Direction of arrival estimation using microphone array processing for moving humanoid robots

V. Tourbabin*, *Student Member, IEEE* and B. Rafaely, *Senior Member, IEEE*

***Abstract*—The auditory system of humanoid robots has gained increased attention in recent years. This system typically acquires the surrounding sound field by means of a microphone array. Signals acquired by the array are then processed using various methods. One of the widely applied methods is direction of arrival estimation. The conventional direction of arrival estimation methods assume that the array is fixed at a given position during the estimation. However, this is not necessarily true for an array installed on a moving humanoid robot. The array motion, if not accounted for appropriately, can introduce a significant error in the estimated direction of arrival. The current paper presents a signal model that takes the motion into account. Based on this model, two processing methods are proposed. The first one compensates for the motion of the robot. The second method is applicable to periodic signals and utilizes the motion in order to enhance the performance to a level beyond that of a stationary array. Numerical simulations and an experimental study are provided, demonstrating that the motion compensation method almost eliminates the motion-related error. It is also demonstrated that by using the motion-based enhancement method it is possible to improve the direction of arrival estimation performance, as compared to that obtained when using a stationary array.**

***Index Terms*—direction of arrival estimation, microphone array, moving array, robot audition, rotation, translation, spherical harmonics.**

EDICS Category: AUD-ASAP

I. Introduction

AUDITION is an essential part of humanoid robots; it facilitates the robot's communication with the environment by exploiting the surrounding sound field. The auditory system of a humanoid robot is typically comprised of a microphone array and a set of signal processing methods. One of the widely applied array processing methods is Direction of Arrival (DoA) estimation. This method is used for sound source localization [1], spatial filtering and dereverberation [2], as well as for preprocessing in Automatic Speech Recognition (ASR) systems [3]. Several different approaches for DoA estimation are used in the literature, including time-delay-based algorithms [4], beamforming [5], maximum likelihood estimators [6] and subspace-based algorithms [7], [8]. These algorithms usually assume that the microphone array is fixed at a given position. However, a microphone array installed on a robot is not necessarily fixed; it can move in accordance with the robot's activity. Motion of the robot poses several challenges for DoA estimation algorithms. One of the problems, associated with motion and addressed in the literature, is the ego-motion noise originating from robot motors and joint movements [9]. Another problem, addressed in the current paper, is related to the motion of the array relative to the sound field. For a moving array, the DoAs to be estimated are continuously changing during the data acquisition. Thus, direct application of existing DoA estimation methods, based on stationary array models, may lead to severe degradation in performance.

One possible approach that can overcome this problem is the "*stop-perceive-act*" principle [10], which suggests stopping the robot while acquiring data for the estimation. This approach can also reduce the ego-motion noise. However, it may prevent the robot from receiving new commands while moving as a response to a previous command, thereby imposing behavioral constraints that may limit natural robot interaction with the environment. An alternative approach is to take the motion into account by utilizing the information about the array position, speed and acceleration, as a function of time. This information can be obtained from the operating system of the robot. Incorporating this information into the processing model can reduce the DoA estimation errors originating from array motion. Moreover, the motion can be utilized to improve the performance beyond that of a stationary array. For example, it is possible to reduce spatial aliasing in beamforming by using a linear array moving with constant acceleration [11] or by using a planar array rotating with constant angular velocity [12]. It is also possible to improve the DoA estimation resolution by using the synthetic aperture technique, which has been applied to underwater linear arrays towed at a constant speed (see, for example, [13]). In addition, when the sampling of a field is considered, the motion of the sensors can be utilized to reduce the reconstruction error [14]. The above mentioned techniques provide an insight into a possible approach for the processing of moving microphone arrays for robot audition. However, as mentioned above, these techniques are usually limited to linear arrays moving along a straight line or planar arrays rotating with constant velocity. Therefore, these techniques cannot be directly applied to DoA estimation using arbitrarily moving arrays of general geometry, which is usually the case in the context of humanoid robot audition. This problem is addressed in the current paper.

The approach adopted here exploits the sphere-like shape of the humanoid robot head, which facilitates processing in the Spherical Harmonic (SH) domain. It is shown that processing in the SH domain, in contrast to the more traditional space-domain processing, enables a convenient representation of robot's motion by a sequence of relatively simple linear transformations. This linear representation is utilized here

The authors are with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Be'er-Sheva 84105, Israel (email: {tourbabv,br}@ee.bgu.ac.il)

to reduce the motion-induced error in DoA estimation. In addition, an approach is presented for periodic signals that improves the DoA estimation beyond that of a stationary array. The improvement is achieved by combining the measurements from different spatial positions in a way similar to the synthetic aperture technique [15]. It is emphasized that in the SH domain, the motion-aware processing is applied to the Plane-Wave Density (PWD) function of the surrounding sound field, which is obtained prior to the DoA estimation. Hence, the proposed methods are generic, i.e., they can be incorporated into various DoA estimation techniques.

The remainder of the paper is organized as follows. Section II introduces the stationary array model and the DoA estimation algorithm to be used for the evaluation of the proposed approach. Then, in sections III and IV, the proposed approach is presented and analyzed. Sections V and VI provide numerical investigations of the two proposed methods. Section VII validates the proposed methods with experimental data using a humanoid robot. Finally, section VIII concludes the paper.

II. Background

This section starts with the derivation of a signal model for a stationary microphone array. A DoA estimation algorithm is also outlined in this section, for completeness. This algorithm will be used for assessing the performance of the proposed methods. Later, in section III, the signal model derived here will be extended to account for array motion.

A. Stationary array model

Consider an array of M microphones distributed on the surface of a robot head. The array center coincides with the origin of a standard spherical coordinate system [16] denoted by (r, θ, ϕ), representing the radius, elevation and azimuth, respectively. For the moment, suppose that the surrounding sound field is produced by a single far-field source located in the direction (θ_s, ϕ_s). In the absence of the array, the sound pressure produced by this source at the origin of the coordinate system is denoted by $a(t, \theta_s, \phi_s)$. Using this notation, the time-sampled pressure at all microphones $\mathbf{p}(t) = [p_1(t)\, p_2(t)\, \cdots\, p_M(t)]^T$ can be expressed as

$$\mathbf{p}(t) = \sum_{\tau=-\infty}^{\infty} \mathbf{h}(t-\tau, \theta_s, \phi_s) a(\tau, \theta_s, \phi_s) + \mathbf{n}(t), \quad (1)$$

where $(\cdot)^T$ denotes the transpose operator and $\mathbf{n}(t)$ represents additive noise. Vector $\mathbf{h}(t, \theta_s, \phi_s) = [h_1(t, \theta_s, \phi_s)\, h_2(t, \theta_s, \phi_s)\, \cdots\, h_M(t, \theta_s, \phi_s)]^T$ contains the impulse responses of the appropriate discrete Linear Time-Invariant (LTI) systems. These systems describe the transformation of the sound field from the origin of the coordinate system to the microphones. In the context of the human auditory system, the impulse response is usually referred to as the Head-Related Impulse Response (HRIR) [17]. In order to generalize the expression in (1), suppose that instead of a single source, the sound field is produced by an arbitrary distribution of far-field sources, with their relative contribution to the sound field at the origin given by $a(t, \theta, \phi)$. In this case, the output of the microphones can be obtained by integrating over all directions, i.e.,

$$\mathbf{p}(t) = \int_0^{2\pi} \int_0^{\pi} \sum_{\tau=-\infty}^{\infty} \mathbf{h}(t-\tau, \theta, \phi) a(\tau, \theta, \phi) \sin\theta d\theta d\phi + \mathbf{n}(t). \quad (2)$$

DoA estimation is usually performed using the Short-Time Fourier Transform (STFT) of the microphone outputs

$$\mathbf{p}(i, \omega) = \sum_{t=0}^{T-1} w(t) \mathbf{p}(t + iD) e^{-j\frac{2\pi}{T}\omega t}, \quad (3)$$

where T is the duration of the transform frame in samples, D is the offset between subsequent frames, $w(t)$ is a window of duration T, $j = \sqrt{-1}$, i is the index of the time frame, and $\omega = 0, 1, ..., T-1$ is the frequency bin index. By applying the STFT on both sides of (2) and using the Multiplicative Transfer Function (MTF) approximation [18], we obtain

$$\mathbf{p}(i, \omega) = \int_0^{2\pi} \int_0^{\pi} \mathbf{v}^*(\omega, \theta, \phi) a(i, \omega, \theta, \phi) \sin\theta d\theta d\phi + \mathbf{n}(i, \omega), \quad (4)$$

where $a(i, \omega, \theta, \phi)$ is the STFT of $a(t, \theta, \phi)$, being the PWD function of the sound field in the time frame i. Vector $\mathbf{v}^*(\omega, \theta, \phi)$ is the Fourier transform of $\mathbf{h}(t, \theta, \phi)$, holding the direction-dependent responses of all the microphones at a given frequency, and is known as the array steering vector in the array processing literature [19]. Finally, $\mathbf{n}(i, \omega)$ is the STFT of $\mathbf{n}(t)$ and $(\cdot)^*$ denotes the complex-conjugate operator. Recall that in (4) we use the MTF approximation, which implies that time-domain convolution between the source signal in each time frame and the impulse response of a system can be approximated as their multiplication in the STFT domain. This imposes a constraint on the duration of the time frames T, which should be sufficiently large compared to the duration of the HRIR, $\mathbf{h}(t)$ [18]. The duration of the human HRIR is in the order of milliseconds. Assuming a similar duration for the humanoid-robot HRIR, it is believed that a time frame longer than 10 ms should result in a reasonable approximation.

