



HAL
open science

Reverberant Sound Localization with a Robot Head Based on Direct-Path Relative Transfer Function

Xiaofei Li, Laurent Girin, Fabien Badeig, Radu Horaud

► **To cite this version:**

Xiaofei Li, Laurent Girin, Fabien Badeig, Radu Horaud. Reverberant Sound Localization with a Robot Head Based on Direct-Path Relative Transfer Function. IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Oct 2016, Daejeon, South Korea. pp.2819-2826, 10.1109/IROS.2016.7759437 . hal-01349771

HAL Id: hal-01349771

<https://inria.hal.science/hal-01349771>

Submitted on 28 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reverberant Sound Localization with a Robot Head Based on Direct-Path Relative Transfer Function

Xiaofei Li[†], Laurent Girin^{†,‡,#}, Fabien Badeig[†] and Radu Horaud[†]

[†]INRIA Grenoble Rhone-Alpes, [‡]GIPSA-LAB, [#]Univ. Grenoble Alpes

Abstract— This paper addresses the problem of sound-source localization (SSL) with a robot head, which remains a challenge in real-world environments. In particular we are interested in locating speech sources, as they are of high interest for human-robot interaction. The microphone-pair response corresponding to the direct-path sound propagation is a function of the source direction. In practice, this response is contaminated by noise and reverberations. The direct-path relative transfer function (DP-RTF) is defined as the ratio between the direct-path acoustic transfer function (ATF) of the two microphones, and it is an important feature for SSL. We propose a method to estimate the DP-RTF from noisy and reverberant signals in the short-time Fourier transform (STFT) domain. First, the convolutive transfer function (CTF) approximation is adopted to accurately represent the impulse response of the microphone array, and the first coefficient of the CTF is mainly composed of the direct-path ATF. At each frequency, the frame-wise speech auto- and cross-power spectral density (PSD) are obtained by spectral subtraction. Then a set of linear equations is constructed by the speech auto- and cross-PSD of multiple frames, in which the DP-RTF is an unknown variable, and is estimated by solving the equations. Finally, the estimated DP-RTFs are concatenated across frequencies and used as a feature vector for SSL. Experiments with a robot, placed in various reverberant environments, show that the proposed method outperforms two state-of-the-art methods.

I. INTRODUCTION

Sound source localization (SSL) is a crucial methodology for robot audition. This paper addresses the problem of real-world SSL using a microphone array embedded into a robot head. The NAO robot (version 5) is used in this paper, whose head and its four embedded microphones are shown on Fig. 1.

Microphone-array processing SSL techniques are widely adopted for robot audition, e.g., [1], [2], [3], [4], [5], [6]. These techniques generally need a large number of microphones and high computational cost. The time difference of arrival (TDOA) techniques [7], [8] are suitable if fewer microphones are available, however they are generally applied to a free-field setup, in which the TDOA is frequency-independent. We address SSL in the more general case, namely when the source-to-sensor sound propagation is affected by the robot’s head and torso, e.g., binaural audition [9], [10], as well as by the room acoustics [11], and these effects are frequency-dependent [12].

As shown in Fig. 1, four microphones are embedded in NAO’s head. The two most discriminative microphone pairs

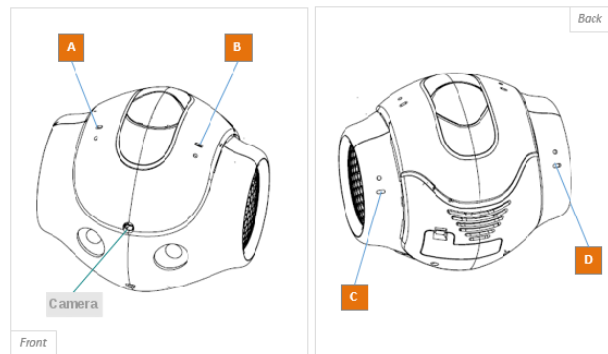


Fig. 1: The version 5 of the NAO head has four microphones, namely A, B, C, and D. This robot-head configuration is used in our experiments to illustrate the proposed SSL method.

in terms of SSL, i.e., the two cross microphone pairs (A-C and B-D) are used in this paper. The acoustic features are extracted separately from these two microphone pairs, and then these *pairwise features* are combined together. Two interaural cues, the interaural time (or phase) difference (ITD or IPD) and the interaural level difference (ILD), are widely used for SSL. When computed using the STFT, the ILD and IPD correspond to the magnitude and phase of a two-channel *relative transfer function* (RTF), which is the ratio between the ATFs of the two microphones [13]. The interaural cues, or equivalently the two-channel RTF, that correspond to the direct-path sound propagation are a function of the source direction, which is to be estimated from noisy and reverberant sensor signals, as they are available in a real environment.

Techniques have been proposed to identify the RTF in noisy environments, such as a speech non-stationary estimator [13], an RTF identification method based on speech presence probability and spectral subtraction [14], and an RTF estimator based on segmental PSD matrix subtraction [15]. In these RTF estimators, the multiplicative transfer function (MTF) approximation [16] is assumed. This approximation is justified only when the length of the room impulse response is shorter than the length of the STFT window, which is rarely the case in realistic acoustic setups. Moreover, the RTF estimated above is the ratio between two ATFs that include the reverberations and hence it is poorly suitable for SSL in echoic environments.

