Proceedings of the 2008 IEEE
International Conference on Robotics and Biomimetics
Bangkok, Thailand, February 21 - 26, 2009

Implementation and Calibration of a Bayesian Binaural System for 3D Localisation*

João Filipe Ferreira, Cátia Pinho, and Jorge Dias
ISR — Institute of Systems and Robotics
FCT-University of Coimbra
Coimbra, Portugal
{jfilipe, catiap, jorge}@isr.uc.pt

***Abstract*— In this text we present a Bayesian system of auditory localisation in distance, azimuth and elevation using binaural cues only; we focus mainly on implementation details and the calibration procedure, and present experimental results. This binaural system is also integrated in a spatial representation framework for multimodal perception of 3D structure and motion — the Bayesian Volumetric Map (BVM). This solution will enable the implementation of an active perception system with great potential in applications as diverse as social robots or even robotic navigation.**

***Index Terms*— Bioinspired Perception, Sound-Source Localisation, Binaural Cues, Bayesian Inference, Occupancy Grids.**

I. Introduction

Although vision might be the dominant sense in humans, we rely on hearing as our only panoramic, long-range sensory system. The ability not only to detect and identify a sound, but also to pinpoint swiftly and accurately the location of its source can bring substantial advantages. This applies equally to a predator stalking its prey in the wild [1] and to robotic applications such as [2], [3] and many others. Moreover, auditory stimulus localisation is also an important component driving attention and gaze shifts, especially when the target is not in sight.

In this text we present a Bayesian system of auditory localisation in distance, azimuth and elevation using binaural cues only. We briefly summarise its driving theoretical background (presented in greater detail, together with preliminary results, in [4]), describe implementation details and its calibration procedure. The binaural system is also integrated in a spatial representation framework for multimodal perception of 3D structure and motion, the Bayesian Volumetric Map (BVM) — for more details, please refer to [5].

To support our research work, an artificial multimodal perception system (IMPEP — Integrated Multimodal Perception Experimental Platform) has been constructed at the ISR/FCT-UC consisting of a stereovision, binaural and inertial measuring unit (IMU) setup mounted on a motorised head, with gaze control capabilities for image stabilisation and perceptual attention purposes — see Fig. 1. The stereovision system is implemented using the STH-MDCS2/-C Stereo Head from Videre Design (`http://www.videredesign.com`), the binaural setup using two AKG Acoustics C417 linear microphones (`http://www.akg.com/`) and a FA-66 Firewire Audio Capture interface from Edirol (`http://www.edirol.com/`), and the motorised head using the pan and tilt unit model PTU-46-17.5

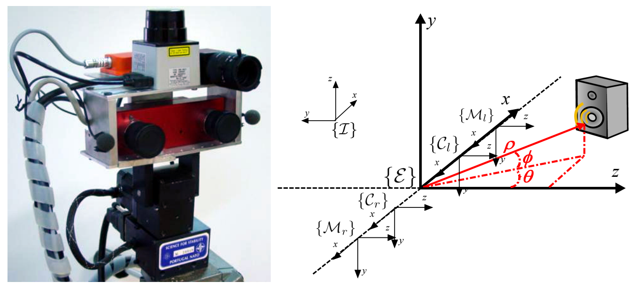

Fig. 1. View of the current version of the Integrated Multimodal Perception Experimental Platform (IMPEP), on the left. On the right, the IMPEP perceptual geometry is shown: $\{\mathcal{E}\}$ is the main reference frame for the IMPEP robotic head, representing the egocentric coordinate system; $\{\mathcal{C}_{l,r}\}$ are the stereovision (respectively left and right) camera referentials; $\{\mathcal{M}_{l,r}\}$ are the binaural system (respectively left and right) microphone referentials; and finally $\{\mathcal{I}\}$ is the inertial measuring unit's coordinate system.