The integral in (4) can be rewritten as a sum in the SH domain using Parseval's theorem for the Spherical Fourier Transform (SFT) [16]:

$$\mathbf{p}(i, \omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \mathbf{v}_{nm}^*(\omega) a_{nm}(i, \omega) + \mathbf{n}(i, \omega), \quad (5)$$

where $a_{nm}(i, \omega)$ and $\mathbf{v}_{nm}(\omega)$ are the SFT coefficients of the PWD function and of the complex conjugate of the steering vector, respectively. Note that in (5) it is assumed that the SH order of the sound field on the robot head surface is limited to N. The maximum SH order contained in the field, N, is a function of frequency and depends on the geometry of the head surface. It may be difficult to derive the expression for N considering a general head geometry. Nevertheless, assuming that the head surface is close to spherical, it is suggested here to use the expression for the effective order of the pressure

on the surface of a rigid sphere, $N = \lceil kr \rceil$ [20], where $k = 2\pi f_s \omega / Tc$ is the wavenumber, $\lceil \cdot \rceil$ is the ceiling operator, c denotes the speed of sound, and f_s is the sampling frequency in Hz. The dependence of N on frequency is omitted for notation simplicity.

Note that (5) implies that the measured sound field is effectively represented by the first N orders of the PWD function. This allows us to rewrite (5) in a matrix form:

$$\mathbf{p}(i,\omega) = \mathbf{V}(\omega)\mathbf{a}(i,\omega) + \mathbf{n}(i,\omega), \tag{6}$$

where $\mathbf{V}(\omega) = [\mathbf{v}^*_{0,0}(\omega)\, \mathbf{v}^*_{1,-1}(\omega)\, \mathbf{v}^*_{1,0}(\omega) \cdots \mathbf{v}^*_{N,N}(\omega)] \in \mathbb{C}^{M\times(N+1)^2}$ and $\mathbf{a}(i,\omega) = [a_{0,0}(i,\omega)\, a_{1,-1}(i,\omega)\, a_{1,0}(i,\omega) \cdots a_{N,N}(i,\omega)]^T \in \mathbb{C}^{(N+1)^2\times 1}$. The model in (6) relates the sound field, represented by its PWD function in the SH domain $\mathbf{a}(i,\omega)$, to the array measurements $\mathbf{p}(i,\omega)$. Processing of the array based on this model requires knowledge of the array steering matrix $\mathbf{V}(\omega)$. In practice, this matrix can be obtained from measurements [21] or by numerical simulation [22]. The model in (6) provides the basis for the discussion in the following sections.

B. SH-MUSIC

This section outlines the Spherical Harmonics MUltiple SIgnal Classification (SH-MUSIC) [23] algorithm for DoA estimation in the SH domain. It is provided here for convenience as a reference algorithm for the evaluation of the methods proposed below.

Assume that a sound field produced by S spatially separated far-field sources is sampled using an array of $M > S$ microphones. The SH-MUSIC algorithm for the estimation of the DoAs of all the sources proceeds as follows:
Step 1*:* Obtain $J \geq S$ snapshots of the sound pressure at all microphones $\mathbf{p}(i,\omega)$, $i = 1, 2, ..., J$, by calculating the STFT of the microphone outputs, as defined in (3).
Step 2*:* Estimate the PWD functions $\mathbf{a}(i,\omega)$. Considering the model in (6), this step can be accomplished by

$$\hat{\mathbf{a}}(i,\omega) = \mathbf{V}^{\dagger}(\omega)\mathbf{p}(i,\omega), \tag{7}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse [24]. Recall that the sound field on the surface of the robot head is assumed to have a limited SH order N, implying that $\mathbf{V}(\omega) \in \mathbb{C}^{M\times(N+1)^2}$. The maximum SH order depends on frequency and, as suggested above, is given by $N = \lceil kr \rceil$. Hence, in the frequency range for which $M > (N+1)^2$, equation (7) will result in the Least-Squares sense estimation, while at higher frequencies, the estimation will be in the Minimum-Norm sense [24].
Step 3*:* Estimate the modal narrow-band covariance matrix at each frequency by averaging over time:

$$\mathbf{Q}(\omega) = \frac{1}{J}\sum_{i=1}^{J} \hat{\mathbf{a}}(i,\omega)\hat{\mathbf{a}}^H(i,\omega), \tag{8}$$

where $(\cdot)^H$ denotes the conjugate-transpose operator. Next, average the narrow-band covariance matrices over frequency:

$$\tilde{\mathbf{Q}} = \frac{1}{T}\sum_{\omega=1}^{T} \mathbf{Q}(\omega). \tag{9}$$

Note that the averaging in (9) can be performed over a selected subset of frequencies. The averaging over frequency (also known as frequency smoothing) preserves the spatial structure of the covariance matrix, due to the decoupling between space and frequency dependent parameters that is inherent to the SH domain [23].
Step 4: Estimate the noise subspace using $\tilde{\mathbf{Q}}$ and construct the MUSIC spectrum $P(\theta,\phi)$ [7]. Note that this step may require whitening of the noise covariance matrix [23].
Step 5: Obtain the DoA estimates by picking S values of (θ,ϕ) corresponding to the S greatest peaks of $P(\theta,\phi)$.

It is emphasized here that DoA estimation in the SH domain will usually start with the first two steps of the above described algorithm. The purpose of these steps is to estimate the SH coefficients of the PWD function to be used in steps $3-5$ for actual DoA estimation. The methods presented below aim to take array motion into account in the second step, thereby accounting for motion in any DoA estimation algorithm that uses the PWD function estimates, while SH-MUSIC serves here as an example of an algorithm for evaluation purposes.

III. Moving array model

The previous section presented an overview of a model for the processing of signals from stationary arrays. However, microphone arrays mounted on a moving humanoid robot are expected to move in accordance with the robot's activity. The current section is concerned with an extension of the model to account for the array motion, while the sources are assumed to be fixed at their positions.

Assume that the origin of the coordinate system is positioned at the array center and moves together with the array. Thus, the sound field, as viewed from this coordinate system, is continuously moving in the direction reciprocal to the array motion. Similarly to in the stationary array case, the STFT of the microphone outputs is calculated by dividing the time-line into frames of T samples each. The effect of motion within each time frame [25] is neglected because the speed of the robot is believed to be very low relative to the speed of sound. Nevertheless, the motion between the time frames is accounted for by adjusting the stationary array model:

$$\mathbf{p}(i,\omega) = \mathbf{V}_i(\omega)\mathbf{W}_i(\omega)\mathbf{a}(i,\omega) + \mathbf{n}(i,\omega), \tag{10}$$

where $\mathbf{a}(i,\omega)$ are the SH coefficients of the PWD function measured by a stationary array located at a reference position and $\mathbf{W}_i(\omega)$ describes the transformation of the sound field due to the array motion between the reference position and its position during the i^{th} time frame. Note that $\mathbf{W}_i(\omega)$ is not, in general, a square matrix, because the transformation may affect the SH order of the field. Moreover, the order of the transformed field may vary between the time frames. Hence, in contrast to the stationary case, here, the number of columns in the steering matrix $\mathbf{V}_i(\omega)$ depends on i, as

indicated by the subscript. Additional details on the SH order of the transformed field are provided in section III-B.

According to Chasles' theorem, the change in the array position induced by the motion of the robot between two time frames, can be divided into a rotation followed (or preceded) by a translation (see, for example, [26], page 42). This implies that the transformation matrix $\mathbf{W}_i(\omega)$ can be decomposed as

$$\mathbf{W}_i(\omega) = \mathbf{T}_i(\omega)\mathbf{R}_i(\omega), \tag{11}$$

where $\mathbf{R}_i(\omega)$ and $\mathbf{T}_i(\omega)$ describe the rotation and translation parts of the transformation, respectively. This idea is illustrated schematically in Fig. 1. In the SH domain, the matrices $\mathbf{R}_i(\omega)$

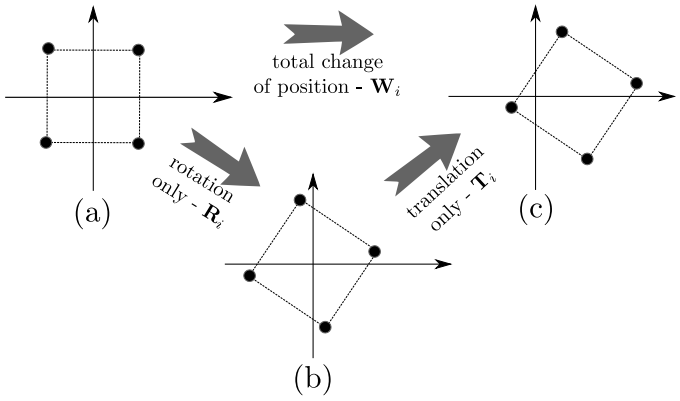

Fig. 1. An illustration of a rectangular 4-microphone array moving from the original position in (a) to the final position in (c). The motion is divided into a rotation from (a) to (b) and a subsequent translation from (b) to (c).

and $\mathbf{T}_i(\omega)$ have closed form expressions as functions of the desired rotation angles and translation vector, respectively. These expressions are provided and discussed in the following two subsections.