Techniques have been proposed to extract the interaural cues that correspond to the direct-path sound propagation,

e.g. based on the detection of time frames with less reverberations. The precedence effect [17] is widely modeled for SSL [18], [19], which relies on the principle that the onset frame is dominated by the direct-path wavefront [20], [21]. In the STFT domain, the coherence test [22] and the direct-path dominance test [23] are proposed to detect the frames dominated by one active source (namely only the direct-path propagation), from which reliable localization cues are estimated. However, in practice, there are always reflection components in the frames selected by these algorithms due to the inaccurate model or an improper decision threshold.

In this paper we propose a direct-path RTF estimator suitable for the localization of a single speech source in the real world. We build on the crossband filter proposed in [24] (actually a simplified CTF approximation proposed in [25]) for system identification. This filter accurately characterizes the impulse response in the STFT domain by a convolutive transfer function instead of the MTF approximation. The first coefficient of the CTF at different frequencies represents the STFT of the first segment of the channel impulse response, which is composed of the impulse response of the direct-path propagation and possibly a few reflections. Therefore, we refer to the first coefficient of the CTF as the direct-path ATF, and the ratio between the coefficients from the two channels as the direct-path RTF (DP-RTF). For the noise-free case, inspired by [26], based on the relation of the CTFs between the two channels, we construct a set of linear equations using the auto- and cross-power spectral density (PSD) of the speech signal received by the microphones.

At each frequency, the DP-RTF is the unknown variable of the linear equations, and can be estimated from these equations using the least square estimator. However, in practice, the sensor signals are always contaminated by noise. The speech PSD constructing the linear equations can be obtained by subtracting the noise PSD from the sensor signal PSD. Finally, the estimated DP-RTFs are concatenated over microphone pairs and frequencies, and mapped to the source direction space using the probabilistic piecewise affine mapping model [27]. Experiments, conducted in various real-world environments, show the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section II formulates the sensor signals based on the crossband filter. Section III presents the DP-RTF estimator. In Section IV, the SSL algorithm based on the probabilistic piecewise affine mapping model is described. Experimental results are presented in Section V, and Section VI draws some conclusions.

II. SIGNAL FORMULATION BASED ON CROSSBAND FILTER

In this work, we process microphone pairs separately. Thence, without loss of generality, only the sensor signals of one microphone pair are defined in this section, analyzed in section III, and the acoustic features of several microphone pairs will be combined for SSL in Section IV.

Let us consider a non-stationary source signal, e.g., a speech source $s(n)$ in the time domain. In a noise-free environment, the microphone-pair signals are

$$x(n) = a(n) * s(n), \quad y(n) = b(n) * s(n), \quad (1)$$

where $*$ denotes convolution, $a(n)$ and $b(n)$ are the room impulse responses from the source to the first and second microphone, respectively. Let T denote the length of $a(n)$ and $b(n)$. Applying the STFT, based on the MTF approximation, microphone signal $x(n)$ is approximated in the time-frequency (TF) domain as $x_{p,k} = s_{p,k} a_k$, where $x_{p,k}$ and $s_{p,k}$ are the STFT of the corresponding signals, p and k are the indexes of time frame and frequency bin, respectively. Let N denote the length of the STFT window (frame). This MTF approximation is only valid when the impulse response length T is lower than N . For a non-stationary acoustic signal, such as speech, a small length N (around 20 ms) is typically chosen to assume ‘local’ stationarity, i.e. in each frame. Therefore the MTF approximation is questionable in a, possibly strongly, reverberant environment with a long room impulse response.

To address this problem, the crossband filter was introduced in [24] to represent a linear system in the STFT domain more accurately. Let $\tilde{\omega}(n)$ and $\omega(n)$ denote the analysis and synthesis STFT windows respectively, and let L denote the frame step. The crossband filter model consists of representing the STFT coefficient $x_{p,k}$ as a summation of multiple convolutions across frequency bands. A CTF approximation is further introduced in [25] to simplify the analysis, i.e. using only band-to-band filters as

$$x_{p,k} = \sum_{p'=0}^{Q_k-1} s_{p-p',k} a_{p',k} = s_{p,k} * a_{p,k}, \quad (2)$$

where convolution is applied to the time variable p . The frequency dependent CTF length Q_k is related to the reverberation at the k th frequency band, which will be discussed in section V. The TF-domain impulse response $a_{p',k}$ is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k} = a(n) * \zeta_k(n)|_{n=p'L}, \quad (3)$$

which represents the convolution with respect to the time index n evaluated at frame steps, with

$$\zeta_k(n) = e^{j\frac{2\pi}{N}kn} \sum_m \tilde{\omega}(m)\omega(n+m). \quad (4)$$

In the next section, the CTF formalism is used to extract the impulse response of the direct-path propagation.

III. DIRECT-PATH RELATIVE TRANSFER FUNCTION

A. Definition of DP-ATF and DP-RTF Based on CTF

In the CTF approximation (2), using (3) and (4) at $p' = 0$, the first coefficient of $a_{p',k}$ can be derived as

$$a_{0,k} = a(n) * \zeta_{k,k}(n)|_{n=0} = \sum_{t=0}^{N-1} a(t)\nu(t)e^{-j\frac{2\pi}{N}kt}, \quad (5)$$

where

$$\nu(t) = \begin{cases} \sum_{m=0}^N \tilde{\omega}(m)\omega(m-t) & \text{if } 1-N \leq t \leq N-1, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, $a_{0,k}$ can be interpreted as the k -th Fourier coefficient of the impulse response segment $a(n)|_{n=0}^{N-1}$ (windowed by $\nu(t)|_{n=0}^{N-1}$). In the sense of transfer function identification, without loss of generality, we assume that the room impulse response $a(n)$ begins with the impulse response of the direct-path sound propagation. If the frame length N is properly chosen, $a(n)|_{n=0}^{N-1}$ is composed of the impulse responses of the direct-path propagation and a few reflections. Particularly, if the initial time delay gap (ITDG) is large compared to the frame length N , $a(n)|_{n=0}^{N-1}$ is mainly composed of the direct-path impulse response. Thence we refer to $a_{0,k}$ as the direct-path ATF.

