Optimal Positioning of a Binaural Sensor on a Humanoid Head for Sound SourceLocalization

Alain Skaf
CNRS ; LAAS ; 7 avenue du colonel Roche, F-31077 Toulouse Cedex 4, France
Université de Toulouse ; UPS, INSA, INP, ISAE ; UT1, UTM, LAAS ; F-31077 Toulouse Cedex 4, France
alain.skaf@laas.fr, patrick.danes@laas.fr

Patrick Danès

Abstract—A generic approach to the placement of a binaural sensor on a humanoid robot head is proposed in order to improve sound localization. After a brief description of binaural and spectral cues, the method is first illustrated on the well-known head spherical approximation, which response can be described analytically. Then, a simplified CAD model of a mannequin is considered. It is argued that recent developments on analytical expansions of HRTFs can help to get a solution even from a limited set of simulated head responses. The obtained conclusions noticeably comply with some antropomorphic statistics.

I. INTRODUCTION

Robot Audition is a research area of growing interest. The auditory modality, which is near-omnidirectional and insensitive to illumination conditions, constitutes an ideal complement to vision [1]. Besides array processing methods, which exploit the redundancy of the data sensed by a microphone array [2], binaural approaches consist in fitting two ears on a head [3], and mimic the human system to various extents. On this basis, bioinspired acoustic cues are generated and serve as a basis for localization.

Our aim is to design and implement a bioinspired sensor for a humanoid robot, which complies with the geometrical and computational foundations involved in humans. As a first step of our overall study, we propose a methodology to the optimal placement of “conchas” on a head. The paper is organized as follows. In Section II, the major cues involved in sound localization are recalled, namely binaural and monaural cues. Section III introduces the classical spherical head model. From the analytical expression of its Head Related Transfer Function (HRTF), the behavior of the consequent binaural localization cues is analyzed. Section IV introduces the proposed placement optimization method. First, an application to the spherical head model is presented. Prerequisite results from [20] on HRTF continuous expansions are then overviewed. They enable the extension to any head-shoulders CAD model.

II. BINAURAL SOUND LOCALIZATION

Humans can localize sound in space with an angular accuracy of 1° [4]. Indeed, the signals sensed by the two ears convey information on the spatial sound source position. Two types of acoustic cues can be distinguished.

A. Binaural Cues

The “Duplex Theory”, developed by Rayleigh [5] to explain sound localization in humans, approximates a human head by a sphere. It also introduces the two main interaural (or binaural) cues used, namely the Interaural Level Difference (ILD) and Interaural Time Difference (ITD). The ITD is defined as the time delay between the signals sensed at the two ears for an incident sound wave. The ear closest to the source is impinged first, and is termed “ipsilateral”. The ITD then comes from the difference in distance between the paths traveled by the wave to this ear and to the other “contralateral” ear. If the head had an exact spherical shape, the ITD for a sound source emitting from an azimuth φ with respect to boresight could be well described by the Woodworth-Schlosberg formula [6]

$$\text{ITD} = \frac{a}{c}(\varphi + \sin \varphi),$$  \hspace{1cm} (1)

with a the head radius and c the speed of sound (Fig. 1). This approximation is excellent in the horizontal plane and is widely used in binaural synthesis. Besides, the ILD describes the ratio of Sound Pressure Levels (SPLs) between the two ears. Logically, the ipsilateral ear SPL is greater because of the attenuation of the sound reaching the contralateral ear induced by the shadowing of the head [7]. The ILD is frequency dependent, the high-frequency sound waves being more attenuated by structures than waves of higher wavelength. However, at high frequencies such that the sound wavelength becomes smaller than the head semi-circumference, the propagation delay between the two ears exceeds one wave time period. This aliasing effect makes the ITD ambiguous. So, the ITD comes as the dominant localization cue at low frequencies, while the ILD is the decisive cue at high frequencies [7].

ITD and ILD take identical values when the source lies on a cone symmetrical around the interaural axis, called “cone of confusion” [7]. So, from binaural cues alone, one cannot determine the elevation of a source, nor discern if it comes from the front or the back. In particular, a source in the median plane leads to zero ILD and ITD. Yet, the cone of confusion as introduced in the Duplex Theory does not exactly reflect the situation in humans. Indeed, the human head is not perfectly spherical and the external ear—or “pinna”—plays an important role in the elevation localization and in the resolution of the front-back confusion [8].