A. Array rotation

The rotation matrix $\mathbf{R}_i$ is given by the Wigner-D matrix [27]. This matrix is block-diagonal and unitary (see the appendix for the proof of unitarity) and is given by

$$\mathbf{R}_i = \begin{pmatrix} \mathbf{D}_0 & \mathbf{0}_{0,1} & \cdots & \mathbf{0}_{0,N} \\ \mathbf{0}_{1,0} & \mathbf{D}_1 & \cdots & \mathbf{0}_{1,N} \\ \vdots & & \ddots & \vdots \\ \mathbf{0}_{N,0} & \mathbf{0}_{N,1} & \cdots & \mathbf{D}_N \end{pmatrix}, \tag{12}$$

where $\mathbf{0}_{m_1,m_2}$ is a zero matrix having $2m_1+1$ rows and $2m_2+1$ columns and $\mathbf{D}_n \in \mathbb{C}^{(2n+1)\times(2n+1)}$ is given by

$$\mathbf{D}_n = \begin{pmatrix} D^n_{-n,-n} & \cdots & D^n_{-n,n} \\ \vdots & \ddots & \vdots \\ D^n_{n,-n} & \cdots & D^n_{n,n} \end{pmatrix}, \tag{13}$$

where $D^{m_3}_{m_1,m_2}$ is the short notation for the Wigner-D function [28] $D^{m_3}_{m_1,m_2}(\alpha_i,\beta_i,\gamma_i)$ with α_i, β_i and γ_i being the Euler angles [16] of the rotation in the i^{th} time frame. Note that $\mathbf{R}_i \in \mathbb{C}^{(N+1)^2\times(N+1)^2}$ is a square matrix, implying that rotation does not affect the SH order of the field. Furthermore, observe that the rotation matrix does not depend on frequency, i.e., the transformation of the SH coefficients due to a rotation by $(\alpha_i,\beta_i,\gamma_i)$ is identical for all frequency components. Hence, the dependence of $\mathbf{R}_i$ on ω will be omitted in the subsequent discussion. Further details on the Wigner-D function can be found in the appendix.

A special case of particular interest is when the head rotates in the horizontal plane. In this case, the rotation is completely specified by α_i, while $\beta_i = \gamma_i = 0$. Substituting into the expression for the Wigner-D function results in

$$\begin{aligned} D^{m_3}_{m_1,m_2}(\alpha_i,0,0) &= e^{-jm_1\alpha_i} d^{m3}_{m_1,m_2}(0) e^{-jm_20} \\ &= e^{-jm_1\alpha_i}\delta_{m_1,m_2}, \end{aligned} \tag{14}$$

where $d^{m3}_{m_1,m_2}(\cdot)$ and δ_{m_1,m_2} are the Wigner-d and Kronecker delta functions, respectively. The result in (14) implies that when the rotation is purely horizontal, the rotation matrix $\mathbf{R}_i$ is a diagonal matrix. Moreover, the diagonal values are obtained by a relatively simple evaluation of the exponential in (14). In addition, note that for $\alpha_i \to 0$ the diagonal term $e^{-jm_1\alpha_i} \to 1$. Hence, as would be expected, $\mathbf{R}_i$ converges to the identity matrix $\mathbf{I}$.

B. Array translation

The translation matrix $\mathbf{T}_i(\omega)$ describes the effect of the translation part of array motion between the reference and the i^{th} time frame. Denote the SH order and degree indices before the translation by n, m and after the translation by n', m'. The element of $\mathbf{T}_i(\omega)$ in row $n'^2+n'+m'$ and column n^2+n+m for far-field sources that are outside the measurement region, is given by [29]

$$[\mathbf{T}_i]^{n^2+n+m}_{n'^2+n'+m'} = \sum_{q=0}^{\lceil kr_i \rceil} j_q(kr_i)\cdot Y_q^{m-m'}(\theta_i,\phi_i)\cdot C^{n,m,q}_{n',m'}, \tag{15}$$

where $j_q(\cdot)$ and $Y_q^{m-m'}(\cdot)$ are the spherical Bessel and spherical harmonic functions, respectively, r_i and (θ_i,ϕ_i) are the distance and the direction of the translation in the i^{th} time window, and the dependence on ω is expressed through the wavenumber k, for convenience. Coefficient $C^{n,m,q}_{n',m'}$ is given by [29]

$$\begin{aligned} C^{n,m,q}_{n',m'} = 4\pi j^{(n'+q-n)}(-1)^m \sqrt{\frac{(2n+1)(2n'+1)(2q+1)}{4\pi}} \\ \times \begin{pmatrix} n & n' & q \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} n & n' & q \\ -m & m' & m-m' \end{pmatrix}, \end{aligned} \tag{16}$$

where $\begin{pmatrix} J_1 & J_2 & J_3 \\ m_1 & m_2 & m_3 \end{pmatrix}$ is the Wigner-3j symbol. The sum in (15) is limited to $\lceil kr_i \rceil$ because $j_q(kr_i) \approx 0$ for $q >> kr_i$ [20]. Note that the translation matrix $\mathbf{T}_i(\omega) \in \mathbb{C}^{(N'+1)^2\times(N+1)^2}$ is, in general, not a square matrix; it transforms a field of order N into a field of order N'. Using the property $C^{n,m,q}_{n',m'} = 0$ for $|n-n'| > q$, the new SH order N' can be expressed as

$$N' = N + \lceil kr_i \rceil. \tag{17}$$

This implies that a translation can effectively increase the SH order of the sound field by up to $\lceil kr_i \rceil$ orders.

Note that the complexity of the calculations required to obtain $\mathbf{T}_i(\omega)$ is proportional to the third power of $\lceil kr_i \rceil$. This is because the upper limit of the sum in (15) depends linearly on $\lceil kr_i \rceil$ and the number of rows in $\mathbf{T}_i(\omega)$ increases quadratically with increasing $\lceil kr_i \rceil$. Hence, especially for large translations, the calculation of all the entries of $\mathbf{T}_i(\omega)$ may become relatively complicated. Fortunately, in practice, the calculation of the translation matrix $\mathbf{T}_i(\omega)$ can be simplified. Taking the first frame $i = 1$ as a reference, the overall translation matrix from the reference frame to the i^{th} frame, can be expressed as

$$\mathbf{T}_i(\omega) = \prod_{l=1}^{i} \mathbf{T}'_l(\omega), \;\; i > 1 \tag{18}$$

where $\mathbf{T}'_l(\omega)$ denotes the translation matrix between the two subsequent frames l and $l-1$. Denote the translation vector between two subsequent time frames l and $l-1$ by $(r'_l, \theta'_l, \phi'_l)$. In practice, a relatively fast motion of 1 m/s results in $kr'_l < 1$ up to 5 kHz, considering an offset of 10 ms between the frames. Hence, for large translations, decomposing $\mathbf{T}_i(\omega)$ as suggested by (18) has the potential to reduce the complexity of the calculations. In particular, the expression for the element of $\mathbf{T}'_l(\omega)$ given in (15) becomes

$$\begin{aligned}[\mathbf{T}'_l]_{n'^2+n'+m}^{n^2+n+m} &= j_0(kr'_l) \cdot Y_0^{m-m'}(\theta'_l, \phi'_l) \cdot C_{n',m'}^{n,m,0} \\ &+ j_1(kr'_l) \cdot Y_1^{m-m'}(\theta'_l, \phi'_l) \cdot C_{n',m'}^{n,m,1}. \end{aligned} \tag{19}$$

Using properties of the Wigner-3j symbol it can be shown that the first term in (19) reduces to

$$Y_0^{m-m'}(\theta'_l, \phi'_l) \cdot C_{n',m'}^{n,m,0} = \delta_{n,n'}\delta_{m,m'}, \tag{20}$$

and for the second term it holds that

$$Y_1^{m-m'}(\theta'_l, \phi'_l) \cdot C_{n',m'}^{n,m,1} = 0, \;\; n = n'. \tag{21}$$

Thus, for $kr'_l < 1$, $\mathbf{T}'_l(\omega)$ can be calculated using the simplified form:

$$\mathbf{T}'_l(\omega) = j_0(kr'_l)\tilde{\mathbf{I}} + j_1(kr'_l)\mathbf{C}, \tag{22}$$

where $\tilde{\mathbf{I}}$ is an $(N'+1)^2 \times (N+1)^2$ matrix with the entries being 1 on the main diagonal and 0 otherwise. Matrix $\mathbf{C}$ is an off-diagonal matrix with the entries given by the second term in (19). It is interesting to note that for $r'_l \to 0$, functions $j_0(kr'_l) \to 1$ and $j_1(kr'_l) \to 0$ and $N' \to N$. This implies that for small translations the second term in (19) vanishes. Hence, consistently, $\mathbf{T}'_l(\omega) \to \mathbf{I}$.

To summarize, a moving array model was proposed in (10). The model accounts for motion by means of a linear transformation $\mathbf{W}_i(\omega)$, as detailed in sections III-A and III-B. This model is the basis for the development of the motion-aware processing methods outlined in the following section.

IV. Motion-aware processing

In the previous section, a signal model was presented that takes into account the motion of the robot. This model is utilized here to develop methods for the processing of signals from microphone arrays installed on a moving robot. First, a method that compensates for array motion is presented; it will be referred to as the *motion compensation* approach. Then, a method is presented that uses the robot motion in order to enhance the performance to a level beyond that of a stationary array. Hereafter, this will be referred to as the *motion-based enhancement* approach.