Similarly, the CTF approximation of $y_{p,k}$ is written as

$$y_{p,k} = s_{p,k} * b_{p,k}, \quad (6)$$

and $b_{0,k}$ is assumed to represent the direct-path ATF from the source to the second microphone. By definition, DP-RTF is given by: $\frac{b_{0,k}}{a_{0,k}}$. Let us remind that this DP-RTF is a relevant cue for SSL.

B. DP-RTF Estimation

Since both channels are assumed to follow the CTF model, we can write:

$$x_{p,k} * b_{p,k} = s_{p,k} * a_{p,k} * b_{p,k} = y_{p,k} * a_{p,k}. \quad (7)$$

In [26], this relation is proposed in time domain for TDOA estimation. Eq.(7) can be written in vector form as

$$\mathbf{x}_{p,k}^\top \mathbf{b}_k = \mathbf{y}_{p,k}^\top \mathbf{a}_k \quad (8)$$

where \top denotes vector or matrix transpose, and

$$\begin{aligned} \mathbf{x}_{p,k} &= [x_{p,k}, x_{p-1,k}, \dots, x_{p-Q_k+1,k}]^\top, \\ \mathbf{y}_{p,k} &= [y_{p,k}, y_{p-1,k}, \dots, y_{p-Q_k+1,k}]^\top, \\ \mathbf{b}_k &= [b_{0,k}, b_{1,k}, \dots, b_{Q_k-1,k}]^\top, \\ \mathbf{a}_k &= [a_{0,k}, a_{1,k}, \dots, a_{Q_k-1,k}]^\top. \end{aligned} \quad (9)$$

Dividing both sides of (8) by $a_{0,k}$ and reorganizing the terms, we can write:

$$y_{p,k} = \mathbf{z}_{p,k}^\top \mathbf{g}_k, \quad (10)$$

where

$$\begin{aligned} \mathbf{z}_{p,k} &= [x_{p,k}, \dots, x_{p-Q_k+1,k}, y_{p-1,k}, \dots, y_{p-Q_k+1,k}]^\top \\ \mathbf{g}_k &= \left[\frac{b_{0,k}}{a_{0,k}}, \dots, \frac{b_{Q_k-1,k}}{a_{0,k}}, -\frac{a_{1,k}}{a_{0,k}}, \dots, -\frac{a_{Q_k-1,k}}{a_{0,k}} \right]^\top. \end{aligned} \quad (11)$$

We see that the DP-RTF appears as the first entry of \mathbf{g}_k . Hence, in the following, we base the estimation of the DP-RTF on the construction of $y_{p,k}$ and $\mathbf{z}_{p,k}$ statistics. More specifically, multiplying both sides of (10) by $y_{p,k}^*$ (* denotes complex conjugation) and taking the expectation (denoted by $E\{\cdot\}$), we obtain:

$$\hat{\phi}_{yy}(p, k) = \varphi_{zy}^\top(p, k) \mathbf{g}_k, \quad (12)$$

where $\hat{\phi}_{yy}(p, k) = E\{y_{p,k}y_{p,k}^*\}$ is the PSD of $y(p, k)$, and

$$\begin{aligned} \varphi_{zy}(p, k) &= [E\{x_{p,k}y_{p,k}^*\}, \dots, E\{x_{p-Q_k+1,k}y_{p,k}^*\}, \\ &E\{y_{p-1,k}y_{p,k}^*\}, \dots, E\{y_{p-Q_k+1,k}y_{p,k}^*\}]^\top \end{aligned} \quad (13)$$

is a vector which is composed of cross-PSD terms between the elements of $\mathbf{z}_{p,k}$ and $y_{p,k}$. In practice, these auto- and cross-PSD terms can be estimated by averaging the corresponding spectra over a number D of frames, i.e.:

$$\hat{\phi}_{yy}(p, k) = \frac{1}{D} \sum_{d=0}^{D-1} y_{p-d,k} y_{p-d,k}^* \quad (14)$$

The elements in $\varphi_{zy}(p, k)$ can be estimated by using the same principle. Consequently, (12) is approximated as

$$\hat{\phi}_{yy}(p, k) = \hat{\varphi}_{zy}^\top(p, k) \mathbf{g}_k. \quad (15)$$

In this equation, the speech PSD $\hat{\phi}_{yy}(p, k)$ and $\hat{\varphi}_{zy}^\top(p, k)$ can be obtained from the noise-free sensor signals. However in the real world, the PSD of speech signals are deteriorated by noise.

C. Speech PSD Estimate in the Presence of Noise

Noise signals are added into the sensor signals in (1) as

$$\begin{aligned} \tilde{x}(n) &= x(n) + u(n) = a(n) * s(n) + u(n), \\ \tilde{y}(n) &= y(n) + v(n) = b(n) * s(n) + v(n), \end{aligned} \quad (16)$$

where $u(n)$ and $v(n)$ are the noise signals in two sensors, respectively, which are supposed to be stationary and uncorrelated to the speech signal $s(n)$.

