A Spherical Cross-Channel Algorithm for Binaural Sound Localization

Carlos Viña, Sylvain Argentieri, and Marc Rébillat

Abstract—This paper proposes a sound localization algorithm inspired by a cross-channel algorithm first studied by MacDonald et. al in 2008. The original algorithm assumes that the Head Related Transfer Functions (HRTFs) of the robotic head under study are precisely known, which is rarely the case in practice. Following the idea that any head is more or less spherical, the above assumption is relaxed by using HRTFs computed using a simple spherical head model with the same head radius as the robot head. In order to evaluate the proposed approach in realistic noisy conditions, an isotropic noise field is also computed and a precise definition of the Signal to Noise Ratio (SNR) in a binaural context is outlined. All these theoretical developments are finally assessed with simulated and experimental signals. Despite its simplicity, the proposed approach appears to be robust to noise and to provide reliable sound localization estimations in the frontal azimuthal plane.

Keywords — Robot audition, binaural cues, sound localization, sound processing.

I. INTRODUCTION

Robots, and more generally intelligent systems, are becoming today more and more reliable as partners in humans everyday life. Thanks to the recent advances in artificial perception, robots are now able to see [1], touch [2], and hear [3], making them able to interact naturally with humans. Among all these sensing abilities, audition is a key sense in humans, playing a crucial role in learning and communication. The same applies to robot audition, with the Robotics Community trying to reproduce the amazing human auditory system abilities to localize sounds-of-interest, extract them from a mixture of noise, and recognize them. This can be first achieved by considering microphone arrays embedded on a robotics platform [4]–[6]. But there is also an increasing demand for symbiotic interaction between humans and robots, thus bringing to the fore the design of humanoid platform endowed with bio-inspired perception. In this field, the binaural paradigm, based on a single pair of microphones placed on a robot head endowed with artificial pinnae, has receive recently more attention.

This work is rooted in this binaural paradigm, and is mainly concerned with the sound localization problem. Sound localization is generally the most important low-level auditory function, and most of the other higher-level auditory tasks (source extraction, source recognition, etc.) are generally highly dependent of it. Most of the existing approaches mainly rely on binaural inter-aural or monaural cues. These cues must be first extracted from the two sensed signals, and then linked to a corresponding source position. Such a connection is generally captured by an analytical model, modeling the sound propagation from the source to the two microphones on the surface of the head, or by experimental identifications of the so-called Head Related Transfer Functions (HRTFs). Many analytical models have been developed for a spherical head in that direction, such as the Auditory Epipolar Geometry (AEG) [7], the Revised AEG [8], or the Scattering Theory (ST) [9], [10]. But whatever the model, the approach is roughly the same: on the basis of the extracted binaural cues, one has to inverse the model to obtain an azimuth estimation. This inversion is not always possible due to the model’s complexity, and might be time consuming and not so robust to noisy conditions [10]. As an alternative, experimental HRTFs-based approaches can be mentioned [11]–[14]. But, since a closed-form HRTFs equation is quite impossible to obtain for a generic robotic head, a prior identification step in an anechoic room is mandatory for this approach. This strongly limits the applicability of the HRTFs approach to robotics.

The approach proposed in this paper is a generalization of an idea first depicted in [15] and extended in [16], [17]. Instead of working with the estimated binaural cues, the algorithm proposed here directly applies to the two binaural signals, using HRTFs derived from a generic spherical head model [18]. The method mainly relies on a product between the frequency content of the signals and the spherical HRTF, followed by the computation of correlation coefficients. Consequently, the proposed algorithm is computationally inexpensive.

The paper is organized as follows. The formalization of the approach is addressed Sec. II, with the focus being put on the careful simulation of realistic noisy conditions. Simulations results are discussed in Sec. III. Experimental results are then exhibited in Sec. IV before concluding the paper.