A. Motion compensation approach

Suppose that J consecutive samples of the STFT, i.e., $\mathbf{p}(i,\omega)$, $i = 1, 2, ..., J$, were acquired using a microphone array installed on a robot. It is assumed here that during the data acquisition the sources remain at the same positions and the motion of the robot does not involve transition between different rooms. As was mentioned above, signals from a microphone array installed on a moving humanoid robot can be processed using the *stop-perceive-act* principle that demands stopping the robot during data acquisition. Following this approach and assuming that the robot is stopped during data acquisition, the PWD function can be computed from $\mathbf{p}(i,\omega)$, $i = 1, 2, ..., J$ by using the stationary array model

$$\begin{aligned}\hat{\mathbf{a}}_s(i,\omega) &= \mathbf{V}^\dagger(\omega)\mathbf{p}_s(i,\omega) \\ &= \mathbf{a}(i,\omega) + \mathbf{V}^\dagger(\omega)\mathbf{n}(i,\omega), \end{aligned} \tag{23}$$

where $\mathbf{p}_s(i,\omega)$ denote the STFTs obtained by a stationary microphone array, as described by the model in (5). The major drawback of the *stop-perceive-act* principle is that it imposes a behavioral constraint on the robot and, therefore, can limit the naturalness of its interaction with the surroundings. An alternative approach, proposed in this paper, is to compensate for the motion by utilizing the moving array model presented in the previous section. Using this model, motion compensated estimation of the PWD function can be performed from the measurements obtained by a moving array:

$$\begin{aligned}\hat{\mathbf{a}}_m(i,\omega) &= \left[\mathbf{V}_i(\omega)\mathbf{W}_i(\omega)\right]^\dagger \mathbf{p}_m(i,\omega) \\ &= \mathbf{a}(i,\omega) + \left[\mathbf{V}_i(\omega)\mathbf{W}_i(\omega)\right]^\dagger \mathbf{n}(i,\omega), \end{aligned} \tag{24}$$

where $\mathbf{p}_m(i,\omega)$ denote the STFTs obtained by a moving microphone array and it is assumed that the measured noise $\mathbf{n}(i,\omega)$ is identical for moving and stationary arrays. It is emphasized that the transformation matrices $\mathbf{W}_i(\omega)$ are known; they can be calculated as explained in sections III-A and III-B based on knowledge of the robot's trajectory. Note that if no compensation for motion is applied, i.e., $\mathbf{p}_m(i,\omega)$ is processed in accordance with (23), then the obtained $\hat{\mathbf{a}}_m(i,\omega) = \mathbf{W}_i(\omega)\mathbf{a}(i,\omega) + \mathbf{V}^\dagger(\omega)\mathbf{n}(i,\omega)$, $i = 1, 2, ..., J$, describe different sound fields, as viewed from the array coordinate system at different points along the array trajectory. Hence, the performance of the subsequent DoA estimation using all of $\hat{\mathbf{a}}_s(i,\omega)$, $i = 1, 2, ..., J$, as outlined in *step 3* of the SH-MUSIC algorithm, will be degraded. This point is further demonstrated by the simulation examples in section V.

Note that the error components that appear when using the *stop-perceive-act* and the motion compensation approaches, i.e., the terms $\mathbf{V}^\dagger(\omega)\mathbf{n}_i(\omega)$ and $\left[\mathbf{V}_i(\omega)\mathbf{W}_i(\omega)\right]^\dagger \mathbf{n}_i(\omega)$, respectively, are, in general, different. Comparison of these terms can be difficult for a general $\mathbf{W}_i(\omega)$. However, for pure rotation the matrix $\mathbf{W}_i(\omega) = \mathbf{R}_i$ is unitary. Moreover, $\mathbf{V}_i(\omega) = \mathbf{V}(\omega)$,

because rotation does not affect the SH order of the field. In this case, it is straightforward to show that the power of the error terms, expressed by the trace of their covariance matrices, is identical. For the general case, recall that the rotation and translation parts, $\mathbf{R}_i$ and $\mathbf{T}_i(\omega)$, closely approximate the identity matrix for relatively small movements between the time frames. Therefore, it may be expected that $\mathbf{W}_i(\omega)$ has only a small effect on the power of the error term. Thus, the motion compensation approach is expected to maintain the same performance as the *stop-perceive-act* approach, while removing the constraints on the robot's behavior.

The following subsection presents a method that exploits the robot's motion in order to improve the array performance to a level beyond that of a stationary array.

B. Motion-based enhancement approach

Section III introduced a signal model for a moving array. Based on this model, an approach has been discussed in section IV-A that compensates for the motion in the estimation of the PWD function. It was assumed that the number of microphones and their distribution on the head surface enable estimation of the PWD function up to the required order N. However, in practice, the number of microphones may be limited. See, for example, the microphone arrays of humanoid robots NAO [30] and Hearbo [3], which have only 4 and 8 microphones, respectively. A small number of microphones limits the SH order of the PWD function that can be inferred; this limits the performance of the DoA estimation algorithms in terms of resolution, robustness, frequency range, and the number of sources that can be localized. In the current subsection, an approach is presented that combines the measurements from several time frames. In this way, the effective number of microphones is increased, thereby improving the performance of the subsequent DoA estimation.

Here, it is assumed that the STFTs of the microphone outputs were computed resulting in I pressure samples at each frequency, i.e., $\mathbf{p}(i,\omega)$, $i = 1, 2, ..., I$. Note that the number of frames is denoted here by I instead of J. The reason for this change of notation is discussed at the end of the section. The method presented here enables the combination of these I samples to produce a single estimate of the PWD function.

In addition to the assumptions made in the previous section, it is assumed here that the amplitudes of the signals produced by the sources at all frequencies of interest are constant in time, i.e., the amplitudes are the same for $i = 1, 2, ..., I$. Although this is a restrictive assumption, it appears to be relatively common in the signal processing literature [13], [31]. In section IV-C, it is demonstrated that the method proposed here is reasonably robust to variations in both the amplitude and the frequency, promoting its application in practice for the localization of sources that produce quasi-periodic signals, such as fire alarms or music.

In the case where the above assumption holds, the PWD function $\mathbf{a}(i,\omega)$, as measured from the coordinate system in a reference position, differs for various values of i only by the phase related to the time offset D between the frames:

$$\mathbf{a}(i,\omega) = \mathbf{a}(1,\omega)e^{j2\pi D\omega(i-1)/T}. \tag{25}$$

Note that in (25), the reference time frame was arbitrarily chosen to be the first frame, for convenience. By substituting (25) into the moving array model in (10), we obtain

$$\mathbf{p}(i,\omega) = \mathbf{V}_i(\omega)\mathbf{W}_i(\omega)\mathbf{a}(1,\omega)e^{j2\pi D\omega(i-1)/T} + \mathbf{n}(i,\omega). \tag{26}$$

Multiplying both sides of (26) by $e^{-j2\pi D\omega(i-1)/T}$ results in the following relation:

$$\tilde{\mathbf{p}}(i,\omega) = \mathbf{V}_i(\omega)\mathbf{W}_i(\omega)\mathbf{a}(1,\omega) + \tilde{\mathbf{n}}(i,\omega), \tag{27}$$

where

$$\tilde{\mathbf{p}}(i,\omega) = \mathbf{p}(i,\omega)e^{-j2\pi D\omega(i-1)/T} \tag{28}$$

and

$$\tilde{\mathbf{n}}(i,\omega) = \mathbf{n}(i,\omega)e^{-j2\pi D\omega(i-1)/T}. \tag{29}$$

The result in (27) relates the time-aligned STFTs, $\tilde{\mathbf{p}}(i,\omega)$, for $i = 1, 2, ..., I$, to the same PWD function $\mathbf{a}(1,\omega)$ that was captured by the array during the reference (first) time frame. Next, by concatenating the time-aligned STFTs into a single measurement vector, a combined system can be constructed:

$$\tilde{\mathbf{p}}(\omega) = \mathbf{A}(\omega)\mathbf{a}(1,\omega) + \tilde{\mathbf{n}}(\omega), \tag{30}$$

where $\tilde{\mathbf{p}}(\omega) = [\tilde{\mathbf{p}}^T(1,\omega), \tilde{\mathbf{p}}^T(2,\omega), ..., \tilde{\mathbf{p}}^T(I,\omega)]^T$ and $\tilde{\mathbf{n}}(\omega) = [\tilde{\mathbf{n}}^T(1,\omega), \tilde{\mathbf{n}}^T(2,\omega), ..., \tilde{\mathbf{n}}^T(I,\omega)]^T$ are column vectors with the dimensions $I \cdot M \times 1$. Matrix $\mathbf{A}(\omega)$ is given by

$$\mathbf{A}(\omega) = \begin{pmatrix} \mathbf{V}_1(\omega)\mathbf{W}_1(\omega) \\ \mathbf{V}_2(\omega)\mathbf{W}_2(\omega) \\ \vdots \\ \mathbf{V}_I(\omega)\mathbf{W}_I(\omega) \end{pmatrix} \in \mathbb{C}^{I\cdot M\times(N+1)^2}, \tag{31}$$

where $\mathbf{W}_1 = \mathbf{I}$ is the identity matrix; this is because the first frame was chosen to serve as the reference. Using the model in (30), the PWD function in the reference frame can be obtained by applying the pseudo inverse of the combined matrix, i.e.,

$$\hat{\mathbf{a}}(1,\omega) = \mathbf{A}^{\dagger}(\omega)\tilde{\mathbf{p}}(\omega). \tag{32}$$

Note that the motion-based enhancement approach described here provides a single estimate of the PWD function using I samples of the sound pressure. Thus, in order to obtain reliable estimates of the covariance matrices, an extended data acquisition time of $J \cdot I$ frames may be required, as compared to J frames that are required in the motion compensation approach. The advantage of the motion-based enhancement method is that it enables us to estimate $(N+1)^2$ SH coefficients of the PWD function by jointly using $I \cdot M$ equations, as compared to only M equations available when using the motion compensation method. Hence, when using the motion-based enhancement approach, the SH order N of the estimated PWD function can be increased. This is discussed in more detail in the next subsection.

C. Analysis of the motion-based enhancement approach

The motion-based enhancement approach presented in (32) enables the estimation of the PWD function by jointly using the information gathered by the array from I different time frames, which, due to the array motion, were obtained

at different array positions. The apparent advantage of this approach over the motion compensation approach presented in (24) is that the number of equations it provides for the estimation of the PWD function is I times greater. This is particularly true for relatively large movements for which $\mathbf{W}_i(\omega)$ significantly differ from each other. On the other hand, for small movements, matrices $\mathbf{W}_i(\omega) \rightarrow \mathbf{I}$, and $\mathbf{A}(\omega) \rightarrow [\mathbf{V}(\omega)^T\, \mathbf{V}(\omega)^T \cdots \mathbf{V}(\omega)^T]^T$, implying that no independent equations are added. Hence, it is important to gain an insight into the degree to which the motion-based enhancement approach can improve the estimation of the PWD function, and into the way in which the improvement depends on the array speed, the trajectory, and the frequency range.