Applying the STFT to the sensor signals in (16): $\tilde{x}_{p,k} = x_{p,k} + u_{p,k}$ and $\tilde{y}_{p,k} = y_{p,k} + v_{p,k}$, respectively, in which each quantity is the STFT coefficient of its corresponding time domain signal. Similar to $\mathbf{z}_{p,k}$, we define

$$\begin{aligned} \tilde{\mathbf{z}}_{p,k} &= [\tilde{x}_{p,k}, \dots, \tilde{x}_{p-Q_k+1,k}, \tilde{y}_{p-1,k}, \dots, \tilde{y}_{p-Q_k+1,k}]^\top \\ &= \mathbf{z}_{p,k} + \mathbf{w}_{p,k} \end{aligned} \quad (17)$$

where

$$\mathbf{w}_{p,k} = [u_{p,k}, \dots, u_{p-Q_k+1,k}, v_{p-1,k}, \dots, v_{p-Q_k+1,k}]^\top. \quad (18)$$

We define the PSD of $\tilde{y}_{p,k}$ as $\hat{\phi}_{\tilde{y}\tilde{y}}(p, k)$. We also define the PSD vector $\varphi_{\tilde{z}\tilde{y}}(p, k)$, which is composed of the auto- or cross-PSD between the elements of $\tilde{\mathbf{z}}_{p,k}$ and $\tilde{y}_{p,k}$. Following the principle in (14), by averaging the auto or cross spectra of multiple frames, these PSDs can be estimated using the STFT coefficients of input signals as $\hat{\phi}_{\tilde{y}\tilde{y}}(p, k)$ and $\hat{\varphi}_{\tilde{z}\tilde{y}}(p, k)$. Because the speech and noise signals are not correlated, they can be represented as

$$\begin{aligned} \hat{\phi}_{\tilde{y}\tilde{y}}(p, k) &= \hat{\phi}_{yy}(p, k) + \hat{\phi}_{vv}(p, k) \\ \hat{\varphi}_{\tilde{z}\tilde{y}}(p, k) &= \hat{\varphi}_{zy}(p, k) + \hat{\varphi}_{wv}(p, k) \end{aligned} \quad (19)$$

where $\hat{\phi}_{vv}(p, k)$ is an estimation of the PSD of $v_{p,k}$, the PSD vector $\hat{\varphi}_{wv}(p, k)$ is composed of the estimated auto or cross PSD between the elements of $\mathbf{w}_{p,k}$ and $v_{p,k}$. The auto- and cross-PSD of noise can be subtracted by using the noise estimator [28] or the inter-frame spectral subtraction

technique [15]. In this work, for simplicity, we assume that noise is stationary (for example, the robot's ego-noise), and the noise-only signal is available, from which the noise PSD $\phi_{vv}(p, k)$ and $\varphi_{wv}(p, k)$ can be computed in advance. Consequently, we approximately compute the speech PSD as

$$\begin{aligned}\hat{\phi}_{yy}(p, k) &\approx \hat{\phi}_{\bar{y}\bar{y}}(p, k) - \phi_{vv}(p, k) \\ \hat{\varphi}_{zy}(p, k) &\approx \hat{\varphi}_{\bar{z}\bar{y}}(p, k) - \varphi_{wv}(p, k).\end{aligned}\quad (20)$$

Because of the temporal sparsity of the speech signal, parts of the frames are dominated by noise, which should be disregarded for DP-RTF estimation. Thence we define the frame index set that comprises the frames with considerable speech power as

$$\mathbf{p}_k = \{p \mid \hat{\phi}_{\bar{y}\bar{y}}(p, k) > \gamma\phi_{vv}(p, k)\}, \quad (21)$$

where γ is a power threshold. Let $P_k = |\mathbf{p}_k|$ denote the cardinal of \mathbf{p}_k .

D. Direct-Path Relative Transfer Function Estimation

Based on the speech PSD estimated in (20), by concatenating across frames, (15) could be written in matrix form

$$\hat{\mathbf{\Phi}}_{yy}(k) = \hat{\mathbf{\Psi}}_{zy}(k)\mathbf{g}_k. \quad (22)$$

where

$$\begin{aligned}\hat{\mathbf{\Phi}}_{yy}(k) &= [\dots, \hat{\phi}_{yy}(p, k), \dots]^\top, \quad p \in \mathbf{p}_k \\ \hat{\mathbf{\Psi}}_{zy}(k) &= [\dots, \hat{\varphi}_{zy}(p, k), \dots]^\top, \quad p \in \mathbf{p}_k\end{aligned}$$

are $P_k \times 1$ vector, $P_k \times (2Q_k - 1)$ matrix, respectively.

A least-square (LS) solution to (22) is given as

$$\hat{\mathbf{g}}_k = (\hat{\mathbf{\Psi}}_{zy}^H(k)\hat{\mathbf{\Psi}}_{zy}(k))^{-1}\hat{\mathbf{\Psi}}_{zy}(k)\hat{\mathbf{\Phi}}_{yy}(k) \quad (23)$$

where H denotes matrix conjugate transpose, $^{-1}$ denotes matrix inverse. The first element of $\hat{\mathbf{g}}_k$ is denoted as \hat{g}_k , which is an estimation of DP-RTF $\frac{b_{0,k}}{a_{0,k}}$.