II. THE SPHERICAL CROSS-CHANNEL APPROACH

This section is mainly devoted to the theoretical concepts required by the proposed approach. First, the original cross-channel algorithm proposed in [15] is quickly recalled. Since it relies on the impracticable hypothesis that the exact HRTFs of the robotic head are known, we propose in Sec. II-B a spherical generalization of this cross-channel approach. The procedure used for the computation of the HRTFs for the spherical model is then given in Sec. II-C. In order to evaluate the proposed approach in realistic noisy conditions, an isotropic noise field associated with a precise definition of the Signal to Noise Ratio (SNR) in a binaural context are finally outlined in Sec. II-D.
A. The cross-channel sound localization algorithm

In all the following, a point position is described by its azimuth $\theta$, its elevation $\varphi$ and its distance $r$ from the head center according to the LISTEN coordinate system [19], as shown in Fig. 1. Let $H_i(r_s, \theta_s, \varphi_s, \omega)$ with $i = \{L, R\}$ denote the left and right HRTFs of the head for a source positioned at coordinates $(r, \theta, \varphi) = (r_s, \theta_s, \varphi_s)$. The very basic idea of the cross-channel localization algorithm [15] is to convolve the left and right perceived signals $m_L(t)$ and $m_R(t)$ with their opposing right and left Head Related Impulse Responses (HRIRs). Formally, one can then write in the frequency domain:

$$M_i(\omega) = H_i(r_s, \theta_s, \varphi_s, \omega)S(\omega), \quad i \in \{L, R\},$$

(1)

where $M_i(\omega)$ represents the Fourier Transform (FT) of the sensed signal $m_i(t)$, $i \in \{L, R\}$, and $S(\omega)$ the FT of the source signal emitted from a punctual sound source. One can then define the left and right cross-channel spectra $I_L(\cdot)$ and $I_R(\cdot)$ as:

$$I_L(r, \theta, \varphi, \omega) \overset{def}{=} M_L(\omega)H_L(r, \theta, \varphi, \omega),$$

(2)

$$I_R(r, \theta, \varphi, \omega) \overset{def}{=} M_R(\omega)H_L(r, \theta, \varphi, \omega).$$

Consequently, the cross-channel spectra are then given by

$$I_L(r, \theta, \varphi, \omega) = H_L(r_s, \theta_s, \varphi_s, \omega)S(\omega)H_L(r, \theta, \varphi, \omega);$$

(3)

$$I_R(r, \theta, \varphi, \omega) = H_R(r_s, \theta_s, \varphi_s, \omega)S(\omega)H_L(r, \theta, \varphi, \omega).$$

(4)

It’s obvious to notice that $I_L(.) = I_R(.)$ for $\theta = \theta_s$, $\varphi = \varphi_s$ and $r = r_s$. Of course, these two terms are no longer exactly equal in the presence of noise. But in that case the correlation between the cross-channel spectra is still expected to be maximal. This leads to the following straightforward source position estimation procedure [15]:

$$\{\hat{r}_s, \hat{\theta}_s, \hat{\varphi}_s\} = \arg \max_{r, \theta, \varphi} \{\text{cov}[(I_L(r, \theta, \varphi, \omega), I_R(r, \theta, \varphi, \omega))]\}$$

(5)

with $\hat{r}_s, \hat{\theta}_s$ and $\hat{\varphi}_s$ the estimated range, azimuth and elevation of the sound source, and

$$\text{cov}(X(\omega), Y(\omega)) = \frac{\text{cov}(X(\omega), Y(\omega))}{\sqrt{\text{var}(X(\omega))\text{var}(Y(\omega))}}$$

(6)

defined as the traditional Pearson’s correlation coefficient.

B. Spherical cross-channel sound localization algorithm

In the original version of the algorithm [15], cross-spectra are computed according to Eq. (4) using the same HRTFs as the considered head. This algorithm is thus based on the strong assumption that the HRTFs of the robotic head are precisely known. We propose to relax this hypothesis and to generalize this approach by assuming that each head is always roughly made of a spherical shape. Consequently, due to the assumed spherical symmetry of the problem, only the distance and azimuth are now of interest in the following.