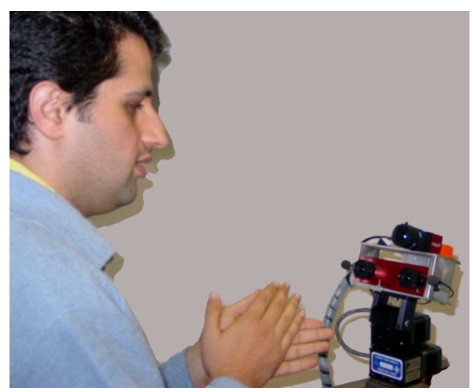

Fig. 2. Typical application context of the IMPEP active perception system.

*This publication has been supported by EC-contract number *FP6-IST-027140, Action line: Cognitive Systems*

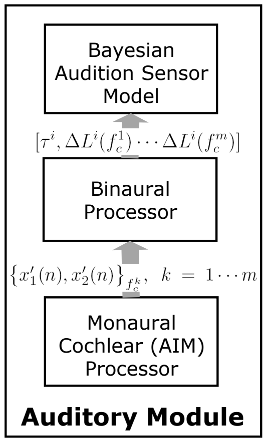

Fig. 3. The IMPEP Bayesian binaural system.

from Directed Perception (`http://www.dperception.com/`). The miniature inertial sensor, Xsens MTi (`http://www.xsens.com/`), provides digital output of 3D acceleration, 3D rate of turn (rate gyro) and 3D earth-magnetic field data for the Inertial Measurement Unit (IMU). This solution will enable the implementation of an active perception system with great potential in applications as diverse as social robots or even robotic navigation (Fig. 2).

II. Bayesian Binaural System Description

The Bayesian binaural system presented herewith is composed of three distinct and consecutive processors (Fig. 3): the *monaural cochlear unit*, which processes the pair of monaural signals $\{x_1, x_2\}$ coming from the binaural audio transducer system by simulating the human cochlea, so as to achieve a *tonotopic* representation (i.e. a frequency band decomposition) of the left and right audio streams; the *binaural unit*, which correlates these signals and consequently estimates the binaural cues and segments each sound-source; and, finally, the *Bayesian 3D sound-source localisation unit*, which applies a Bayesian sensor model so as to perform localisation of sound-sources in 3D space.

A. Cochlear and auditory periphery processing

The first stages of auditory processing consist of cochlear and auditory periphery processing, which produces what is called an *auditory image model* (AIM) [6]. The AIM processor implements a functional model of a cochlea that simulates the phase-locked activity that complex sounds produce in the auditory nerve.

Spectral analysis, the first stage of the AIM, is performed by a bank of auditory filters which converts each digitised wave that composes the stereo signal into an array of filtered waves. This processing is done using *gammatone filters* [7], [8], linearly distributed along a frequency scale measured in *equivalent rectangular bandwidths* (ERBs), as defined by [9] for simulating the cochlea, obtaining a model of *basilar membrane motion* (BMM) through frequency band decomposition.

The second stage of the AIM simulates the mechanical/neural transduction process performed by the inner haircells. It converts the BMM into a *neural activity pattern* (NAP), which is the AIM's representation of the afferent activity in the auditory nerve [6]. In this stage the envelopes of the signals are first compressed, and then subjected to halfwave rectification followed by a squaring and lowpass filtering, resulting in m stereo audio signal pairs corresponding to m frequency channels with respective frequency centre f_c^k, $\{x_1'(n), x_2'(n)\}_{f_c^k}$, $k = 1 \cdots m$.

B. Binaural cue processing

Sound waves arising from a source on our left will arrive at the left ear first. This small, but perceptible, difference in arrival time (known as an ITD, interaural time difference) is an important localisation cue and is detected by the *inferior colliculus* in primates, which acts as a temporal correlation detector array, after the auditory signals have been processed by the cochlea. Similarly, for intensity, the far ear lies in the head's "sound shadow", giving rise to interaural level differences (ILDs) [1], [10]. ITDs vary systematically with the angle of incidence of the sound wave relative to the interaural axis, and are virtually independent of frequency, representing the most important localisation cue for low frequency signals ($< 1500\,$Hz in humans). ILDs are more complex than ITDs in that they vary much more with sound frequency as a function of distance and elevation. Low-frequency sounds easily travel around the head and produce negligible ILDs. ILD values produced at higher frequencies are larger, and are increasingly influenced by the filter properties of each external ear, which imposes peaks and notches on the sound spectrum reaching the eardrum. Instead of being centred on the interaural axis, cones of confusion associated with particular ILD values take a different shape for each sound frequency.