IV. SOUND SOURCE LOCALIZATION METHOD

The amplitude and the phase of DP-RTF is equivalent to the IPD and ILD interaural cues corresponding to the direct-path propagation. As discussed in [29], [30], when the reference transfer function $a_{0,k}$ is much smaller than $b_{0,k}$, the amplitude ratio estimation is sensitive to the noise in the reference channel. Therefore, we normalize \hat{g}_k as

$$\hat{c}_k = \frac{\hat{g}_k}{\sqrt{|\hat{g}_k|^2 + 1}}. \quad (24)$$

It is clear that the phase is retained, and the amplitude is normalized as $0 < |\hat{c}_k| < 1$.

The quantity \hat{c}_k is the estimated DP-RTF for one microphone pair, where the index of microphone pair is omitted. Concatenating the estimated DP-RTF of microphone pairs A-C and B-D, yields $\hat{\mathbf{c}}_k = [\hat{c}_{k,AC}, \hat{c}_{k,BD}]^\top$.¹ Then, concatenating $\hat{\mathbf{c}}_k$ across frequencies, we obtain a global feature vector in \mathbb{C}^{2K} :

$$\hat{\mathbf{c}} = [\hat{\mathbf{c}}_0^\top, \dots, \hat{\mathbf{c}}_k^\top, \dots, \hat{\mathbf{c}}_{K-1}^\top]^\top, \quad (25)$$

¹For NAO version 5, a total of six microphone pairs are available. However, experiments show that it is sufficient to consider two microphone pairs.

where K denotes the number of frequencies involved in SSL.

To map the high-dimensional feature vector $\hat{\mathbf{c}}$ to a low-dimensional source direction $\mathbf{o} \in \mathbb{R}^O$ (O denote the dimension of source direction), we adopt the regression method proposed in [27]. Briefly, a probabilistic piecewise-linear regression $f : \mathbb{C}^{2K} \rightarrow \mathbb{R}^O$ is learned from a training dataset $\{\mathbf{c}_i, \mathbf{o}_i\}_{i=1}^I$, where \mathbf{c}_i is a feature vector and \mathbf{o}_i is the corresponding sound-source direction. Then, for a test DP-RTF feature vector $\hat{\mathbf{c}}$ extracted from the microphone signals, the source direction is predicted with $\hat{\mathbf{o}} = f(\hat{\mathbf{c}})$.

Due to the sparsity of speech signals in the STFT domain, it is possible that there are only a few significant speech frames at frequency k for one microphone pair, especially in the case of low SNR. In other words, P_k could be small, which makes the estimated \hat{c}_k unreliable. To disregard the unreliable \hat{c}_k in the regression procedure, we introduce a missing data indicator vector $\mathbf{h} \in \mathbb{R}^{2K}$. If the matrix $\hat{\mathbf{\Psi}}_{zy}(k)$ in (22) is underdetermined, i.e., $P_k < 2Q_k - 1$, its corresponding element in \mathbf{h} is set to 0, and 1 otherwise. The regression method that we use [27] makes use of such an indicator vector \mathbf{h} and the element in $\hat{\mathbf{c}}$ with a 0 indicator is disregarded. The revised prediction is $\hat{\mathbf{o}} = f(\hat{\mathbf{c}}, \mathbf{h})$.

V. EXPERIMENTS WITH THE NAO ROBOT

In this section several experiments using the NAO robot (version 5) are conducted in various real-world environments. From Fig. 1, one can see that four microphones are nearly coplanar, and that the angle between the microphone plane and the horizontal plane is small. The microphones are close to the head's fan (the circular ear in Fig. 1), thence the microphone recording include ego-noise due to the fan. As mentioned in [31], the fan noise is stationary and spatially correlated. In addition, its spectral energy mainly concentrates in a frequency range of up to 4 kHz, thence the recorded speech signal will be contaminated by the fan noise significantly.

A. The Datasets

The data are recorded in four real world environments: meeting room, laboratory, office, e.g., Fig. 2, and cafeteria, whose reverberation time T_{60} are approximately 1.04 s, 0.52 s, 0.47 s and 0.24 s, respectively.

Two **test datasets** are recorded in these environments:

- The *Audio-only* dataset: in the laboratory, the speech recording from the TIMIT dataset [32] are emitted by a loudspeaker. Two groups of data are recorded with a fixed robot-to-source distance of 1.1 m and 2.5 m, respectively. Besides T_{60} , ITDG and direct-to-reverberation ratio (DRR) are also important to measure the intensity of the reverberation. In general, the larger the robot-to-source distance the less ITDG and DRR. Obviously, the two cross microphone pairs allow a 360° azimuth localization. However, because of the limitation of NAO's head joint, NAO's head can not rotate in a 360° azimuth range. Thence, for each group, 174 sounds are emitted from directions uniformly distributed



Fig. 2: A typical *audio-only* localization experiment in the office environment. The robot turns its head towards the speaking person shown on the screen (please see the supplementary video).



Fig. 3: The *audio-visual* training dataset contains sound sources emitted by a loud-speaker that correspond to sound directions materialized by image locations (marked as blue circles).

in the range -120° to 120° (azimuth), and -15° to 25° (elevation).

- The *Audio-visual* dataset: Fig. 1 shows the NAO head camera, with a field-of-view of $60.97^\circ \times 47.64^\circ$; speech sounds are emitted by a loudspeaker lying in the camera's field of view. The image resolution is of 640×480 pixels, so 1° of azimuth/elevation corresponds to about 10.5 horizontal/vertical pixels. For this dataset, the source direction corresponds to a pixel in the image. The ground-truth source direction is obtained by localizing in the image the visual marker fixed on the loudspeaker. Four groups of data are recorded in four rooms, respectively. For each group, about 230 sounds are emitted from directions uniformly distributed in the camera field-of-view. As an example, Fig. 3 illustrates the 228 directions shown as blue dots in the image plane. The robot-to-source distance is approximately fixed as 2 m in this dataset.