Let’s now denote $H^*_i(\theta, \omega)$ the HRTFs of a spherical head, with $i \in \{L, R\}$. We then propose to compute the spherical cross-channel spectra as follows:

$$I_L(r, \theta, \omega) = M_L(\omega)H^*_R(r, \theta, \omega),$$

(7)

$$I_R(r, \theta, \omega) = M_R(\omega)H^*_L(r, \theta, \omega),$$

(8)

According to our assumption, the spherical HRTF pair $\{H^*_L(r, \theta, \omega), H^*_R(r, \theta, \omega)\}$ maximizing the correlation coefficient between the two cross-channel spectra $I_L(r, \theta, \omega)$ and $I_R(r, \theta, \omega)$ is expected to correspond to the source position $(r_s, \theta_s)$. The sound source position is thus estimated as before, thanks to:

$$\{\hat{r}_s, \hat{\theta}_s\} = \arg \max_{r, \theta} \{\text{cov}[(I_L(r, \theta, \omega), I_R(r, \theta, \omega))]\}. \quad (9)$$

This new formulation of the original Cross-Channel algorithm [15], denoted as Spherical Cross-Channel (SCC) algorithm in the following, is very appealing as it does not need the knowledge of the full HRTFs set of the used robotic head. The only required parameter is the spherical head model’s radius, which should be set to half the distance between the two ears endowing the robotics head. Additionally, the approach is computationally inexpensive, since it only requires a product in the frequency domain, followed by a straightforward correlation coefficient computation.

C. Spherical model HRTFs computation

As outlined in the previous subsection, the exact expression of the HRTF of a perfectly spherical head is needed by the proposed approach. This model will be now quickly recalled. Let $\beta$ be the so-called incidence angle, i.e. the angle between the ray from the center of the sphere to an arbitrary source position $(r, \theta)$, and the ray from the center of the sphere to a measurement point placed on the sphere. The transfer function $H^e(r, \beta, \omega)$ linking the sound pressure $P_s(r, \beta, \omega)$ received at the measurement point and the free-field pressure $P_f(\omega)$, i.e. the sound pressure existing at point 0 in the absence of the spherical head, for the angular frequency $\omega$, can be expressed as

$$H^e(r, \beta, \omega) = \frac{P_s(r, \beta, \omega)}{P_f(\omega)}.$$ 

(10)
In the case of a rigid perfectly spherical head, the expression of the diffracted sound pressure wave received at the measurement point allows to write [18]:

\[
H^s(r, \beta, \omega) = \frac{r e^{-j r \omega/c}}{a^2 \omega} \sum_{m=0}^{\infty} (2m+1) P_m [\cos(\beta)] b_m(r \omega/c) h_m'/(\omega a/c) \tag{11}
\]

with \(c\) the speed of sound in air and \(a\) the head radius.

In this expression, \(P_m(\cdot)\) and \(h_m(\cdot)\) stand for the Legendre polynomial of degree \(m\) and the \(m\)th-order spherical Hankel functions respectively. \(b_m'(\cdot)\) denotes the derivative of the function \(h_m(\cdot)\). Importantly, the spherical Hankel functions \(h_m(\cdot)\) can be expressed in terms of an intermediate auxiliary function easing its numerical evaluation, see [18]. Assuming that the two ears are placed on the surface of the sphere at \(r = a\) and \(\theta = \pm \frac{\pi}{2}\), as shown in Fig. 1, the incidence angle \(\beta\) can be replaced by \(\beta_L = -\pi/2 - \theta\) and \(\beta_R = \pi/2 - \theta\) in Eq. (11). The subsequent left and right HRTFs, denoted \(H^s_L(r, \theta, \omega)\) and \(H^s_R(r, \theta, \omega)\), respectively, for an arbitrary source located at \((r, \theta)\) are then given by

\[
H^s_L(r, \theta, \omega) = H^s \left( r, \frac{\pi}{2} - \theta, \omega \right), \tag{12}
\]

\[
H^s_R(r, \theta, \omega) = H^s \left( r, \frac{\pi}{2} - \theta, \omega \right). \tag{13}
\]

Note that to avoid the front-back confusion problem during localization, the source azimuth will be limited to lie between \(-\pi/2\) and \(\pi/2\), thus assuming that the source-sound is always located in front of the head.