Moreover, when considering sound sources within $1-2$ meters of the listener, binaural cues alone can even be used to fully localise the source in 3D space (i.e. azimuth, elevation and distance). Iso-ITD surfaces form hollow cones of confusion with a specific thickness extending from each ear in a symmetrical configuration relatively to the medial plane. On the contrary, iso-ILD surfaces, which are spherical surfaces, delimit hollow spherical volumes, symmetrically placed about the medial plane and centred on a point on the interaural axis [11]. Thus, for sources within 2 meters range, the intersection of the ILD and ITD volumes is a torus-shaped volume [11]. If the source is more than 2 meters away, the change in ILD with source position is too gradual to provide spatial information (at least for an acoustically transparent head), and the source can only be localised inside a volume within the cone of confusion delimited by the respective iso-ITD surfaces [11].

Given this background, we have decided to adapt the solution by Faller and Merimaa [12] to implement the binaural processor. Using this algorithm, interaural time difference and interaural level difference cues are only considered at time instants when only the direct sound of a specific source has nonnegligible energy in the critical band and, thus, when the evoked ITD and ILD represent the direction of that source (corresponding to the process involving the *superior olivary complex* (SOC) and the *central nucleus of the inferior colliculus* (ICc) in mammals). They show how to identify such time instants as a function of the *interaural coherence* (IC). The source localisation suggested by the selected ITD and ILD cues are shown to imply the results of a number of published psychophysical studies related to source localisation in the presence of distractors, as well as in precedence effect conditions [13]. This algorithm thus amplifies the signal-to-noise ratio and facilitates auditory scene analysis for multiple auditory object tracking, and is briefly summarised in the following paragraphs — for more details, please refer to [12].

The ITD and IC, denoted respectively by $\tau(n)$ and $c_{12}(n)$, where n indexes the sample currently being processed, are estimated from the normalised cross-correlation functions of the signals from left and right ear for each centre frequency f_c, respectively x_1' and x_2'. The normalisation of the cross-correlation function is introduced in order to get an estimate of the IC, defined as the maximum value of the instantaneous normalised cross-correlation function. This estimate describes the coherence of the left and right ear input signals. In principle, it has a range of $[0;1]$, where 1 occurs for perfectly coherent x_1' and x_2'. However, due to the DC offset of the halfwave rectified signals, the values of c_{12} are typically higher than 0 even for independent (nonzero) x_1' and x_2'. Thus, the effective range of the interaural coherence c_{12} is compressed to $[a;1]$ by the neural transduction. The compression is more pronounced (larger a) at high frequencies, where the low pass filtering of the half-wave rectified critical band signals yields signal envelopes with a higher DC offset than in the signal wave forms [12].

The ILD, denoted as $\Delta L(n)$, is then computed using the signal levels at the corresponding offsets [12]. Note that due to the envelope compression the resulting ILD estimates will be smaller than the level differences between the ear input signals. For coherent ear input signals with a constant level difference, the estimated ILD (in dB) will be 0.23 times that of the physical signals [12].

When several independent sources are concurrently active in free field, the resulting cue triplets $\{\Delta L(n), \tau(n), c_{12}(n)\}$ can be classified into two groups [12]: (1) Cues arising at time instants when only one of the sources has power in that critical band. These cues are similar to the free-field cues — localisation is represented in $\{\Delta L(n), \tau(n)\}$, and $c_{12}(n) \approx 1$. (2) Cues arising when multiple sources have non-negligible power in a critical band. In such a case, the pair $\{\Delta L(n), \tau(n)\}$ does not represent the direction of any single source, unless the superposition of the source signals at the ears of the listener incidentally produces similar cues. Furthermore, when the two sources are assumed to be independent, the cues are fluctuating and $c_{12}(n) < 1$. These considerations motivate the following method for selecting ITD and ILD cues. Given the set of all cue pairs, $\{\Delta L(n), \tau(n)\}$, only the subset of pairs is considered which occurs simultaneously with an IC larger than a certain threshold, $c_{12}(n) > c_0$. This subset is denoted