In both of these two datasets, the external noise is much lower than the fan noise, thence noise in the recorded signal is almost composed of the fan noise. The signal to noise ratios (SNR) are approximately 14 dB, 11 dB for

Audio-only dataset with 1.1 m and 2.5 m robot-to-source distance, respectively, and 2 dB for *audio-visual* dataset ². As mentioned in Section III-C, the fan noise PSD $\phi_{vv}(p, k)$ and $\varphi_{wv}(p, k)$ are precomputed.

The **training dataset** $\{c_i, o_i\}_{i=1}^I$ for *Audio-only* experiments is generated by the anechoic head-related impulse responses (HRIR) of 1002 directions uniformly distributed in the same range as the test dataset. The training dataset for *Audio-visual* experiments is generated by the HRIR of 378 directions uniformly distributed in the camera field-of-view. The anechoic HRIR is obtained by truncating the room impulse response before the first reflection. White Gaussian noise (WGN) signals are emitted from each direction, and the cross-correlation between the microphone signal and source WGN signal gives the room impulse response of each direction.

B. Parameter Setup

The sampling rate of the microphone signals is 16 kHz. The window length of STFT is 16 ms (256 samples) with 8 ms overlap (128 samples). Only the frequency band from 300 Hz to 4 kHz is taken into account for speech source localization, i.e., the corresponding frequency bins are from 5 to 63, so the number of frequencies is $K=59$. The number of frames D for PSD estimation is set to 25 (0.2 s). The power threshold γ is set to 1.8. We set the length of CTF Q_k to be equal for all the frequency bins for simplicity, and denote it as Q , which is set to $0.25T_{60}$.

C. Method Comparison

The crucial point of binaural localization is to extract the reliable binaural cues from the noisy and reverberant sensor signals. Two state-of-the-art binaural feature estimation methods with good capability to reduce noise or reverberations are tested for comparison.

- A variation of the unbiased RTF estimator proposed in [14], in which the MTF approximation is adopted. The noise PSD is recursively estimated in the original work, while is more accurately precomputed using the noise-only signal in this work. We refer to this method as RTF-MTF.
- The coherence test (CT) method in [22]. The coherence test is used for searching the rank-1 time-frequency bins, which are supposed to be dominated by one active source. In this work, it is adopted for single speaker localization, in which one active source denotes the direct-path source signal. The TF bins that involve considerable reflections have low coherence. We first detect the maximum coherence over all the frames at each frequency bin, and then set the coherence test threshold for each frequency bin to 0.9 times its maximum coherence. In our experiments, this threshold achieves the best performance. The covariance matrix is estimated by taking a 120 ms (15 adjacent frames) averaging. The auto and cross PSD of all the frames that

²Note that the loudspeaker volume is different for two datasets.

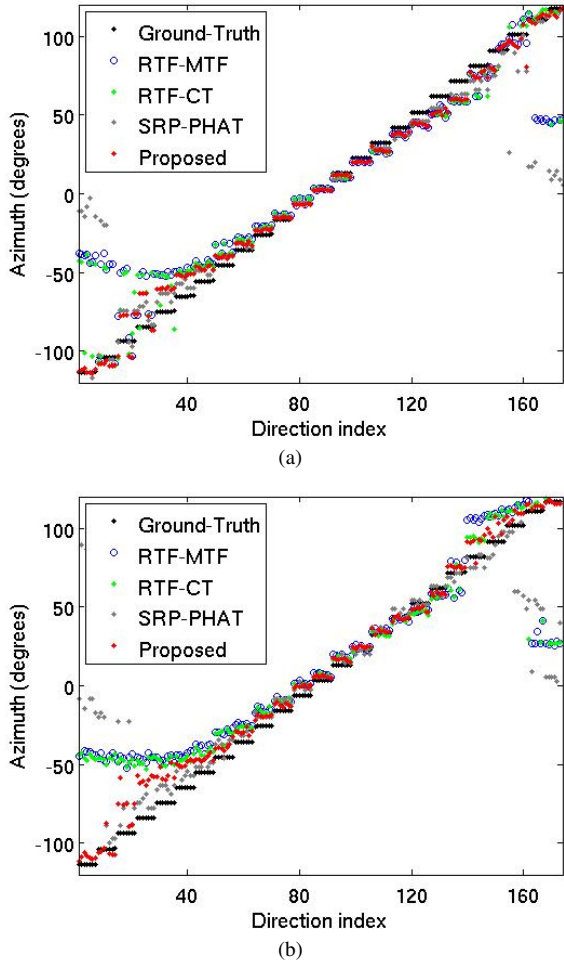


Fig. 4: Localization results for *Audio-only* dataset. (a) 1.1 m robot-to-source distance. (b) 2.5 m robot-to-source distance. The elevations of multiple source directions corresponding to each azimuth uniformly distribute from -15° to 25° .

have a coherence larger than the threshold are applied the spectral subtraction with the same principle in (20), and then are averaged over frames for acoustic feature extraction. We refer to this method as RTF-CT.

- In addition, a conventional beamforming SSL method: the steered-response power (SRP) utilizing the phase transform (PHAT) [33], [34] is also tested. The source directions in the training set of the proposed method are taken as the steering directions, and their HRIRs are taken as the steering vector.