Fig. 1. ITD for a spherical head.
B. Monaural Cues

Spectral cues, such as normal modes or spectral notches, can be separately extracted from each sensed signal. They are mainly caused by the pinna [9]. They are essential in elevation determination and for localization in the median plane [10].

The relationship between the signal emitted by a sound source at a certain location in space and the signal sensed by an ear is captured by the “Head Related Transfer Function” (HRTF). The HRTF accounts for the way the incident sound wave is modified by the listener’s morphology, mainly through the reflections and diffractions caused by the pinna, head and torso [7]. Free field propagation is assumed, so that neither room reflections nor reverberations are considered.

Blauert [4] defines the HRTF as the ratio of the complex signal at the ear canal to the complex signal that would be caused by the same sound source at the center of the head when no listener is present. It is a function of the source range, elevation, azimuth and frequency. Its equivalent in the range-elevation-azimuth-time domain is the Head Related Impulse Response (HRIR). The HRTF encompasses spectral (monaural) cues, and its data for the left and right ears enable the computation of both ITD and ILD. For instance, the ILD (monaural) cues, and its data for the left and right ears enable the computation of both ITD and ILD. For instance, the ILD (

In (3), $\alpha$ is the angle of incidence—i.e. the angle between the ray from the sphere center to the source and the ray to the measurement point on the sphere surface, with $\alpha = 0^\circ$ the normal incidence—, $P_m$ terms the Legendre polynomial of degree $m$, and $h_m, h'_m$ respectively stand for the $m^{th}$-order spherical Hankel function and its derivative. For an “average” radius $a = 8.75\, \text{cm}$, $\mu = 1$ corresponds to a frequency $f \approx 625\, \text{Hz}$.

For an infinitely distant source and at low frequencies, (3) can be approximated by $H(\infty, \alpha, \mu) = 1 - i \frac{12}{\mu} \cos \alpha$ [13], so that its magnitude is essentially unity. In the special case of normal incidence and at high frequencies, i.e. when the wavelength is small compared to the radius of the sphere, the solution is reduced to that of a plane wave impinging normally on a rigid plane surface. So, the pressure at this interface becomes twice the free-field pressure, i.e. $|H(\infty, 0, \infty)| = 2$ [13].

Fig. 2 plots the magnitude of the HRTF (3) as a function of $\alpha$ for various monochromatic farfield sources. As stated above, the magnitude is close to 1 for all $\alpha$ at low frequencies, while it tends up to 2 at high-frequencies for $\alpha = 0$. The response towards the back of the head decreases as a result of shadowing. However, the minimum response does not occur at $\alpha = 180^\circ$. Instead, the so-called “bright spot” appears, as the scattered waves get there in phase [13].

To study the behavior of binaural cues based on this same model, the ears positions must first be defined. According to [4], the ears are set back about 10° w.r.t. the head diameter in humans. Fig. 3 shows the variation of the ILD along range and frequency when they are located at either ±90° or ±100° (using the same convention as $\alpha$). For a distant source, the ILD is quite small at low frequencies, and thus cannot be a dominant cue. By contrast, when the source gets closer to the listener, the low-frequency ILD becomes significant. A large low-frequency ILD thus seems to be a major cue indicating the proximity of a source and of the listener [13]. Another conclusion is that non-antipodal ear locations breaks the symmetry, see Fig. 3-right. Fig. 3 also plots the ITD, computed as the difference between the phases of the HRTFs at the two aforementioned locations of the left and right ears. Note that in general the ITD remains roughly independent of distance even for close sources.

As mentioned above, less symmetric head models are better suited to human morphology, therefore spheroids or ellipsoids have been studied [14].