$$\{\Delta L(n), \tau(n) | c_{12}(n) > c_0\} \tag{1}$$

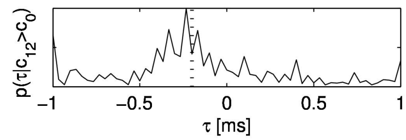

Fig. 4. Example of the use of an adaptation of the cue selection method proposed by [12] using a 1 s "multiple looks" buffer. Represented in the figure is a histogram of collected ITD cues corresponding to high IC levels for a particular frequency channel of a 1 s audio snippet. This histogram is interpreted as a distribution corresponding to the probability of the occurrence of ITD readings, which is then used as a conspicuity map in order to perform a *summary cross-correlogram* over all frequencies (see main text for more details).

The same cue selection method is applicable for deriving the direction of a source while suppressing the directions of one or more reflections. When the "first wave front" arrives at the ears of a listener, the evoked ITD and ILD cues are similar to the free-field cues of the source, and $c_{12}(n) \approx 1$. As soon as the first reflection from a different direction arrives, the superposition of the source signal and the reflection results in cues that do not resemble the free-field cues of either the source or the reflection. At the same time IC reduces to $c_{12}(n) < 1$, since the direct sound and the reflection superimpose as two signal pairs with different ITD and ILD. Thus, IC can be used as an indicator for whether ITD and ILD cues are similar to free-field cues of sources or not, while ignoring cues related to reflections.

Faller and Merimaa's cue selection method, as the authors point out, can be seen as a "multiple looks" approach for localisation, which provides the motivation for our implementation. Multiple looks have been previously proposed to explain monaural detection and discrimination performance with increasing signal duration [14]. The idea is that the auditory system has a short-term memory of "looks" at the signal, which can be accessed and processed selectively. In the context of localisation, the looks would consist of momentary ITD, ILD, and IC cues. With an overview of a set of recent cues, ITDs and ILDs corresponding to high IC values are adaptively selected and used to build a histogram that provides a statistical description of gathered cues (see Fig. 4).

Finally, the binaural processor capitalises on the multiple looks configuration and implements a simple auditory scene analysis algorithm for detection and extraction of important auditory features to build conspicuity maps and ultimately a saliency map, thus providing a functionality similar to the role of the *external nucleus of the inferior colliculus* (ICx) in the mammalian brain. The first stage of this algorithm deals with figure-ground (i.e. foreground-background) segregation and signal-to-noise ratio. In signal processing, the energy of a discrete-time signal $x(n)$ is given by [15]

$$E = \sum_{-\infty}^{\infty} |x(n)|^2$$

Using this notion, a simple strategy can be followed to selectively apply the multiple looks approach to a binaural audio signal buffer so that only relevant audio snippets are analysed. This strategy goes as follows: given a binaural signal buffer of N samples represented by the tuple $\{x'_1(n), x'_2(n)\}$, the average of the energies of the component signals $x'_1(n)$ and $x'_2(n)$ is

$$E_{avg} = \frac{\sum_1^N |x'_1(n)|^2 + \sum_1^N |x'_2(n)|^2}{2} \qquad (2)$$

and can be used as a noise gate so that only when $E_{avg} > E_0$ ITDs, ILDs and ICs triplets are collected for the buffer, yielding multiple looks values only for relevant signals (just the ITD-ILD pairs corresponding to high IC values are kept in conspicuity maps per frequency channel), while every other buffer instantiation is labelled as irrelevant noise. E_0 can be fixed to a reasonable empirical value or be adaptive, as seems to happen with human hearing.