D. Simulation of an isotropic noise field

Noise is probably one of the most common problems that localization algorithms have to face. Ideally, localization approaches should be efficient even for very low SNR conditions, thus allowing a robust sound source localization. When trying to assess this robustness, most authors propose to add two independent noises on the two binaural sensors, with the SNR being evaluated on one arbitrarily chosen reference signal, i.e. the left or right channel [20]. But when performing experiments in a realistic robotics environment, involving multiple noise sources in a reverberant space, the noise measured at the two binaural sensors is shown to be highly correlated, breaking down the independent noise assumption [21], [22]. Such correlations can be partially explained by the statistical nature of the noise field perceived by the binaural sensors. Indeed, an isotropic noise field, e.g. a wave field made of spatially uncorrelated plane-waves arriving with equal probability from any direction [21], is a better way to account for noise in simulations.

As shown in [21], an isotropic noise field can be in practice approximated by a finite discrete set of \(N\) punctual sound sources uniformly distributed over a sphere. The noise source angular positions \(\{\theta_n, \varphi_n\}_{n \in [1, N]}\) over this sphere can be computed according to

\[
\theta_n = \begin{cases} 
\text{mod} \left( \theta_{n-1} + \frac{3.6}{\sqrt{N(1-f(n)/N)^2}}, 2\pi \right), & \text{for } k = 2, \ldots, N-1, \text{ and } \theta_0 = 0, \theta_N = 0, \\
\varphi_n = \arccos (f(n)), \text{ with } n = 1, \ldots, N, \\
\text{with } f(n) = -1 + 2 \frac{n-1}{N-1}.
\end{cases}
\]

Then, the simulated noises \(N_L(\omega)\) and \(N_R(\omega)\) in the frequency domain, for the left and right channel respectively, come as:

\[
N_L(\omega) = \sum_{n=1}^{N} N_n(\omega) H_L(r_n, \theta_n, \varphi_n, \omega), \tag{15}
\]

\[
N_R(\omega) = \sum_{n=1}^{N} N_n(\omega) H_R(r_n, \theta_n, \varphi_n, \omega),
\]

where \(N_n(\omega)\) represents the \(n\)th source spectrum and \(r_n\) the noise source distance to the head. Depending on the used head in simulation, \(H_i(\cdot), i = \{L, R\}\), will be adjusted to the corresponding HRTF. For the spherical head, due to the symmetry of the problem, and assuming a farfield propagation, one has \(H_i(r_n, \theta_n, \varphi_n, \omega) = H^s_i(\theta_n, \omega)|_{r \rightarrow \infty}\). With other kinds of head, \(H_i(\cdot)\) is set to the values provided by the HRTFs database. Importantly, in all the following, the SNR will be defined as the ratio (expressed in dB) between the free-field signal power \(P_{\text{signal}}\) and the free-field noise power \(P_{\text{noise}}\) at the center of the head, i.e.

\[
\text{SNR}_{\text{dB}} = 10 \log_{10} \left( \frac{P_{\text{signal}}}{P_{\text{noise}}} \right), \tag{16}
\]

with \(P_{\text{noise}} = K \int_{-\infty}^{\infty} \left| \sum_{n=1}^{N} N_n(\omega) \right|^2 d\omega\).

Consequently, the SNR can be easily modified by simply scaling the global power of all the noise sources through the coefficient \(K\). Note that with such a definition, the effective SNR estimated for the two left and right channels will be different from \(\text{SNR}_{\text{dB}}\), and will be a function of the azimuth \(\theta\) of the sound source. This will be discussed in Sec. III-A.