IV. Optimal Ear Placement for Sound Localization

A. Basics

Interaural cues (ILD & ITD) provide substantial information not available monaurally. Their sensitivity to source location strongly depends on the ears positions. Assimilating the human head to an ellipsoid [14], the ears are positioned symmetrically w.r.t. the median plane, but rather backwards and downwards [4][10]. According to anthropometric statistics [15], the pinna offset down is about 3.03 cm (standard deviation $\sigma \approx 0.66\, \text{cm}$) and its offset back is about 0.46 cm ($\sigma \approx 0.59\, \text{cm}$). This may explain, together with the important role of motion and of the pinna (resp. the torso) at high (resp. low) frequencies, why humans can localize sounds accurately and do not suffer from serious perceptual ambiguities such as the front-back confusion and elevation errors.
with \( \Psi_{\text{pos}} = \Psi(\text{ILD}_{\text{pos}}, \text{ITD}_{\text{pos}}) \) a function of ILD and ITD for the ear position \( \text{pos} \). Elementary criteria can be built with 
\[ \Psi_{\text{pos}}(r, \theta, \phi, f) = |\text{ILD}_{\text{db}, \text{pos}}(r, \theta, \phi, f)| \text{ or } |\text{ITD}_{\text{pos}}(r, \theta, \phi, f)|. \]
Practically, the continuous integration over a sphere (4) is approximated by a formula of the form [21]
\[ \tilde{J}_{\text{pos}}(f) = \sum_{i,j,k} A_i B_j C_k \Psi_{\text{pos}}(r_i, \theta_j, \phi_k, f), \] (5)
where \( r_i, \theta_j, \phi_k \) term discrete values of the range, elevation and azimuth, and \( A_i, B_j, C_k \) are adequate constant weights. The criterion (4) can be also be frequency-averaged, e.g.,
\[ J_{\text{pos}} = \int_{f_{\text{min}}}^{f_{\text{max}}} \tilde{J}_{\text{pos}}(f) \, df. \] (6)

B. Application to the Spherical Head Model

The above method has been applied to the spherical head model, the ITD and ILD being deduced from Section III. 121 candidate positions \( \text{pos} \) of the left ear have been considered, namely, 11 elevations (from 15° at the top of the head to 165° at the bottom) times 11 azimuths (from 15° at the front to 165° at the back). The right ear is placed by symmetry.

Fig. 4 plots a numerical approximation of (6) vs \( \text{pos} \), defining (4) either from \( \Psi_{\text{pos}} = |\text{ILD}_{\text{db}, \text{pos}}| \) or \( \Psi_{\text{pos}} = |\text{ITD}_{\text{pos}}| \). The mean value is computed over a spherical shell of 35 cm interior radius (nearfield source) and 3 m exterior radius (farfield source), and over the frequency range [100 Hz–20 kHz]. Here, the antipodal configuration predictably leads to the maximum performance for both definitions of \( \Psi_{\text{pos}} \).

Restricting (6) to narrowband frequency bins leads to interesting results. For \( \Psi_{\text{pos}} = |\text{ILD}_{\text{db}, \text{pos}}| \), (6) is close to 0 dB over low frequency bins for all \( \text{pos} \), because the HRTF magnitude is close to 1 whatever the source and sensor locations. Over higher frequencies, (6) increases towards the shape of Fig. 4, and shows a maximum for diametrically opposite ears. As for \( \Psi_{\text{pos}} = |\text{ITD}_{\text{pos}}| \), (6) keeps the same shape over all frequency bins. It is more pronounced over very low frequencies, but then becomes stable, in agreement with Fig. 3.

C. A Useful Result for General Heads: HRTF Continuous Expansion

As aforementioned, the HRTF of each individual reflects his/her morphology. In practice, experimental measurements are gathered by placing a microphone in his/her ear canal

A first attempt to the placement of a binaural sensor which provides good average localization performance can consist in computing the conchae locations which maximize the mean value of a function of the binaural cues over a sphere (source positions at a fixed range from the head) or a spherical shell (source positions at variable range from the head). The candidate positions of the conchae (microphones) on the head of the studied subject (mannequin, robot) should keep the symmetry w.r.t. the median plane as in humans and animals. The criterion to be maximized can be expressed as follows:
\[ J_{\text{pos}}(f) = \int_{r_{\text{min}}}^{r_{\text{max}}} \int_{0}^{2\pi} \int_{0}^{\pi} \Psi_{\text{pos}}(r, \theta, \phi, f) \, r^2 \sin\theta \, dr \, d\theta \, d\phi, \] (4)
and recording the pressure for a discrete set of sound source (range, elevation, azimuth, frequency) tuples. A continuous HRTF expansion along space and frequency variables is then obtained from these data [16][17][18][19]. Zhang et al. [20] have developed a general modal decomposition of HRTFs along (range, elevation, azimuth, frequency). Interestingly, it enables the interpolation of the HRTF from measurements gathered at a single source distance. It writes as

$$ H(r, \theta, \phi, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \sum_{l=1}^{L} A_{m,n,l} j_{n}(Z_{l}(n)) R_{n}(kr) Y_{m}^{n}(\theta, \phi), $$