The effective increase in the number of independent equations in (30) is assessed here through the *effective rank* of $\mathbf{A}(\omega)$ [32]. This quantity is based on the uniformity of the singular values of a matrix and, as opposed to the discrete values of the actual matrix rank, it provides a continuous estimate of the effective dimension of a system. This property makes it suitable for assessing the dimension of a system as a function of continuous parameters such as frequency, radius of translation, and angle of rotation [33]. In addition, this measure was shown to be related to array beamforming and DoA estimation performance [22], making it particularly suitable for the analysis presented here. First, the effective rank of $\mathbf{A}(\omega)$, combining $I = 2$ frames, was calculated as a function of the array displacement between the time frames. For simplicity, we used a rigid equiangular spherical array [20] of 13 microphones distributed with a $45°$ spacing in elevation and a $90°$ spacing in azimuth. The array radius is $r_a = 6$ cm and the SH order of the field is assumed to be limited to $N = 6$, implying that the number of columns in $\mathbf{A}(\omega)$ is $(N+1)^2 = 49$. In order to compute $\mathbf{A}(\omega)$, the steering matrix $\mathbf{V}(\omega)$ is required (see (6)). For a rigid spherical array, this matrix can be computed using an analytic expression [20]. The rotation and translation matrices were calculated using (12) and (15).

Two different modes of motion were considered: (a) rotation about the z axis by an angle α and (b) translation in the direction $(90°, 90°)$ by a distance r. The effective rank of $\mathbf{A}(\omega)$ at 2 kHz as a function of α and of r is presented in Fig. 2. The effective rank of the array in only the reference position is also plotted in order to provide a reference value. It can be seen that the effective rank of $\mathbf{A}(\omega)$ increases significantly for larger displacements (either rotation or translation). The increase in the effective rank is consistent with the fact that increasing the displacement is similar to increasing the effective array aperture. Therefore, it is expected to increase the spatial information about the surrounding sound field gathered by the array. In the rotation case, the improvement is maximal at $\alpha = 45°$. This is due to the spatial symmetry of the rotations by α and $90° - \alpha$ for this particular array geometry. Note that combining two time frames may be thought of as doubling the array to obtain 26 virtual microphones. This has the potential to increase the rank of $\mathbf{A}$ to 26. However, in the configuration considered here, the rank can be effectively increased to about 17 dimensions, as it is demonstrated by Fig. 2.a.

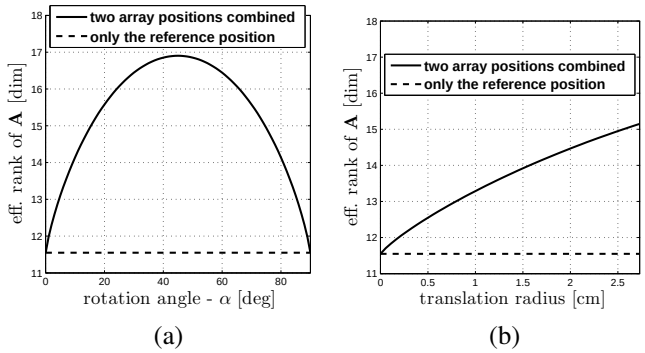

Fig. 2. Effective rank of the combined model matrix $\mathbf{A}(\omega)$ at 2 kHz (see (30)) as a function of (a) rotation angle α, (b) translation distance r.

The increase in the effective rank for a particular displacement may depend on frequency. Fig. 3 provides an example of the dependence of the effective rank of $\mathbf{A}(\omega)$ on frequency. Three different modes of motion are considered: (i) rotation by $\alpha = 3°$, (ii) translation by 5 mm, and (iii) a combination of both, with the rotation followed by the translation. It can be seen that in all three cases the effective rank increases with frequency. Recall that the rotation matrix $\mathbf{R}_i$ does not depend on frequency. Thus, the dependence on frequency of $\mathbf{A}(\omega)$ that results in the increase of its effective rank is due to the frequency dependence of $\mathbf{V}(\omega)$. Note that, for this particular set of parameters, the improvements in the effective rank due to the rotation or the translation alone are similar. The effect of the complex movement (rotation followed by translation) is greater than that of the rotation or the translation alone.

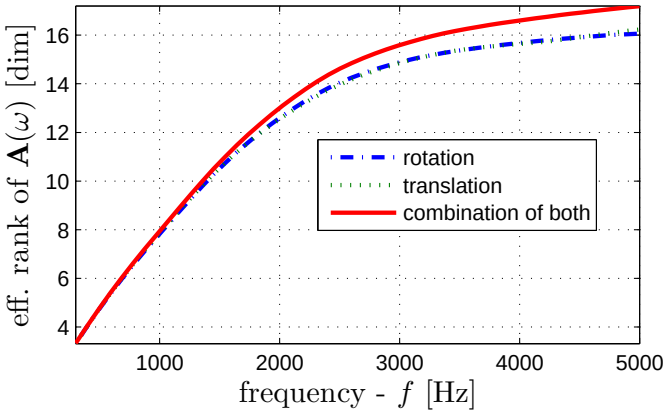

Fig. 3. Effective rank of $\mathbf{A}(\omega)$ as a function of frequency for three different modes of motion. The rotation and translation curves almost overlap.

To summarize, the effective rank of the combined system matrix $\mathbf{A}(\omega)$ increases for larger displacements and higher frequencies. Thus, the benefit of combining the different frames is expected to be more pronounced with an increase in the speed of the robot motion and the processing frequency range. This tendency is expected to be similar for different types of motion, including rotation, translation, and a combination of both. Hence, for conciseness, the investigation in the following sections considers a typical example of a common robot head motion - rotation about the z axis. This type of motion in humanoid robots represents, for example, left and right head rotations during a conversation with multiple speakers, steering

of the head in response to a sudden acoustic event, and scanning of the surroundings.

V. Simulation study: motion compensation

This section presents an investigation of the motion compensation method. This investigation demonstrates the DoA estimation error induced by array motion with realistic parameters and the ability of the motion compensation method to reduce this error. The investigation is based on simulated microphone outputs from a moving microphone array and on the SH-MUSIC algorithm. An investigation of the motion-based enhancement approach is presented in the next section. Note that in both sections V and VI, the effect of the robot's torso on the sound propagation is neglected at present. However, the effect may be significant and may depend on the posture of the robot. A study of the appropriate adjustments of the Head-Related Transfer Function (HRTF) [34] that would be required to take this effect into account is suggested for future work. Nevertheless, section VII demonstrates experimentally that DoA estimation at high frequency may be robust to this effect.

A. Simulated setup

The investigation in this section uses a simulated rigid spherical array [20]. The array radius is 6 cm. The array consists of 24 nearly uniformly [35] distributed microphones. This array enables the estimation of the PWD function up to 3 kHz [20]. As noted above, the investigation is based on a relatively common head motion mode - rotation about the z axis with constant angular velocity denoted by α_z. A single source in a free field was simulated; it was initially positioned in the direction $(\theta, \phi) = (\pi/2, 0)$ in the array coordinate system. The array, the source, and the rotation direction are illustrated schematically in Fig. 4. An example

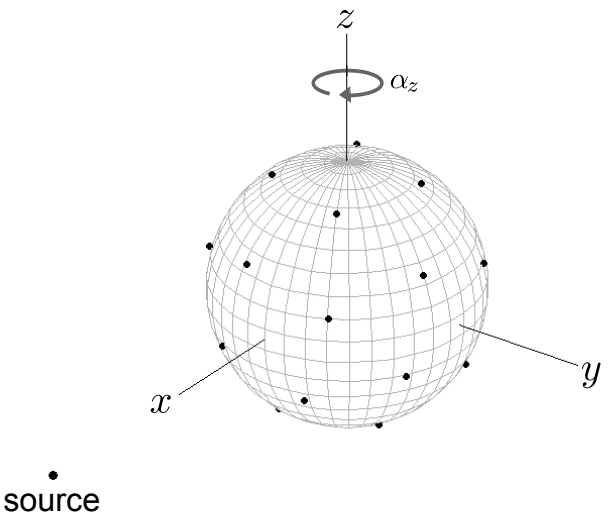

Fig. 4. A schematic illustration of the microphone array, the initial source position, and the rotation direction that were used for the investigation of DoA estimation performance using the motion compensation approach. Microphone positions are indicated by the black dots on the sphere surface.

involving the head geometry of a real humanoid is presented in section VI. In addition, the investigation in this section assumes that the source and the array are positioned in a free field in order to simplify the simulation and the interpretation of the results. A more realistic acoustic scenario may include room reverberation, which is expected to have a significant effect on DoA estimation performance. However, it should be emphasized that the motion compensation method presented here only aims to remove the effect of motion from the PWD estimates. Hence, the degree to which reverberation affects performance is expected to be determined solely by the robustness of the DoA estimation algorithm that uses these PWD estimates [36].

The outputs of all 24 microphones were simulated by filtering the source signal with the frequency responses of individual microphones. The responses were obtained analytically by using the expression for the array steering vector in the SH domain [20]. The filtering was carried out using the overlap-save technique. The rotation was simulated by changing the angle between the source and the array and updating the filter each millisecond. Considering the maximum angular velocity of 180 deg/s that was used in the simulations, updating the filter each millisecond corresponds to a spatial resolution of 0.18 deg. A speech signal from the TIMIT database [37] downsampled to 10 kHz was used as a source signal. Prior to the DoA estimation, white Gaussian noise with appropriate wideband Signal-to-Noise Ratio (SNR) levels was added to the microphone outputs in order to simulate the additive noise.

B. DoA estimation parameters and performance measures

The DoA estimation was performed using the SH-MUSIC algorithm described in section II-B. The STFT in *step 1* of the algorithm was calculated using Hamming windows of length 256 samples with 50% overlap. A block of 60 consecutive frames was used for the estimation of the covariance matrix in *step 3*. Hence, the overall data acquisition time required to produce a DoA estimate was about 0.75 s. Note that the array used here enables aliasing-free estimation of the PWD function up to the SH order of $N = 3$. The PWD function of this order was estimated in the frequency range of $1800 - 2700$ Hz. The lower frequency limit ensures that the simulated sound pressure contains significant energy from the required SH order.