D. Localization Results with the Audio-Only Dataset

Our experiments on *Audio-only* dataset show that, in the elevation range $[-15^\circ, 25^\circ]$, the elevation localization results are completely unreliable for all the three methods. This can be easily explained by the fact that the angle between the microphone plane and the horizontal plane is small, hence the microphone array has a low resolution for the elevation direction. Therefore, in Fig. 4, we only present the azimuth localization results.

From Fig. 4-(a), we observe that both the proposed method and the RTF-MTF and RTF-CT methods work well in the azimuth range $[-50^\circ, 50^\circ]$. The proposed method achieves slightly better results in this range. The performance drops drastically for the source directions out of this range. This indicates that the NAO’s microphone array has a better localization capability for the azimuth range $[-50^\circ, 50^\circ]$. From the results for the azimuth range $[-120^\circ, -50^\circ]$ and $[50^\circ, 120^\circ]$, it can be seen that RTF-MTF has the largest localization error and many localization outliers caused by the reverberations. By selecting frames that involve less reverberations, RTF-CT performs better than RTF-MTF, evidently, which can be observed from the fact that RTF-CT has less outliers than RTF-MTF. However, it is difficult to automatically set a coherence test threshold that could perfectly select the desired frames. Many frames that have a coherence larger than the threshold include reflections. Therefore, RTF-CT also has a relatively large localization error and some localization outliers. There are also many outliers for SRP-PHAT, which indicates that the steered response power is influenced by the reverberation. The proposed method achieves the best performance by properly extracting the direct-path RTF.

Fig. 4-(b) shows the localization results for the data with 2.5 m robot-to-source distance. Compared to the robot-to-source distance of 1.1 m, both ITDG and DRR are smaller. Consequently, the performance degrades for both the proposed method and the two state-of-the-art methods compared to Fig. 4-(a). The reasons for this degradation are the followings: for both RTF-MTF and RTF-CT the reflections are large relative to the direct-path impulse response, which makes the feature estimated from the reverberated signals more different than the feature corresponding to the direct-path propagation. In addition, concerning RTF-CT, the early reflection is closer to the direct-path impulse response, which makes less reverberation-free TF bins to be available. SRP-PHAT also has more outliers than the case in Fig. 4-(a) due to the lower DRR. For the proposed method, (i) the early reflections in the impulse response segment $a(n)|_{n=0}^N$ increase and (ii) in vector \mathbf{g}_k , the DP-RTF $\frac{b_{0,k}}{a_{0,k}}$ plays a more unimportant role relative to the other elements with the decreasing of DRR, which makes the DP-RTF estimation error larger. We can see that the proposed method still achieves the best performance, and most of its localization results are reliable.

E. Localization Results with the Audio-Visual Dataset

The source directions of *audio-visual* dataset distribute in the camera field-of-view, which is a small range in front of NAO’s head (azimuth range $[-30.5^\circ, 30.5^\circ]$). As shown in Fig. 4, good azimuth localization results are obtained in this range. Table I shows the localization error for both the azimuth (Azi.) and elevation (Ele.) directions. The localization error is computed by averaging all the absolute errors between the localized directions and their corresponding ground truth (in degrees).

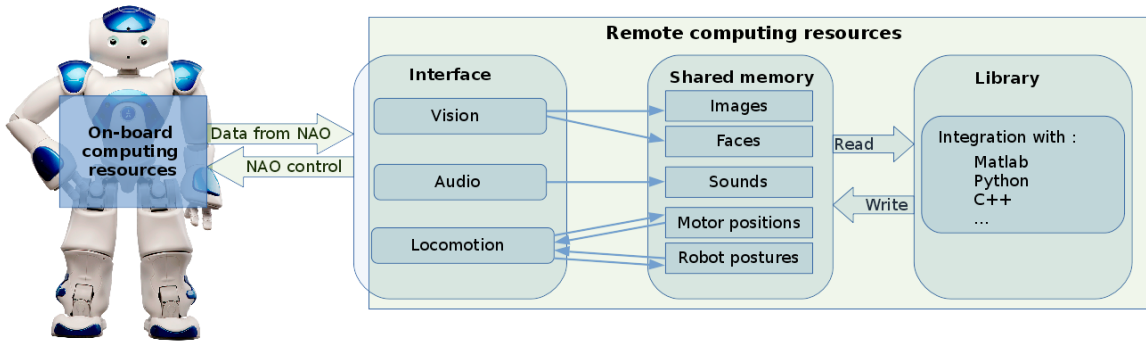


Fig. 5: Overview of the proposed distributed architecture that allows fast development of interactive applications using the humanoid robot NAO [35].

It can be seen that the elevation errors are always much bigger than the azimuth errors, due to the low elevation resolution of the microphone array that we already mentioned. In the cafeteria, the reverberation time T_{60} is 0.24 s, generally speaking, which is a low reverberation time. The RTF-MTF and RTF-CT methods yields performance comparable with the proposed method in the cafeteria environment. The reason is: the MTF approximation is relatively proper for this case, while the proposed method has a higher model complexity which needs more reliable data. In the office and laboratory, the reverberation times are larger, so the MTF approximation is not accurate anymore. As a result, Table I shows that the proposed method achieves evidently better performance than the two other methods in the office and laboratory environments. The performance of RTF-MTF is even better than RTF-CT, the reason is probably that the coherence test doesn't work well under low SNR conditions (the SNR is about 2 dB). In the meeting room, the reverberation time is high (1.04 s). SRP-PHAT achieves the worst performance due to the intense noise, especially the noise is spatially correlated. The proposed method still evidently performs better than the other methods. These further validates that the proposed method is more efficient in reverberant environments.