(7)

$$ R_{n}(kr) = \sqrt{n+1} k e^{-ikr} h_{n}^{(1)}(kr), $$

(8)

with \( k (k_{\max}) \) the (maximum) wavenumber, and \( \theta, \phi \) the elevation and azimuth. Herebelow, \( \theta \) is defined from \( 0^\circ \) (top) to \( 180^\circ \) (bottom), and \( \phi \) rotates counterclockwise, with \( 0^\circ, 90^\circ, 180^\circ \) and \( 270^\circ \) the front, left, back and right directions, respectively. \( \{Y_{m}^{n}\} \) terms the family of spherical harmonics of degree \( n \) and order \( m \), which constitutes a basis of orthogonal functions. \( \{R_{n}\} \) are normalizing functions, with \( h_{n}^{(1)} \) the spherical Hankel function of the first kind. The so-called “modal coefficients” \( \{\sum_{l=1}^{L} A_{m,n,l} j_{n}(Z_{l}(n))\} \) entail the spherical Bessel functions of the first kind \( \{j_{n}\} \) and their \( l \)-th positive roots \( \{Z_{l}^{(n)}\} \). The required truncation order \( N \) until which all spatial modes need to be included, is given by

$$ N = \left\lceil \frac{ek_{\max}}{2} \right\rceil, $$

(9)

where the value of \( s \) relates to the typical size of scattering elements. For the spherical head model, the value of \( s \) is just the radius \( a \), while for mannequins or humans, the torso should be enclosed as well. However, the torso only contributes to the HRTF at low frequencies [10]. So, above \( 3 \) kHz, \( s \) can be selected so as to include only the head contribution [20].

The spatial dimensionality of an HRTF, i.e. the number of basis functions \( \{Y_{m}^{n}\} \) to be considered in its expansion, is

$$ \text{Dim}(H) = (N + 1)^{2}. $$

(10)

In light of (9), the spatial dimensionality increases with the considered upper frequency. This is because at low frequencies, HRTFs require fewer modes since the spatial variation of waves is slower. At higher frequencies, more modes are needed as the smaller wavelength indicates faster spatial changes [20]. So, (7) is fully described by the set of \( (N + 1)^{2}L \) coefficients \( A_{m,n,l} \). These can be computed from \( (N + 1)^{2}L \) HRTF discrete measurements. A practical low-complexity algorithm is developed in [20] to obtain them. The truncation order \( L \) depends on the number of HRTF frequency samples.

As outlined before, the advantage of the expansion (7) is that it can map a discrete data set of measured HRTFs to a set of coefficients \( \{A_{m,n,l}\} \) enabling the expansion of the HRTF for any (range, elevation, azimuth, frequency) tuple. To test the efficiency of this method, the theoretical HRTF for a spherical head (3) has been compared with its reconstruction from a finite values set, for four distinct source positions and a frequency bandwith extending up to 20 kHz. Fig. 5 shows a good matching, except at high frequencies.

\[\text{Fig. 5. Comparison between the analytical HRTFs on a rigid sphere and the reconstructed ones using the continuous model of Zhang et al. [20].}\]
D. Optimal Ear Placement for a Simple Mannequin Model

The results for the spherical head model were expected, but their only aim was to assess the meaningfulness of some criteria. Now, the same method is applied to a head model much closer to a real human body. For this purpose, a simple mannequin model was designed, that consists of an ellipsoidal head with a nose, connected to an ellipsoidal torso through a cylindrical neck. The head is placed more ahead compared to the torso with respect to the frontal plane as in humans. Fig. 6 shows the three-dimensional model created using the finite element analysis software COMSOL Multiphysics®. The dimensions of the head and torso reported on Fig. 7 are based on anthropometric data [15].