The performance of the algorithm was assessed by observing the STandard Deviation (STD) and the average of the DoA estimation error angle Δ, which is defined as the angle between the true and the estimated DoA. The STD and the average of the error were calculated using 60 consecutive DoA estimation trials that differed by the noise component.

C. Results and discussion

Here, we examine the degree to which the DoA estimation performance is degraded when motion compensation is not applied to a moving array. For this purpose two different ways of estimating the PWD function are compared: (i) using the motion compensation method and (ii) using the stationary array model, in spite of the fact that the array is moving. The STD and the average error values obtained using the two

methods are presented in Figs. 5 and 6 as a function of the array angular velocity α_z. These results were obtained at an SNR of 10 dB.

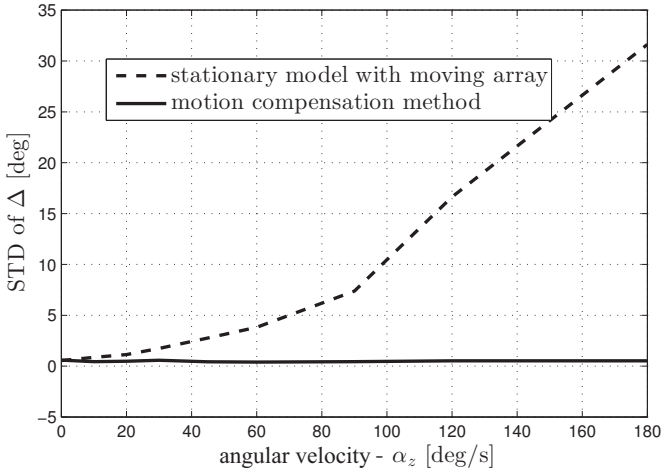

Fig. 5. STD of the DoA estimation error Δ as a function of angular velocity α_z.

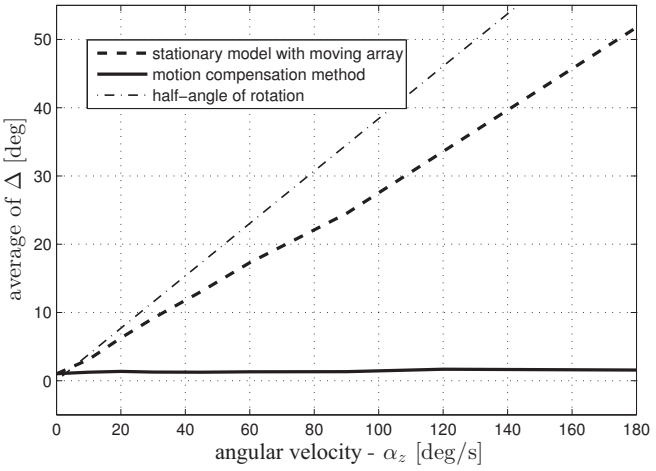

Fig. 6. Average of the DoA estimation error Δ as a function of angular velocity α_z. The curve denoted "half-angle of rotation" displays half of the angle that the array covers during the acquisition of the block of 60 STFT frames required to produce a single DoA estimate. See text for the details.

It can be seen that when no compensation for the motion is employed, both the DoA estimation variance and the bias increase with increasing angular velocity. This is due to the fact that during the data acquisition the source position relative to the array changes continuously, introducing an error into the estimated PWD function. It is interesting to note that, especially at lower angular velocities, the bias in the DoA estimation roughly follows the half-angle of rotation covered during the data acquisition, calculated as $\alpha_z \cdot 0.75/2$ [deg] (see Fig. 6). This is due to the averaging of the PWD function samples obtained at different array positions when estimating the covariance matrix in (8). Hence, at least for rotation, the bias can be partially corrected by a simple addition of half of the covered angle. Note that for $\alpha_z = 0$ deg, Figs. 5 and 6 display the performance of a stationary array. As expected, when motion compensation is employed this level of performance is maintained for arrays rotating at all α_z values in the range.

VI. Simulation study: motion-based enhancement

The previous section presented an investigation of the motion compensation approach using an array with a relatively large number of microphones. In practice, the arrays installed in humanoid robots may have a limited number of microphones. This motivated the development of the motion-based enhancement approach described in section IV-B, which proposes to combine the observations from different time frames (see (30)) for the estimation of the PWD function in *step 2* of the SH-MUSIC algorithm. The ability of this approach to increase the information gathered by the array was analyzed in section IV-C. The current section demonstrates that the approach is capable of improving actual DoA estimation performance under realistic conditions.

A. Simulated setup

The study in this section is based on the head of the existing humanoid robot NAO [30], with an average radius of about 6.25 cm. This radius is similar to the array used in the previous section. However, here, the number of microphones is limited to only $M = 4$, which is the case in the existing Phase-I and Phase-II models of this robot. The geometry of the head surface and the distribution of microphones used here are schematically illustrated in Fig. 7. As in the previous section,

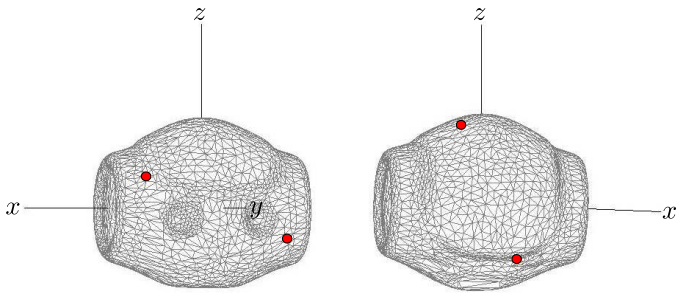

Fig. 7. Schematic illustration of the 4-microphone array used for the demonstration of the performance of the motion-based enhancement approach.

this study also focuses on array rotation about the z axis with a constant angular velocity α_z. Recall that the motion-based enhancement approach assumes a periodic signal, i.e., a signal with a constant amplitude at a given frequency, at least over the sound acquisition time. It was pointed out that, in practice, a periodic signal may be produced by an alarm or a music source. Hence, in this section, we use a pure tone source of frequency 3100 Hz, imitating a fire alarm system [38].

Microphone outputs were simulated using the overlap-save filtering technique, as before. The array steering vectors used in the filtering procedure were obtained by means of a Boundary Element Method (BEM) simulation that is based on the geometry of the head [22]. One source direction was simulated at a time. In total, 20 nearly-uniformly distributed source directions [35] were simulated. The DoA estimation

performance described below represents the average over all simulated directions. An additive noise component was simulated by adding white Gaussian noise with various narrowband SNR levels.

B. DoA estimation parameters

DoA estimation was performed using the SH-MUSIC algorithm. The algorithm was based on the STFT with the same parameters as in section V. A total of 180 time frames were used in order to produce a single DoA estimate, resulting in an acquisition time of 2.3 seconds. Estimation of the PWD function was based on the combined model in (30) with various numbers of the combined frames I. The narrowband covariance matrix $\mathbf{Q}(\omega)$ in *step 3* of the algorithm was estimated using $J = \lfloor 180/I \rfloor$ consecutive estimates of the PWD function. For example, for $I = 30$, the covariance matrix was estimated by averaging over 6 PWD estimates, i.e., $i = 1, 2, ..., 6$.

Recall that, as demonstrated in section IV-C, the effective rank of the combined steering matrix $\mathbf{A}(\omega)$ can be significantly lower than the actual rank of the matrix. Therefore, some of its singular values can be close to zero. Hence, for robustness purposes, the pseudo-inverse in (32) was calculated using the Singular Value Decomposition approach [24], while inverting only the singular values greater than $1/3$ of the largest singular value.

C. Results and discussion

Recall that the array simulated in this section is configured around a humanoid head with an average array radius of 6.25 cm. This array is used to process a sound field at the frequency of 3100 Hz. These parameters define the effective SH order of the sound field, which is $\lceil kr \rceil = 4$. At the same time, the array used here consists of only 4 microphones implying that without combining frames, as dictated by the motion-based enhancement approach, the array is capable of aliasing-free estimation of the PWD function of only the first order. Therefore, severe aliasing is expected to reduce the DoA estimation performance. Nevertheless, using the motion-based enhancement approach and increasing the number of combined frames is expected to improve the effective rank of $\mathbf{A}(\omega)$. This allows us to increase the estimated SH order of the PWD function in (32) to $N = 4$, thereby achieving the effective SH order of the field.

The DoA estimation performance as a function of the SNR and for different numbers of combined frames I is presented in Fig. 8. The reduced DoA estimation accuracy for $I = 1$ is, as expected, due to the detrimental effect of spatial aliasing. Nevertheless, increasing the number of combined frames, consistently increases the effective rank and the number of significant singular values of $\mathbf{A}(\omega)$, as summarized in Table I. Note that the maximum effective rank obtained in this simulation is 19.6 dimensions, which does not realize the full potential of the 25 columns of $\mathbf{A}(\omega)$. Nonetheless, the increase in the effective rank is significant enough to produce a substantial improvement in the DoA estimation accuracy, as demonstrated in Fig. 8 for $I = 15, 30, 45$, and 90. A similar improvement is obtained when increasing the angular velocity, α_z, while keeping the same number of combined frames, as demonstrated in Fig. 9. This is because increasing α_z for a given I increases the effective array aperture, which, in turn, increases the effective rank of $\mathbf{A}(\omega)$. Increasing the effective rank leads to improved DoA estimation performance, as before.