Once the multiple looks information is gathered, since ITDs are proven to be stable across frequencies for a specific sound source at a given azimuth regardless of range or elevation, the ITD conspicuity maps may be summed over all frequencies, in a process similar to what is believed to occur in the ICx, in computational terms known as a *summary cross-correlogram* (again, see Fig. 4). From the resulting one-dimensional signal, the largest peaks may be taken as having been effected by the most important sound-sources represented in the auditory image. Then, a search is made across each frequency band to find the closest ITD and its ILD pair, for each reference ITD, thus building n-sized measurement vectors (for $m = n-1$ frequency channels) for each relevant sound source of the form

$$Z = [\tau, \Delta L(f_c^1) \cdots \Delta L(f_c^m)] \qquad (3)$$

TABLE I

PROBABILITY TABLE FOR $P(S_C|O_C\,C) \equiv P(S_C|O_C)$

	$[O_C = 0]$	$[O_C = 1]$
$[S_C = 0]$	1	.5
$[S_C = 1]$	0	.5
$\sum P(s_c\|O_C)$	1	1

C. *Bayesian sensor model*

Finally, regarding the Bayesian 3D sound-source localisation unit, auditory sensor space is defined as a log-spherical volumetric occupancy grid $\mathcal{Y}$, with each cell being indexed by its far corner $C \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{C} \subset \mathcal{Y}$ — this configuration follows the same formalism as the Bayesian Volumetric Map (BVM) framework, described in [5], and has the advantage of providing a natural setting for the integration of auditory cues, since the latter are directly a function of egocentric spherical coordinates. Moreover, logarithmic partitioning of distance accounts for the increasing just-noticeable differences (JND) of auditory distance cues corresponding to sound-sources at increasing distances, thus promoting an efficient use of memory resources.

The binaural sensor model Bayesian Program (BP), a formalism first defined by Lebeltel [16] that describes all relevant variables and distributions and the decomposition of the corresponding joint distribution, according to Bayes' rule and dependency assumptions[1], is summarised by the Bayes network presented on Fig. 5 — for an in-depth description of the model, please refer to [4]. The use of the auxiliary binary random variable S_C, which signals the presence or absence of a sound-source in cell C, and the corresponding family of probability distributions $P(S_C|O_C\,C) \equiv P(S_C|O_C)$, promotes the assignment of probabilities of occupancy close to 1 for cells for which the binaural cue readings seem to indicate a presence of a sound-source and close to .5 otherwise (i.e. the absence of a detected sound-source in a cell doesn't mean that the cell is empty). This family of distributions is defined in Table I.

Three separate, essential problems (in Lebeltel's formalism referred to as "questions") can be solved through Bayesian inference: (1) $P(o_c|z\,c)$; (2) $\max, \arg\max_C P([S_C = 1]|z\,C)$; and (3) $P(z|o_c\,c)$. The first question corresponds to the classical occupancy grid formulation, which will be used in the results section; the second question corresponds to the estimation of the position of cells most probably occupied by sound sources, yielding a gaze direction of interest in terms of auditory features for the multimodal attention system; finally, the third question represents the direct binaural sensor model, where the influence of S_C has been removed through

[1]Using this formalism, random variables are denoted through upper-case (e.g. S_C), and specific instantiations either stated explicitly using square brackets (e.g. $[S_C = 1]$) or implicitly through lower-case (e.g. c instead of C).

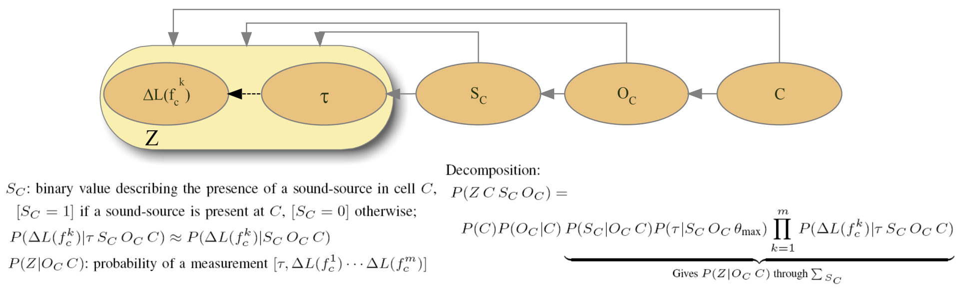

Fig. 5. Bayes network corresponding to the Bayesian Program for the binaural sensor model presented in [4]. The model's random variables are defined as follows: C indexes a cell in log-spherical space; O_C denotes the occupancy of a specific cell (either 0 for empty or 1 for occupied — this general case is needed for the BVM [5]); S_C signals the special case of the occupancy of a specific cell *with* a sound-source (either 0 or 1 for "not occupied by a sound-source" or "occupied by a sound-source", respectively); the remainder denote the binaural cue readings, which have been defined earlier on.

marginalisation (i.e. sum over all possible values of S_C), and is used as a sub-BP for the BVM in [5].