III. RESULTS: SIMULATIONS

All the needed theoretical developments have now been introduced. In this section, the proposed localization algorithm will be assessed using simulations. To begin, all the parameters needed to perform the simulations are presented in Sec. III-A. Then, localization results for a spherical head for various SNR conditions are discussed in Sec. III-B. Finally, the proposed approach is evaluated with a more realistic head in Sec. III-C.

A. Generation of the synthetic binaural signals

From now, only the far-field case will be investigated. This literally means that the distance \(r\) to the source will be set to a value close to infinity in all the previous equations. Consequently, only the source azimuth estimation will be studied in the following.
**a) Simulation setup:** The speed of sound has been set to \( c = 330 \text{ m/s} \) and the head radius to \( r = 8.5 \text{ cm} \). In order to compute the spherical head HRTFs, the infinite sum in Equation (11) must be truncated. Since the higher-order Legendre and Hankel polynomials only capture decreasing variations in the angular and \( \omega \) frequency domains respectively, the sum can be easily truncated [18]. More precisely, a frequency-independent threshold – whose role is to bound the error between the infinite and truncated versions of the HRTFs – is chosen equal to \( 10^{-5} \) (see [18]), thus limiting in practice the number of terms in Equation (11) from about 10 to 40. Depending on the scenario, the sound source of interest is a white noise or a speech signal, sampled at \( f_s = 44.1 \text{ kHz} \).

All the results are obtained after applying the algorithm on 512-points windows. 30 windows are considered to perform a statistical analysis of the localization results in the white noise source case.

**b) Simulation of the isotropic noise field:** The simulation of the isotropic noise field has been performed with \( N = 200 \) punctual noise sources, each of them emitting an independent white Gaussian noise with the same variance. This variance is then modified in order to obtain the desired \( \text{SNR}_\text{dB} \) value, see Eq. (16). Because \( \text{SNR}_\text{dB} \) is defined in the free-field case, the effective SNR measured on the left and right signals, denoted \( \text{SNR}_L \) and \( \text{SNR}_R \) respectively, are dependent of the source azimuth \( \theta_s \), as illustrated in Figure 2. For instance, let’s consider a sound source at \( \theta_s = 45^\circ \), i.e. at the right of the spherical head. In this case, the sound source is directly in the sight of the right ear, while being hidden from the left ear. This head shadowing effect thus significantly degrades the left SNR value: one can see on Figure 2 that \( \text{SNR}_L \approx -9 \text{dB} \) while \( \text{SNR}_R \approx 2.5 \text{dB} \) for \( \text{SNR}_\text{dB} = 0 \text{dB} \). This phenomenon has to be kept in mind in the following, since the free-field \( \text{SNR}_\text{dB} \) exhibits quite optimistic values in comparison with the ones effectively encountered on the left and right channel. Finally, notice that \( \text{SNR}_\text{dB} \) is approximately obtained on the left and right channels when the farfield sound source is emitting from the front of the head (\( \theta_s = 0^\circ \)).

**B. SCC approach with a spherical head**

In this subsection, the proposed cross-channel approach depicted in Sec. II-B is assessed by simulations. This first study aims at demonstrating the fundamental limits of the method, since both the HRTFs and the used robotics head are assumed to be perfectly spherical. A non-spherical head will be used in the next subsection.

**a) Study of the correlation function:** As outlined in Sec. II-B, the proposed approach mainly relies on the computation of the correlation coefficient defined in Eq. (6) for several azimuth candidates. This coefficient should then exhibit its maximal value for \( \theta = \theta_s \). This is confirmed in Figure 3, where the correlation function is represented as a function of the azimuth \( \theta \) with 3 different SNR conditions, for a source emitting a white noise from the azimuth \( \theta_s = 45^\circ \). The correlation function exhibits a very sharp peak at the source azimuth for \( \text{SNR}_\text{dB} = 5 \text{dB} \), thus confirming the ability of the proposed approach to precisely localize a sound source. Logically, the maximal value of the correlation function decreases when the SNR conditions get worst. This is the case for \( \text{SNR}_\text{dB} = 0 \text{dB} \), but note that the angular position of the maximal correlation coefficient still remains at the azimuth \( \theta_s \), thus still allowing a good estimation of the source position. But for \( \text{SNR}_\text{dB} = -5 \text{dB} \), the peak related to the sound source position is no longer visible, or at best not placed at the good azimuth. But one must keep in mind that \( \text{SNR}_\text{dB} \) is not the effective SNR of the left and right signals (see Figure 2), and having \( \text{SNR}_\text{dB} < 0 \) corresponds some extreme cases (for instance, \( \text{SNR}_\text{dB} = 0 \text{dB} \) corresponds to a \(-12 \text{dB} \) SNR on the left or right channel, e.g. a signal approximately 16 times less powerful than noise).