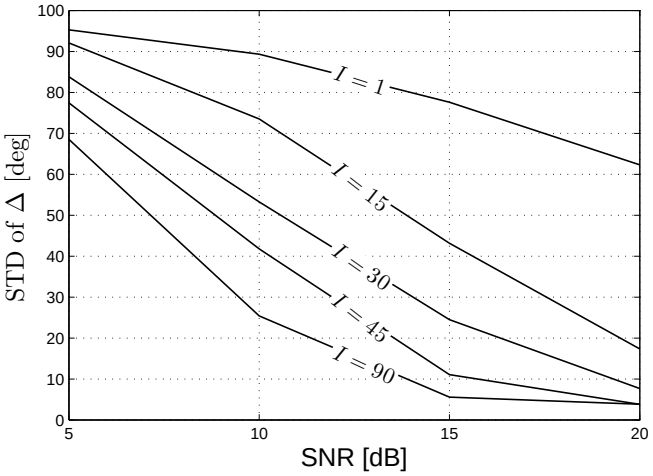

Fig. 8. Performance of the SH-MUSIC DoA estimation algorithm using the motion-based enhancement approach as a function of the number of combined frames I, with constant angular velocity $\alpha_z = 180$ deg/s.

TABLE I
THE EFFECTIVE RANK AND THE NUMBER OF SIGNIFICANT SINGULAR VALUES OF $\mathbf{A}(\omega)$ FOR VARIOUS NUMBERS OF COMBINED FRAMES I AT FREQUENCY 3100 HZ AND ANGULAR VELOCITY $\alpha_z = 180$ DEG/S.

I	1	15	30	45	90
eff. rank of $\mathbf{A}(\omega)$	4.0	7.8	10.8	13.4	19.6
norm. sing. val. $> 1/3$	4	5	8	10	17

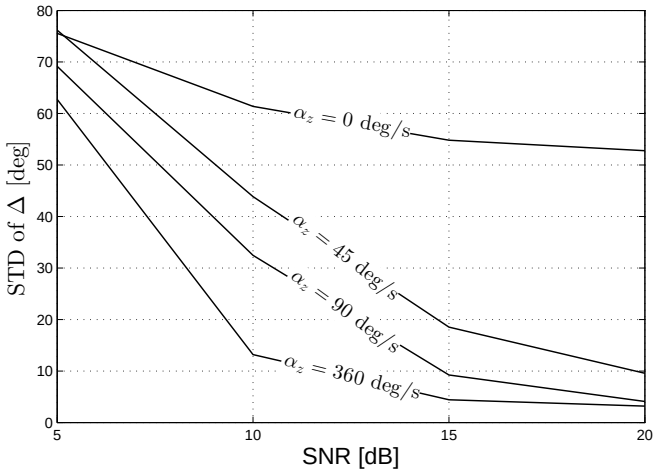

Fig. 9. Performance of the SH-MUSIC DoA estimation algorithm using the motion-based enhancement approach for different values of the array angular velocity α_z and a fixed number of combined frames, $I = 90$.

Recall that the motion-based enhancement method assumes that the amplitude of the source does not change from frame to frame during a sound acquisition time period. Hence, the above analysis was carried out using a pure tone source. In order to assess the sensitivity of the method to deviations from this constraint, two variations of the pure tone were considered: Amplitude Modulation (AM) and Frequency Modulation (FM). Note that both variations will affect the instantaneous amplitude of the source component in the STFT domain, which explicitly violates the above assumption.

The performance of the motion-based enhancement method for various levels of AM and FM modulations is plotted in Fig. 10. It can be seen that increasing the amount of variation

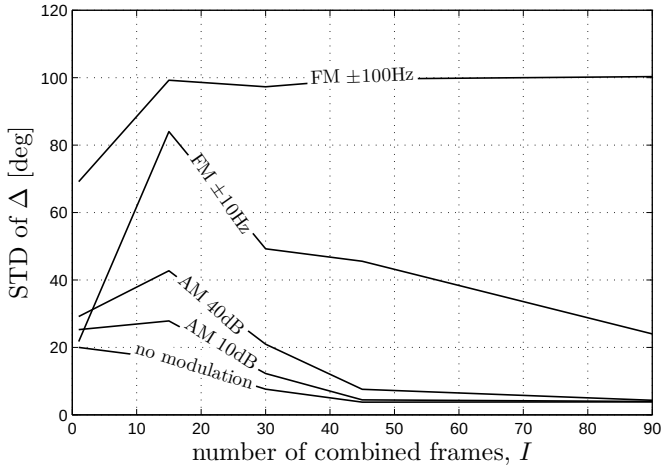

Fig. 10. Analysis of the effect of the source amplitude variation on the performance of the motion-based enhancement method. For the AM case, the ratio between the maximum and minimum amplitudes is specified in dB. For the FM case, the deviation from the tone frequency is specified in Hz. In both cases, the modulation sine wave had a frequency of 3 Hz, completing a cycle in about 13 time frames. The results were obtained under a SNR level of 20 dB.

either in amplitude or in frequency has a significant effect on the performance, mainly for low values of I. The method seems to be more robust to variations in the amplitude of the source than to the variations in frequency, especially for large I. The sensitivity to variations in frequency can be attributed to the fact that a sufficiently large change in the instantaneous frequency causes most of the signal energy to shift to the neighbouring frequency bins, thereby significantly reducing the SNR in the frequency bins that are used for the DoA estimation. A more detailed investigation of the robustness of the method to amplitude and frequency variations is proposed for future work.

Recall that the simulations in this section were carried out assuming sources in a free field. In practice, reverberation may reduce the performance of the motion-based enhancement method. In order to improve robustness to reverberation, the method can be combined with the Direct Path Dominance (DPD) test [36]. The investigation of this approach is out of the scope of the current paper and is left for future work.

VII. Experimental Study

In the current section, the processing methods proposed in this paper are applied to experimentally obtained data, based on the full-body humanoid robot NAO. Recall that a mismatch between the head-only based numerically-simulated steering vectors and the true array steering vectors is therefore expected in this case. Imprecise geometric modelling of the head and imprecise microphone positioning may also contribute to the expected mismatch in the steering vectors. The main purpose of the current section is, therefore, to demonstrate that the proposed methods are robust to these and other real-world related mismatches.

The experiment was performed in an anechoic chamber with dimensions of $2\times2\times2$ m. A loudspeaker (KRK systems, Rokit 6) was set to produce a tone at a frequency of 3100 Hz. The humanoid robot NAO was positioned in the chamber facing the loudspeaker. The head of the robot was set to rotate from left to right with constant angular velocity of 180 deg/s. An array of four microphones (AKG, C417PP) was positioned on the head of the robot as illustrated in Fig. 11. The positions of the microphones were chosen to be as close as possible to those used in the simulations in the previous section. The signals picked by the microphones were fed into a multichannel sound card (Focusrite, Scarlett 18i20) and sampled at a 10 kHz sampling rate. The methods used in sections V and VI with the same parameters were applied to the obtained microphone signals.

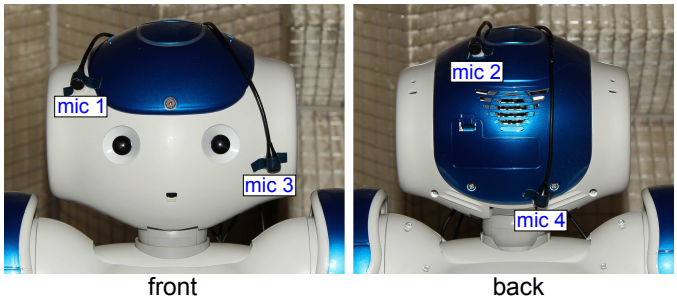

Fig. 11. Positioning of the microphones on the head of the humanoid robot NAO used in the experiment.

Figure 12 shows spectra of the SH-MUSIC algorithm applied to the estimated PWD function using three different methods: (a) no compensation for motion, (b) with motion compensation, and (c) motion-based enhancement with $I = 35$ frames. It can be seen that when no compensation for motion is applied (Fig. 12.a), the estimated DoA lags behind the true DoA in the direction of rotation by about 40 deg. Moreover, the spectrum in this case contains numerous additional peaks which may further reduce the performance of the estimation. When the motion compensation method is applied (Fig. 12.b), the estimated DoA is much closer to the true DoA. However, the spectrum still contains several additional peaks, due to spatial aliasing. The motion-based enhancement method (Fig. 12.c) overcomes the effect of aliasing and produces a much cleaner spatial spectrum, as compared to both previous cases. The spectrum contains a single peak in the vicinity of the true DoA, as expected in the case of a single source in an anechoic

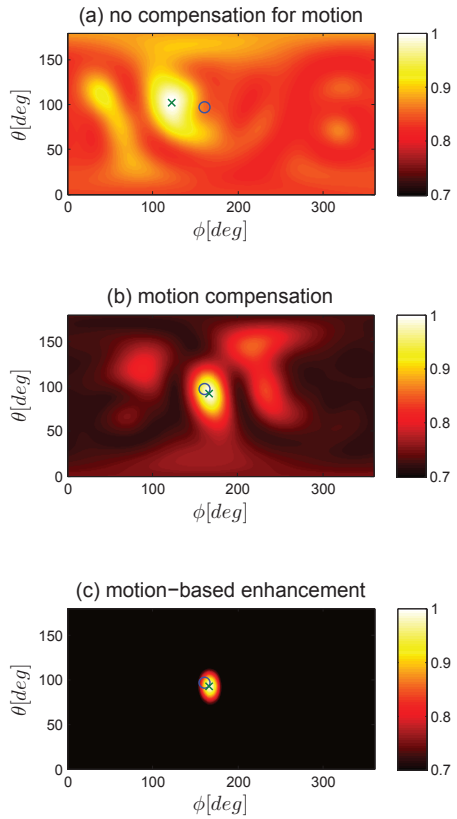

Fig. 12. Spatial spectra of the SH-MUSIC algorithm applied to the PWD function estimated from the experimental data using three different methods: (a) no compensation for motion, (b) with motion compensation, and (c) motion-based enhancement with $I = 35$ frames. The ground-truth DoA is indicated by $\circ$, the estimated DoA is indicated by $\times$.

chamber.