III. Calibration, Implementation and Results

As can be seen on the BP in Fig. 5, calibration of the binaural system involves the characterisation of the families of normal distributions $P(\tau|S_C\,O_C\,\theta_{\max})$ and $P(\Delta L(f_c^k)|\tau\,S_C\,O_C\,C) \approx P(\Delta L(f_c^k)|S_C\,O_C\,C)$ through descriptive statistical learning of their central tendency and statistical variability. This is done in an equivalent manner as with commonly used head-related transfer function (HRTF) calibration processes (see, for example, [17]) and is described in the following paragraphs.

A set M_c of n-dimensional measurement vectors such as defined in equation (3) is collected per cell $c \in \mathcal{C}$. The full set of collected measurement vectors for all cells in auditory sensor space $\mathcal{Y}$ is expressed as $M = \bigcup M_c$. Denoting $M_{\bar{c}} = M \setminus M_c$ as the set of measurements for all cells other than c, the statistical characterisation process of each family of distributions is effected for each cell c through

$$P(\tau|[S_c = 1]\,O_c\,\theta_{\max}) \equiv \mathcal{N}(\tau, \mu_\tau(M_c), \sigma_\tau(M_c)) \tag{4a}$$

$$P(\tau|[S_c = 0]\,O_c\,\theta_{\max}) \equiv \mathcal{N}(\tau, \mu_\tau(M_{\bar{c}}), \sigma_\tau(M_{\bar{c}})) \tag{4b}$$

$$P(\Delta L(f_c^k)|[S_c = 1]\,O_c\,c) \equiv \mathcal{N}(\Delta L(f_c^k), \mu_{\Delta L(f_c^k)}(M_c), \sigma_{\Delta L(f_c^k)}(M_c)) \tag{4c}$$

$$P(\Delta L(f_c^k)|[S_c = 0]\,O_c\,c) \equiv \mathcal{N}(\Delta L(f_c^k), \mu_{\Delta L(f_c^k)}(M_{\bar{c}}), \sigma_{\Delta L(f_c^k)}(M_{\bar{c}})) \tag{4d}$$

Auditory calibration is performed by presenting a broadband audio stimulus through a loudspeaker positioned in well-known spatial coordinates corresponding to the geometric centre of each cell $c \in \mathcal{C}$ so as to sample space according to the auditory sensor space $\mathcal{Y}$. The experimental setup used for this purpose is described in Fig. 6.

The acquisition method may be simplified by a factor of 4 by taking into account the spatial redundancies of auditory sensing, namely the symmetry enforced by the back-to-front ambiguity and the left-to-right antisymmetry for both ITDs and ILDs, to reduce calibration space to the front-left quadrant. A further simplification of the procedure consists in positioning the loudspeaker, for each of the N_d considered distances from the binaural system, precisely in front of the active perception head (i.e. $(\theta, \phi) = (0, 0)$) and to *rotate the active head* so that the whole range of azimuths and elevations of the auditory sensor space is covered. This replaces the several minutes taken to reposition the loudspeaker by hand (which this way only happens N_d times) by a few seconds of head motions for each cell. The full procedure is depicted in Fig. 7 for a typical calibration context.

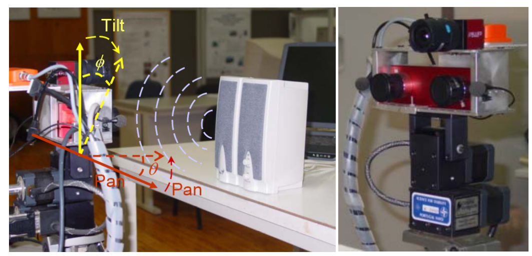

Fig. 6. Experimental setup for the binaural system calibration procedure.