The HRTFs for this model cannot be determined in closed-form as in the case of the sphere. So, some HRTF samples were simulated. Unfortunately, we were limited in this simulation to a frequency range of 5 kHz because for frequencies above this limit, the required number of mesh elements (at least five elements per wavelength) becomes very large and requires much more computer resources than available.

The HRTFs simulations were performed considering a farfield source. 176 ear positions were considered on the dummy head which give us 88 pairs of ears disposed symmetrically with respect to the median plane as shown on Fig. 6. 184 simulations corresponding to different source directions were carried out. The elevations angles were taken from 0° to 170° and the azimuth angles from 0° to 180°, with the conventions of Section IV-C. For symmetry reasons, the HRTFs for the left ears corresponding to azimuth angles between 180° and 360° were deduced from those of the right ears, which eventually leads to 337 HRTFs corresponding to 337 different source directions. This number satisfies the dimensionality condition required by (10) to reconstruct the HRTF up to a frequency of 5 kHz.

From the simulated HRTFs and using the continuous expansion presented in Section IV-C, the HRTF (relative to a specific ear position on the mannequin head) can be expanded for any source position. Fig. 8 shows the magnitude and the phase of the simulated HRTF samples and the reconstructed expansion up to 5 kHz, relative to the ear positioned at an 90° elevation and 90° azimuth, considering four distinct wave incidences. The reconstructed HRTF reasonably matches with the original data, especially for the phase.

Furthermore, from this continuous HRTF expansion, the ILDs and ITDs corresponding to any source location can be deduced at any frequency \( f < 5 \text{ kHz} \) and for each of the 88 pairs of ears. We can then proceed to an averaging of these binaural cues along (4) for \( \Psi_{pos} = |\text{ILD}_{\text{dB}}|_{pos} \) or \( \Psi_{pos} = |\text{ITD}_{pos}| \), respectively, over a sphere that represents all the admissible source directions (elevations and azimuths). In this work, the integrals (6)–(4) have been limited to farfield sources. Fig. 9 shows the results. At first glance, one can see...
that for both cues, the zone of maximum average performance is located around the antipodal configuration, which corresponds to azimuth and elevation angles of 90° as was the case for the spherical head model. But if we observe more closely this area, we notice that when exploiting ILD through (4) for $\Psi_{pos} = |ILD_{DB_{pos}}|$, the maximum is shifted downwards and slightly backwards. This can be explained by the fact that for the ITD, the only factor that counts is the difference in the paths traveled by the wave to reach each of the two ears; and for any position of the source, this factor is maximal when the ears are at the antipodal positions. Contrarily, in the case of the ILD, things are more complicated because the factor that matters here is the magnitude of the wave, and this factor is influenced by several effects like the shadowing, reflection and scattering effects caused by the head and the torso. So the antipodal configuration does not correspond necessarily (as is the case for the ITD) to the optimal positions of the ears which allow the best average performance over all the admissible source incidences. According to the proposed method of optimization, the ear should be positioned behind and below the center of the head. This result agrees with aforementioned antropomorphic statistics.

V. Conclusion

In conclusion, the proposed method appears to be successful in defining an optimal placement for a pair of ears on a humanoid robot head in order to obtain the maximum average performance of binaural cues w.r.t. source locations. For a body having a spherical or an ellipsoidal head shape, the ITD does not seem to be the determining cue because it always leads to the antipodal configuration whatever the torso shape. Contrarily, the ILD seems to be more important as it takes into account several effects to which every parts of the body (head, shoulder, torso, pinna, etc) contribute.

This paper presented preliminary results. In the future, the method will be investigated in more depth. Several factors should be studied such as the effect of range and frequency on the optimal placement of the ears. Concerning simulation aspects, it would be useful to refine the frequency steps especially at low frequencies in order to obtain a more accurate reconstruction of the HRTF magnitude in an effort to reduce possible errors. Boundary Element Methods could also be investigated in order to limit the computational cost while allowing to include frequencies higher than 5 kHz. A comparison of simulations method could be evaluated against some experimental measurements. Last, the problem will be extended to chiral pinnae, whose orientation will constitute an additional decision variable.

Acknowledgments

This work was partially 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. The authors would like to thank Dr. Zhang, Profs. Abhayapala and Kennedy for fruitful scientific discussions.

References