Methods	Cafeteria		Office		Laboratory		Meeting room	
	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.	Azi.	Ele.
RTF-MTF	0.45	1.57	0.62	2.14	1.44	2.31	1.87	3.66
RTF-CT	0.44	1.50	0.64	2.25	1.61	2.36	1.77	3.44
SRP-PHAT	0.77	1.95	1.03	2.80	1.41	3.33	2.04	3.52
Proposed	0.47	1.47	0.55	1.87	0.82	1.84	0.95	2.12

TABLE I: Localization error (degrees) for the *audio-visual* dataset. The best results are shown in **bold**.

F. Software Architecture

Ideally, one would like to implement the SSL method just presented using the embedded computing resources available with a robot such as the NAO companion humanoid. However, NAO like any other commercially available robot, has two limitations. Firstly, the on-board computing resources are restricted which implies that it is difficult to implement sophisticated audio signal processing and analysis algorithms needed by SSL in particular and by robot audition in general.

Secondly, robot programming implies the development of embedded software modules and libraries, which is a difficult task in its own right necessitating specialized knowledge.

We have developed a distributed software architecture that attempts to overcome these two limitations and which allows fast experimental validation of proof-of-concept demonstrators [35]. Broadly speaking, NAO's on-board computing resources are networked with external (or remote) computing resources. The latter is a computer platform (laptop or desktop) with its CPU's, GPU's, memory, operating system, libraries, software packages, internet access, etc. This configuration enables easy and fast development in Matlab, C, C++, Python, etc. Moreover, it allows the user to combine on-board libraries (motion control, face detection, etc.) with external toolboxes, such as Matlab's signal processing toolbox.

An overview of the proposed software architecture is shown on Fig. 5. Data coming from NAO (motor positions, images, microphone signals, or data produced by on-board computing modules) are fed into the external computer. Conversely, the latter can control the robot. Currently we developed three internal-to-remote interfaces: *vision*, *audio*, and *locomotion*. The role of these interfaces is twofold: (i) to feed the data into a memory space that is subsequently shared with existing software modules or with modules under development and (ii) to send back to the robot commands generated by the external software modules. Although these modules may be developed in a variety of programming languages, special emphasis was put to allow integration with the Matlab programming environment.

The proposed SSL method is implemented in Matlab, which offers the possibility to use Matlab's signal processing toolbox, e.g., the STFT. The Matlab computer vision toolbox is used for image processing. The on-board robot controller is invoked to rotate the robot head in the direction of the detected sound source.

VI. CONCLUSIONS

We have proposed a direct-path RTF estimator for SSL, and tested it on NAO robot. Instead of the MTF approximation, the method takes the CTF approximation, which is more precise when the impulse response is too long. Compared with the conventional RTF, the ratio between two direct-path

ATFs is more reliable for SSL. Because the training dataset is generated using the anechoic HRIR, the SSL module can operate for various room configurations, which is important for robot audition. Experiments have shown that the proposed method performs well for azimuth localization under difficult acoustic conditions, however poorly for elevation localization because of the microphone geometry of NAO robot head version 5. Thence, for the next version of NAO, a more reasonable microphone topology is expected.

REFERENCES

- [1] S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2009–2014, IEEE, 2007.
- [2] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 694–699, IEEE, 2012.
- [3] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 1033–1038, IEEE, 2004.
- [4] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.
- [5] Y. Sasaki, N. Hatao, K. Yoshii, and S. Kagami, "Nested iGMM recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3930–3936, IEEE, 2013.
- [6] R. Gomez, K. Nakamura, T. Mizumoto, and K. Nakadai, "Temporal smearing compensation in reverberant environment for speech-based human-robot interaction," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 3347–3353, IEEE, 2015.
- [7] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2033–2038, IEEE, 2009.
- [8] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1082–1095, 2014.
- [9] J. Hornstein, M. Lopes, J. S. Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the HRTF," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1170–1176, IEEE, 2006.
- [10] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 1, pp. 68–77, 2010.
- [11] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25, no. 1, 2015.
- [12] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.
- [13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Signal Processing, IEEE Transactions on*, vol. 49, no. 8, pp. 1614–1626, 2001.
- [14] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.
- [15] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *40th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.
- [16] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.
- [17] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.
- [18] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 981–985, IEEE, 2005.
- [19] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," *Journal of the Audio Engineering Society*, vol. 61, no. 7/8, pp. 508–520, 2013.
- [20] T. May, S. Van De Par, and A. Kohrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.
- [21] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.
- [22] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.
- [23] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.
- [24] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.
- [25] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.
- [26] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.
- [27] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 718–731, 2015.
- [28] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
- [29] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.
- [30] X. Li, R. Horaud, L. Girin, and S. Gannot, "Local relative transfer function for sound source localization," in *The European Signal Processing Conference*, 2015.
- [31] H. W. Loellmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Proceedings of Speech Communication*, pp. 1–4, VDE, 2014.
- [32] J. S. Garofolo *et al.*, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.
- [33] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180, Springer, 2001.
- [34] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pp. 1–121, IEEE, 2007.
- [35] F. Badeig, Q. Pelorson, S. Arias, V. Drouard, I. D. Gebru, X. Li, G. Evangelidis, and R. Horaud, "A distributed architecture for interacting with NAO," in *ACM International Conference on Multimodal Interaction*, (Seattle, WA), November 2015.