Some results obtained for a single classical occupancy grid inference step as described earlier on are presented on Figs. 8 and 9, using a uniform prior on occupancy. These are outcomes of offline processing using MATLAB; the algorithms are currently being ported to C++ for realtime processing in active perception applications.

IV. Conclusions

Full 3D auditory localisation has rarely been explored in robotic applications (see, for example, [17] for a review on this subject); this work contributes with a novel probabilistic solution that produces localisation estimates based on binaural cues yielded by a robust binaural processing unit. This solution has been designed so as to provide a sensor model to be used by a multimodal perception framework, the Bayesian Volumtric Map, described in [5]. Further details can be found at `http://paloma.isr.uc.pt/~jfilipe/BayesianMultimodalPerception`.

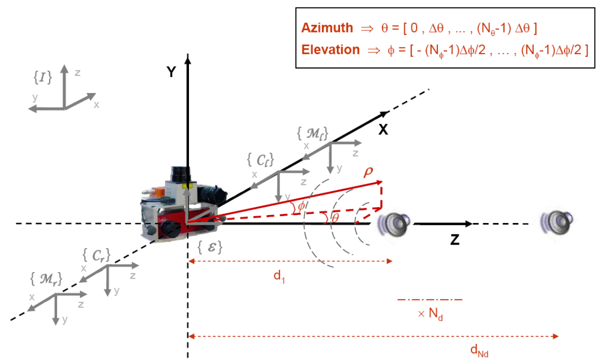

Fig. 7. Schematic of the experimental acquisition method of the auditory calibration procedure. A typical auditory sensor space was used, roughly based both on known human auditory precision ratings [17] and the specifications of the pan and tilt unit model PTU-46-17.5 from Directed Perception, characterised by an azimuthal range of $[0^o; 90^o]$ with $\Delta\theta = 2^o$ resolution and an elevation range of $[-30^o; 30^o]$ with $\Delta\phi = 10^o$ resolution, and a number of partitions $N = 4$, of which only the furthest $N_d = 2$ partitions would be used — the number of cells to sample amount to $[(90/2) \times (60/10) \times 2] = 540$. Thus, for 20 readings of a 1 s broadband stimulus per cell and approximately 1 s to set up each position for the loudspeaker, calibration would take about $10800 \times 2\,\mathrm{s} = 6\,\mathrm{h}$. This procedure allows partitioning calibration into one session per loudspeaker distance, in this case of $\approx 3\,\mathrm{h}$.

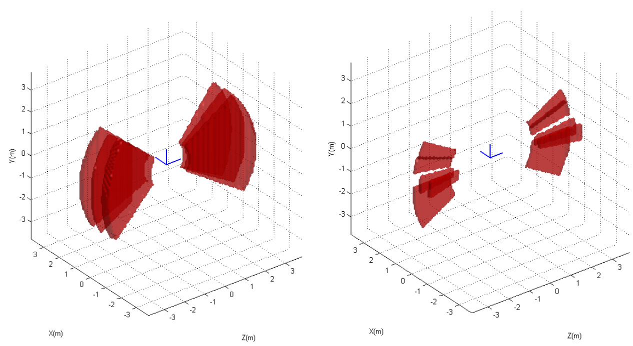

Fig. 8. Occupancy results for the processing of an audio snippet of a human speaker placed in front of the binaural perception system — cells within the log-spherical sensor-space with probabilities of occupancy greater than .75 are depicted in red, and the egocentric referential in blue (X-axis, Y-axis and Z-axis indicate right-to-left, upward and forward directions, respectively). On the left, result of inference using ITDs only; on the right, result of adding ILDs: note the effects on distance and elevation.