VIII. Conclusion

In this paper, an approach was developed for DoA estimation using microphone arrays installed on moving humanoid robots. A signal model was presented that can account for the motion of the robot using the SH domain. Based on this model, two different processing methods were proposed. The first is the *motion compensation* method, which was shown to reduce the motion-related error in the DoA estimation when applied to the data acquired by a moving robot. The second is the *motion-based enhancement* method. This method was shown to exploit the motion of the robot to improve the DoA estimation performance to a level beyond that of a stationary array. Future work may include the extension of the *motion-based enhancement* method to address non-periodic sources and reverberation.

Acknowledgment

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.

Appendix
Proof of the Unitarity of the Rotation Matrix

Although rotation matrices such as those that belong to $SO(3)$ are orthonormal, rotation in this paper is performed directly in the SH domain. Therefore, for completeness, it is shown in this appendix that the Wigned-D matrices used for such rotation are, indeed, unitary.

A general rotation Wigner-D matrix of order N is denoted here by $\mathbf{R}(\alpha,\beta,\gamma)$. Recall, from (12), that $\mathbf{R}$ is a block-diagonal matrix. The blocks on the diagonal are square matrices denoted by $\mathbf{D}_n(\alpha,\beta,\gamma) \in \mathbb{C}^{(2n+1)\times(2n+1)}$, $n = 0,1,...,N$. A general element in row m_1 and column m_2 of $\mathbf{D}_n(\alpha,\beta,\gamma)$ is given by

$$\begin{aligned}[\mathbf{D}_n(\alpha,\beta,\gamma)]_{m1}^{m2} &= D^n_{m_1,m_2}(\alpha,\beta,\gamma) \\ &= e^{-jm_1\alpha} d^n_{m_1,m_2}(\beta) e^{-jm_2\gamma}, \end{aligned} \tag{33}$$

where $d^n_{m_1,m_2}(\beta)$ is the Wigner-d function, which is real-valued and is given by [28]

$$\begin{aligned} d^n_{m_1,m_2}(\beta) = \eta_{m_1,m_2}\sqrt{\frac{s!(s+\mu+\nu)!}{(s+\mu)!(s+\nu)!}}\times \\ \sin^\mu(\beta/2)\cos^\nu(\beta/2)P_s^{\mu,\nu}(\cos\beta), \end{aligned} \tag{34}$$

where $\mu = |m_1 - m_2|$, $\nu = |m_1 + m_2|$, $s = n + (\mu+\nu)/2$, $P_s^{\mu,\nu}(\cdot)$ denotes the Jacobi polynomial, and coefficient η_{m_1,m_2} is given by

$$\eta_{m_1,m_2} = \begin{cases} 1, & m_2 \geq m_1 \\ (-1)^{m_2-m_1}, & m_2 < m_1 \end{cases}. \tag{35}$$

Note that the inverse, $\mathbf{R}^{-1}(\alpha,\beta,\gamma)$, represents a reciprocal rotation and is, indeed, given by a rotation in the reciprocal direction $(-\alpha,-\beta,-\gamma)$:

$$\mathbf{R}^{-1}(\alpha,\beta,\gamma) = \mathbf{R}(-\alpha,-\beta,-\gamma). \tag{36}$$

Hence, by substituting $(-\alpha,-\beta,-\gamma)$ into (33), we obtain

$$[\mathbf{D}_n(-\alpha,-\beta,-\gamma)]_{m1}^{m2} = e^{jm_1\alpha} d^n_{m_1,m_2}(-\beta) e^{jm_2\gamma}. \tag{37}$$

Using (36), the term $d^n_{m_1,m_2}(-\beta)$ can be expressed as

$$d^n_{m_1,m_2}(-\beta) = (-1)^{|m_1-m_2|} d^n_{m_1,m_2}(\beta). \tag{38}$$

Next, note that $(-1)^{|m_1-m_2|}\eta_{m_1,m_2} = \eta_{m_2,m_1}$. Now, observe that η_{m_1,m_2} is the only non symmetric term with respect to m_1 and m_2 in (34). This implies that

$$d^n_{m_1,m_2}(-\beta) = d^n_{m_2,m_1}(\beta), \tag{39}$$

which, by substituting into (37) and recalling that Wigner-d is real valued, yields

$$[\mathbf{D}_n(-\alpha,-\beta,-\gamma)]_{m1}^{m2} = \left([\mathbf{D}_n(\alpha,\beta,\gamma)]_{m2}^{m1}\right)^*. \tag{40}$$

The result in (40) implies that

$$\mathbf{D}_n(-\alpha,-\beta,-\gamma) = \mathbf{D}_n^H(\alpha,\beta,\gamma), \tag{41}$$

which completes the proof, implying that $\mathbf{R}^{-1}(\alpha,\beta,\gamma) = \mathbf{R}^H(\alpha,\beta,\gamma)$. ■

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] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 6, pp. 1082–1095, June 2014.

[5] A. Gershman, V. Turchin, and V. Zverev, "Experimental results of localization of moving underwater signal by adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2249–2257, Oct 1995.

[6] C. Zhang, D. Florencio, D. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, April 2008.

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

[8] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul 1989.

[9] G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura, and H. Nakajima, "Assessment of single-channel ego noise estimation methods," in *IEEE/RSJ Int. Conference Intelligent Robots and Systems (IROS)*, Sept 2011, pp. 106–111.

[10] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *17th National Conference on Artificial Intelligence (AAAI)*. AAAI, 2000, pp. 832–839.

[11] E. Chang, "Irregular array motion and extended integration for the suppression of spatial aliasing in passive sonar," *J. Acoust. Soc. Am.*, vol. 129, no. 2, pp. 765–773, 2011.

[12] A. Cigada, M. Lurati, F. Ripamonti, and M. Vanali, "Moving microphone arrays to reduce spatial aliasing in the beamforming technique: Theoretical background and numerical investigation," *J. Acoust. Soc. Am.*, vol. 124, no. 6, pp. 3648–3658, 2008.

[13] N. Yen and W. Carey, "Application of synthetic-aperture processing to towed-array data," *J. Acoust. Soc. Am.*, vol. 86, no. 2, pp. 754–765, 1989.

[14] J. Unnikrishnan and M. Vetterli, "Sampling and reconstruction of spatial fields using mobile sensors," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2328–2340, May 2013.

[15] 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.

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

[17] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[18] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[19] H. L. Van Trees, *Optimum Array Processing (Detection, Estimation and Modulation Theory, Part IV)*. New York: Wiley Interscience, 2002.

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

[21] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Adaptive blind source separation with HRTFs beamforming preprocessing," in *IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, June 2012, pp. 269–272.

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

[23] 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.

[24] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[25] M. A. Poletti, "Series expansions of rotating two and three dimensional sound fields," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3363–3374, 2010.

[26] W. B. Heard, *Rigid Body Mechanics: Mathematics, Physics and Applications*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany., 2006.

[27] P. J. Kostelec and D. N. Rockmore, "FFTs on the rotation group," *J. Fourier Anal. and Appl.*, vol. 14, no. 2, pp. 145–179, 2008.

[28] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskiĭ, *Quantum Theory of Angular Momentum*. World Scientific Publishing Company, Incorporated, 1988.

[29] T. Peleg and B. Rafaely, "Investigation of spherical loudspeaker arrays for local active control of sound," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. 1926–1935, 2011.

[30] G. Rump, "Embedded sound localization on a humanoid robot," in *Joint Workshop Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014.

[31] E. Tzoreff, B. Bobrovsky, and A. Weiss, "Single receiver emitter geolocation based on signal periodicity with oscillator instability," *IEEE Trans. Signal Process.*, vol. 62, no. 6, pp. 1377–1385, March 2014.

[32] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *European Signal Processing Conference (EUSIPCO)*, Sep. 2007, pp. 606–610.

[33] V. Tourbabin and B. Rafaely, "Objective measure for sound localization based on head-related transfer functions," in *IEEE 27th Convention Electrical Electronics Engineers Israel (IEEEI)*, Nov 2012, pp. 1–5.

[34] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064, 2002.

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

[36] 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.

[37] 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.

[38] D. Bruck and I. Thomas, "Waking effectiveness of alarms(auditory, visual and tactile) for adults who are hard of hearing," The fire protection research foundation, June 2007.

Vladimir Tourbabin (S'12) received the B.Sc. degree (summa cum laude) in materials science and engineering and the M.Sc. degree (cum laude) in electrical and computer engineering from Ben-Gurion University of the Negev, Israel, in 2005 and 2011, respectively. He is currently working towards the Ph.D. degree in electrical and computer engineering at Ben-Gurion University. His current research focuses on audition of humanoid robots.

Mr. Tourbabin is a recipient of the Negev Faran Fellowship.

Boaz Rafaely Boaz Rafaely (SM01) received the B.Sc. degree (cum laude) in electrical engineering from Ben-Gurion University, Beer-Sheva, Israel, in 1986; the M.Sc. degree in biomedical engineering from Tel-Aviv University, Israel, in 1994; and the Ph.D. degree from the Institute of Sound and Vibration Research (ISVR), Southampton University, U.K., in 1997. At the ISVR, he was appointed Lecturer in 1997 and Senior Lecturer in 2001, working on active control of sound and acoustic signal processing. In 2002, he spent six months as a Visiting Scientist at the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), Cambridge, investigating speech enhancement for hearing aids. He then joined the Department of Electrical and Computer Engineering at Ben-Gurion University as a Senior Lecturer in 2003, and appointed Associate Professor in 2010, and Professor in 2013.

He is currently heading the acoustics laboratory, investigating sound fields by microphone and loudspeaker arrays. During 2010-2014 he is served as an associate editor for IEEE Transactions on Audio, Speech and Language Processing, and since 2013 as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He is currently serving as an associate editor for IEEE Signal Processing Letter and for Acta Acustica united with Acustica, and as a chair of the Israeli Acoustical Association.

Prof. Rafaely was awarded the British Councils Clore Foundation Scholarship.