{v}_2\|^2} \right) \mathbf{v}_1 \right\|^2 . \end{aligned} \tag{9}$$

It can be shown that $d(\cdot,\cdot)$ is symmetric and measures the tangent of the angle between the vectors. Using this measure, the *space-domain distance* between the measured STFT in a given direct-path TF bin $\mathbf{p}(i,\nu)$ and the STFT in the same bin that would be obtained if the arrival direction was (θ,ϕ) can be evaluated as $d(\mathbf{p}(i,\nu), \mathbf{V}(i,\nu)\mathbf{y}(\theta,\phi))$. The minimization over a scalar α is included to account for the unknown source amplitude. Using this measure, a spatial spectrum can be defined that combines the information from all of the direct-path TF bins:

$$Z(\theta,\phi) = \left(\sum_{(i,\nu)\in\mathcal{D}} d\left(\mathbf{p}(i,\nu), \mathbf{V}(i,\nu)\mathbf{y}(\theta,\phi)\right) \right)^{-1} . \tag{10}$$

Using (10), the DOA estimate can be obtained by performing a 2D search and selecting the direction that corresponds to the maximum of $Z(\theta,\phi)$:

$$(\hat{\theta},\hat{\phi}) = \underset{(\theta,\phi)}{\operatorname{argmax}}\, Z(\theta,\phi). \tag{11}$$

B. The effect of spatial aliasing

Recall that the motivation for the proposed modification was to extend the operation frequency range and to exploit the high SH order information, while avoiding the spatial aliasing associated with the transformation to the SH domain. For this purpose, the SDD algorithm has a major advantage. In order to illustrate this, first define the array order N' as the maximum SH order of the sound field that can be estimated without aliasing using the array. Next, note that the SH order of the steering vector $\mathbf{y}(\theta,\phi)$ in (10) can be chosen to incorporate the desired order information in the estimation process. Denote this SH order by N_e. Hence, by increasing N_e in (10) beyond N', it is possible to exploit the high SH order information. At the same time, the algorithm operates directly on the STFT vectors in the space domain. Hence, regardless of frequency, the spatial aliasing associated with the transformation to the SH domain is avoided.

In contrast to the SDD algorithm, the DPD test performed in the third step of the proposed algorithm still requires estimates of $\mathbf{a}(i,\nu)$. To analyse the effect of aliasing on the performance of the DPD test, suppose that the neighbourhood $\mathcal{A}_{i,\nu}$ of the TF bin (i,ν) contains only the direct sound. In this neighbourhood, the PWD function is given by:

$$\mathbf{a}(k,l) = s_0(k,l)\mathbf{y}(\theta_0,\phi_0),\ (k,l) \in \mathcal{A}_{i,\nu}, \tag{12}$$

where (θ_0,ϕ_0) denotes the arrival direction of the direct sound and $s_0(k,l)$ denotes the source amplitude in the time frame k and at frequency l. Ignoring the noise, the STFT of the array output in the same neighbourhood is given by:

$$\mathbf{p}(k,l) = s_0(k,l)\mathbf{V}(k,l)\mathbf{y}(\theta_0,\phi_0),\ (k,l) \in \mathcal{A}_{i,\nu}. \tag{13}$$

The PWD function is estimated from the measurements using (4):

$$\tilde{\mathbf{a}}(k,l) = s_0(k,l)\mathbf{V}^{\dagger}(k,l)\mathbf{V}(k,l)\mathbf{y}(\theta_0,\phi_0),\ (k,l) \in \mathcal{A}_{i,\nu}. \tag{14}$$

Note that, in general, $\mathbf{M}(k,l) = \mathbf{V}^{\dagger}(k,l)\mathbf{V}(k,l)$ is not an identity matrix. This results in an aliasing error indicated by the tilde in $\tilde{\mathbf{a}}(k,l)$. Now, assume that within the neighbourhood $\mathcal{A}_{i,\nu}$ the matrix $\mathbf{M}(k,l)$ is independent of k and l, i.e.

$$\mathbf{M}(k,l) = \mathbf{M},\ (k,l) \in \mathcal{A}_{i,\nu}. \tag{15}$$

The implications of this assumption will be discussed shortly. In this case the following holds:

$$\tilde{\mathbf{a}}(k,l) = s_0(k,l)\mathbf{M}\mathbf{y}(\theta_0,\phi_0). \tag{16}$$

Using the result in (15), it is straightforward to show that the rank of $\hat{\mathbf{Q}}(i,\nu)$ estimated using the aliased PWD function in (14) is unity. Hence, in the case where the assumption in (15) holds, the DPD test is expected to accurately provide the direct-path TF bins.