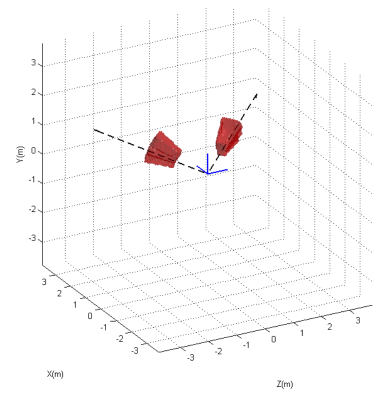

Fig. 9. Occupancy results for the processing of an audio snippet of a sound-source placed at $\rho = 1320\,\mathrm{mm}$, $\theta = 36^o$, $\phi = 20^o$. All that is depicted has the same meaning as in Fig. 8; two dashed directional lines at (θ, ϕ) and $(180^o - \theta, \phi)$ have been additionally plotted to demonstrate the effect of front-to-back confusion. This phenomenon can be countered in two different ways: either by rotating the perceptual system, causing the occupancy probabilities of the correct cells to be confirmed and of incorrect cells to be decreased by accumulated evidence with subsequent inference steps, respectively, or by using artificial *pinnae* so as to enforce asymmetry in the HRTF readings and performing calibration using a half-sphere instead of only a quadrant. The fact that $\theta >> 0^o$ means that precision in elevation and distance is improved as compared to Fig. 8.

References

[1] A. J. King, J. W. Schnupp, and T. P. Doubell, "The shape of ears to come: dynamic coding of auditory space," *TRENDS in Cognitive Sciences*, vol. 5, no. 6, pp. 261–270, 2001.

[2] S. B. Andersson, A. A. Handzel, V. Shah, and P. Krishnaprasad, "Robot Phonotaxis with Dynamic Sound-source Localization," in *ICRA*, 2004.

[3] J. Hörnstein, M. Lopes, J. Santos-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*, Beijing, China, October 2006.

[4] C. Pinho, J. F. Ferreira, P. Bessière, and J. Dias, "A Bayesian Binaural System for 3D Sound-Source Localisation," in *International Conference on Cognitive Systems (CogSys 2008)*, University of Karlsruhe, Karlsruhe, Germany, April 2008, pp. 109–114.

[5] J. F. Ferreira, P. Bessière, K. Mekhnacha, J. Lobo, J. Dias, and C. Laugier, "Bayesian Models for Multimodal Perception of 3D Structure and Motion," in *International Conference on Cognitive Systems (CogSys 2008)*, University of Karlsruhe, Karlsruhe, Germany, April 2008, pp. 103–108.

[6] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. AcoustS. Soc. Am*, pp. 1890–1894, 1995.

[7] E. de Boer, "Synthetic whole-nerve action potentials for the cat," *Acoustical Society of America Journal*, vol. 58, pp. 1030–1045, Nov. 1975.

[8] L. V. Immerseel and S. Peeters, "Digital implementation of linear gammatone filters: Comparison of design methods," *Acoustics Research Letters Online*, vol. 4, no. 3, pp. 59–64, July 2003, Published online by the Acoustical Society of America.

[9] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and Rice, "SVOS Final Report: The auditory filterbank," MRC Applied Psychology Unit, Cambridge., Tech. Rep. 2341, 1988.

[10] B. Kapralos, M. R. M. Jenkin, and E. Milios, "Auditory Perception and Spatial (3D) Auditory Systems'," York University, Tech. Rep. CS-2003-07, July 2003.

[11] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, March 2000.

[12] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, November 2004.

[13] P. M. Zurek, "The precedence effect," in *Directional Hearing*, W. Yost and G.Gourevitch, Eds. Sprigler-Verlag, 1987.

[14] N. F. Viemeister and G. H. Wakefield, "Temporal integration and multiple looks," *The Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 858–865, August 1991.

[15] A. V. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, N. P.-H. E. Cliffs), Ed., 1989.

[16] O. Lebeltel, "Programmation Bayésienne des Robots," Ph.D. dissertation, Institut National Polytechnique de Grenoble, Grenoble, France, September 1999.

[17] P. T. Calamia, "Three-dimensional localization of a close-range acoustic source using binaural cues," Master's thesis, Faculty of the Graduate School of The University of Texas at Austin, 1998.