**b) Localization performances:** Thanks to the correlation function depicted above, the localization efficiency of the approach can now be assessed on the whole azimuth range. For different \( \text{SNR}_\text{dB} \) values, the real source azimuth \( \theta_s \) can be compared with the estimated one \( \hat{\theta}_s \), and an estimation error can be deduced, as shown in Figure 4. The algorithm performs very well for good to poor SNR conditions. Indeed, for \( \text{SNR}_\text{dB} = 5 \text{dB} \) and above, the localization error is very close to \( 0^\circ \). Of course, using the spherical cross-channel approach together with a spherical
A. Experimental setup

In order to evaluate the proposed cross-channel algorithm with real binaural signals, experiments were conducted in an acoustically prepared room, equipped with 3D pyramidal pattern studio foams on the roof and the walls. Two different kinds of heads have been used: a 8.5cm-radius spherical head endowed with two antipodal surface-microphones, and a humanoid-like KEMAR head equipped with two microphones inside pinnas, see Figure 6. The microphone outputs have been synchronously acquired at $f_s = 44.1\text{kHz}$ from an Apogee acquisition card operating with a 24 bits resolution. A loudspeaker, placed at a constant distance $r_s = 1.5\text{m}$...
from the head, emits a white noise or a speech signal from the azimuth $\theta = \{-90, -60, -30, 0, 30, 60, 90\}^\circ$. For each position, a 1.2s-long binaural signal is recorded and then split in successive 512-points windows, resulting in about 100 localization estimations for each tested angular position.

### B. Experimental results

The localization results for the two different types of head are shown in Figure 7. The obtained localization accuracy is consistent with the simulations presented in the previous section. As expected, using the spherical head clearly leads to a better angular accuracy, with a 4° mean angular error on the 7 tested positions. As shown before, the localization gets worse in the lateral direction, with a mean error reaching 9°. The same applies for the KEMAR head, but with a higher angular error of about 7.5°. As anticipated, the angular precision is worse than in the full-spherical case, but remains definitely acceptable, especially for the azimuth angles between $-60^\circ$ and $60^\circ$ exhibiting an only 3.5° mean angular error. Again, localization precision is worse in the lateral directions when using a realistic head, as is the case with humans [25].

### V. Conclusion

An original cross-channel algorithm for binaural sound localization has been proposed in this paper. It relies on a spherical model to compute a correlation coefficient whose value is shown to be maximal at the actual source position. Importantly, the approach is not based on the knowledge of the robotic head’s HRTFs and is computationally inexpensive, thus allowing a real-time implementation on an arbitrary robotic platform. The mismatch between the spherical model and the used head is shown to decrease the angular estimation accuracy in comparison with the full-spherical case. Nevertheless, the localization error remains as small as 3.5° for sound sources emitting from a wide azimuth range. The approach is now being evaluated in simulation with other HRTFs databases, in order to test the generality of the approach w.r.t. various head shapes. Other ongoing works are more concerned with the source signal: preliminary results show good estimation results when working with speech signals. Finally, the robustness of the algorithm to early reflections and to reverberation will be evaluated in future works.

### ACKNOWLEDGMENT

This work was conducted within the French/Japan BINAHR (BINaural Active Audition for Humanoid Robots) project under Contract n°ANR-09-BLAN-0370-02 funded by the French National Research Agency.

### REFERENCES