In order to analyse the validity of the assumption in (15), recall that, by definition, for $N \leq N'$ the estimates obtained using (14) are aliasing-free. In this case, the matrix $\mathbf{M}(k,l)$ is the identity matrix for all (k,l) in the neighbourhood $\mathcal{A}_{i,\nu}$ and the assumption in (15) holds. On the other hand, for $N >> N'$ the matrices $\mathbf{M}(k,l)$ differ significantly from the identity matrix and the validity of the assumption in (15) is not guaranteed. Nevertheless, as is shown by the simulation example below, there is a range of $N > N'$ for which the assumption holds with reasonable accuracy, thereby enabling the use of elevated frequencies.

IV. Simulation example

In this section, we present the results of a numerical performance analysis of the proposed algorithm. The analysis was based on an array of 6 microphones distributed nearly-uniformly [13] on the surface of a rigid sphere. The radius of the sphere was 7 cm. The array was assumed to be positioned inside an enclosure with the dimensions $6 \times 6 \times 4$ meters with the walls having reflection coefficients of 0.85. Microphone outputs were simulated by summing up the direct path and the appropriately delayed and attenuated reflections. Only the reflections that arrived within the first 50 ms were summed. The remaining reflections were considered to be a part of the additive noise component. The delays and the attenuations were obtained using the image method [14]. The direct path and each reflection were generated by filtering the source signal with the appropriate free-field microphone impulse response. The responses were calculated analytically [10]. Speech signals from the TIMIT database [15] were downsampled to 10 kHz and used as the source signals.

Fig. 3 presents an example of DOA estimation. This example is based on the STFTs in the frequency range of $4000 - 4500$ Hz. In this frequency range, the effective SH order of the field is $N = 6$. This is well beyond the SH order of the array, which is only $N' = 1$. Hence, the SH-domain distances $\{d\left(\tilde{\mathbf{a}}(i,\nu), \mathbf{y}(\theta_0,\phi_0)\right) \mid (i,\nu) \in \mathcal{D}\}$ are relatively large, as demonstrated by the histogram in the upper-left corner of Fig. 3. Nevertheless, as can be seen from the histogram in the upper-right corner of Fig. 3, the distances calculated for the same TF bins in the space domain, $\{d\left(\mathbf{p}(i,\nu), \mathbf{V}(i,\nu)\mathbf{y}(\theta_0,\phi_0)\right) \mid (i,\nu) \in \mathcal{D}\}$ using $N_e = 6$, are considerably lower. This is supported by the relatively small DOA estimation error of the SDD algorithm, $\delta = 1^\circ$, as compared to the large error, $\delta = 82^\circ$, of the SH MUSIC algorithm performed on the aliased direct-path TF bins.

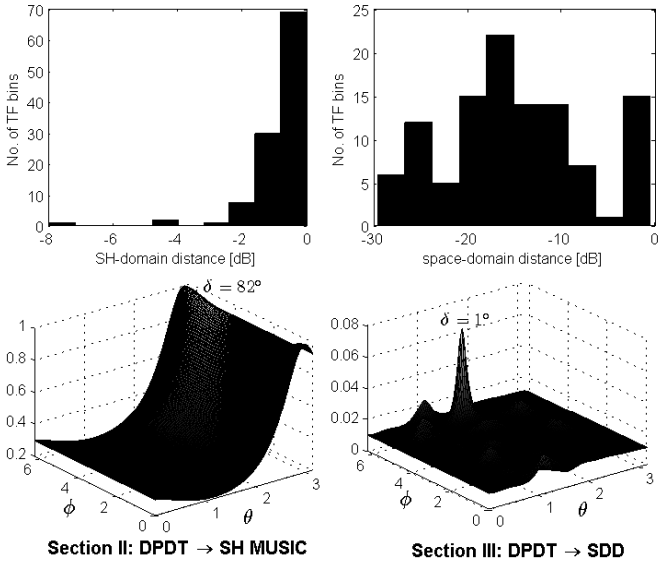

Fig. 3. An example of DOA estimation results for $N > N'$: left column - histogram of the SH-domain distances and the SH MUSIC spectrum, right column - histogram of the space-domain distances and the SDD spectrum. Error δ is defined as the angle between the true and the estimated direction.

The performance of the proposed algorithm was assessed by calculating the standard deviation of the angle between the estimated and true arrival directions, δ, as a function of frequency and the SH order N_e used for the estimation. The results are compared to the performance of the SH MUSIC algorithm. The DPD test in both algorithms used the PWD

function of an order equal to the array order $N' = 1$. The dimensions of the TF neighbourhood in (6) were 2×2. The value of THR in (7) was chosen such that 1.5% of all TF bins were recognized as the direct-path TF bins. The power of the additive noise was selected such that the wide-band signal to noise ratio was 20 dB.

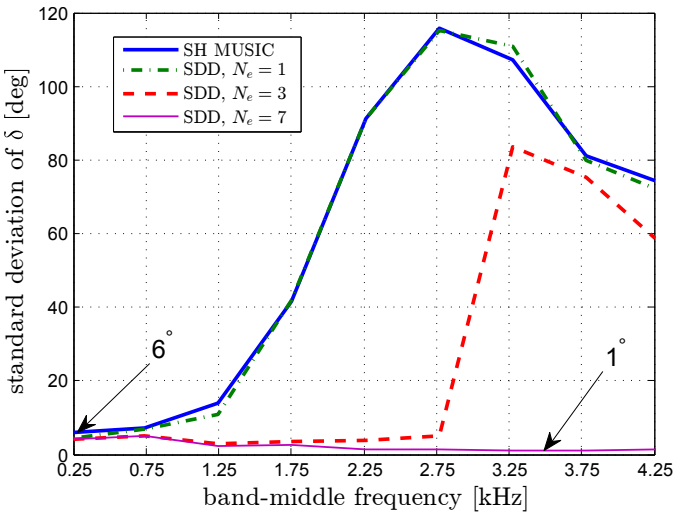

Fig. 4. DOA estimation performance of the MUSIC and SDD algorithms as a function of frequency and the SH order.

It can be seen that the MUSIC algorithm performs well at low frequencies, where the field is relatively simple and can be effectively described by the first SH order. The performance degrades at higher frequencies, as expected. Regarding the performance of the SDD algorithm, two important observations can be made. The first is that at low frequencies, the SDD algorithm performs similarly to the MUSIC algorithm. However, at higher frequencies, when the high order information is incorporated by choosing $N_e = 3$ and $N_e = 7$, the SDD algorithm significantly outperforms the MUSIC algorithm. The second observation is that the error of the SDD algorithm at high frequencies is about 1°. This is significantly lower than the error of the MUSIC algorithm at low frequencies, which is about 6°. This is believed to be due to the more detailed information that is available at higher frequencies.

V. Conclusion

In the current paper, a speaker localization algorithm is proposed that can be used by humanoid robots in highly reverberant environments. This is an improved version of the algorithm introduced previously [4]. The improved algorithm aims to extend the operation frequency range when applied to arrays with a limited number of microphones. The results of the numerical simulation showed that the proposed algorithm can significantly extend the operation frequency range and improve the localization performance. Future work may include an extension of the algorithm to account for multiple sources and a more thorough investigation of the localization performance.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

References

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

[2] Y. Peled and B. Rafaely, "Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2532–2540, Dec 2013.

[3] K. Nakamura, K. Nakadai, and H. G. 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.

[4] 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 Trans. Audio, Speech, Language Process.*, no. 99, 2014.

[5] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *IEEE Workshop Applications Signal Processing Audio and Acoustics (WASPAA).*, Oct 2009, pp. 221–224.

[6] V. Tourbabin and B. Rafaely, "Utilizing motion in humanoid robots to enhance spatial information recorded by microphone arrays," in *Joint Workshop Hands-free Speech Communication and Microphone Arrays (HSCMA), Nancy, France*, May 2014, pp. 147–151.

[7] A. Skaf and P. Danes, "Optimal positioning of a binaural sensor on a humanoid head for sound source localization," in *2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Oct 2011, pp. 165–170.

[8] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Adaptive blind source separation with HRTFs beamforming preprocessing," in *SAM 2012*, June 2012, pp. 269–272.

[9] G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*. Elsevier, 2005.

[10] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan 2005.

[11] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1803–1814, Dec 2014.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[13] R. H. Hardin and N. J. A. Sloane, "Mclarens improved snub cube and other new spherical designs in three dimensions," *Discrete and Computational Geometry*, vol. 15, pp. 429–441, 1996.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. S. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus," CD-ROM, 1993.