IMPEP – Integrated Multimodal Perception Experimental Platform

A Bayesian Binaural System for 3D Sound-Source Localisation

February 2008

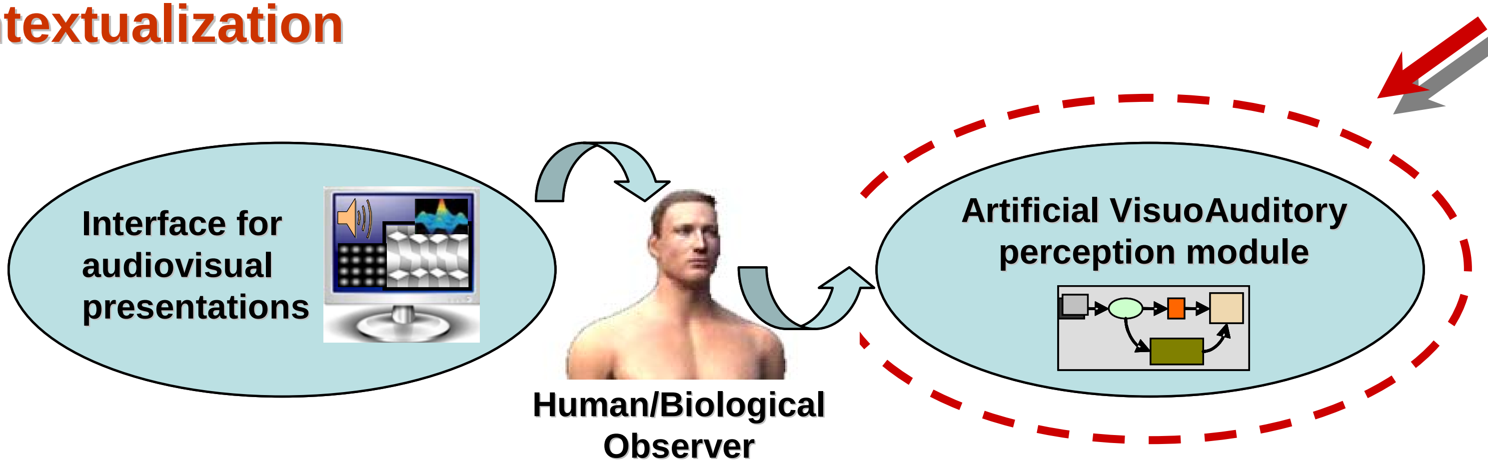
Contact Persons:

Cátia Pinho, João Filipe Ferreira, Jorge Dias
Email: {catiap; jorge; jfilipe}@isr.uc.pt

BACS
BAYESIAN APPROACH TO COGNITIVE SYSTEMS

Mobile Robotics Laboratory
FCT-UC
ISR – Coimbra

Contextualization

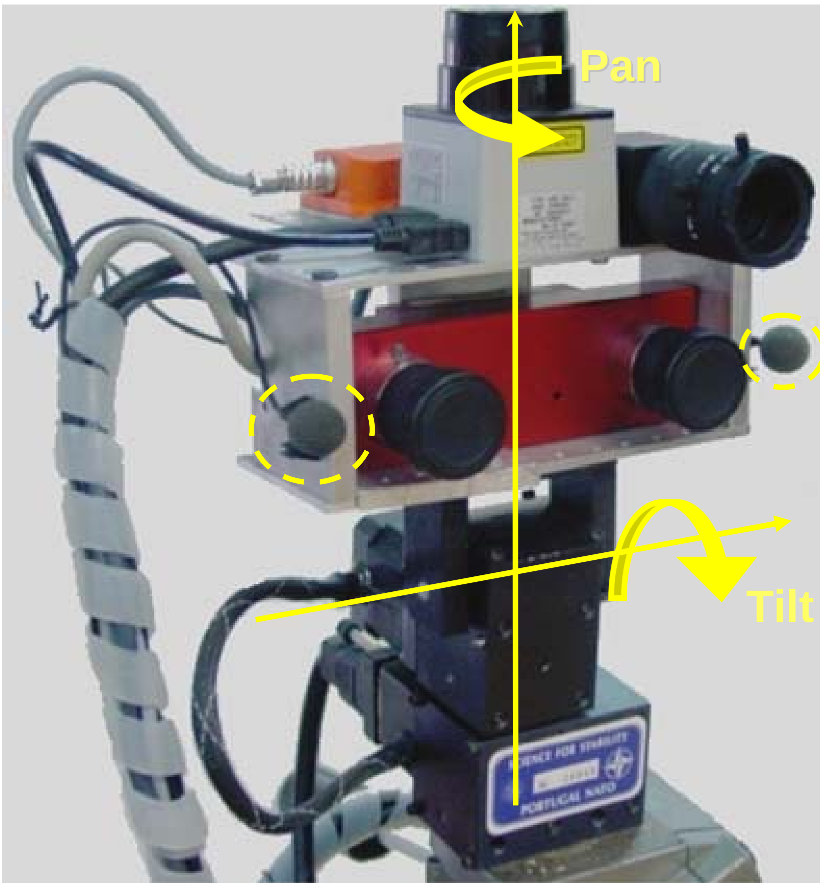


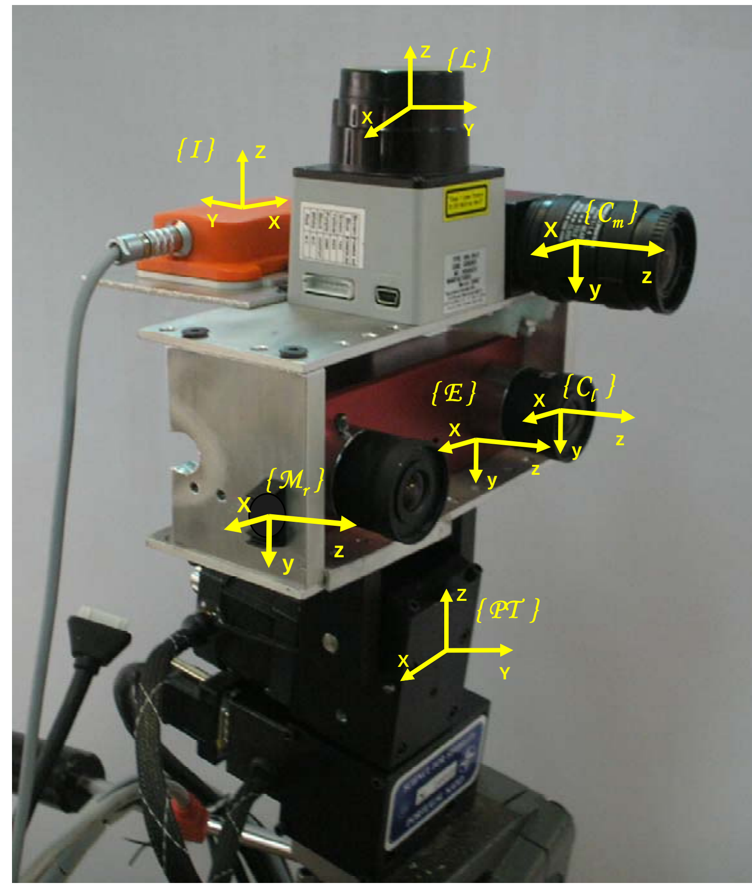
Modelling Approach

- To develop a **visuoauditory perception solution** with the ability to yield **probabilistic estimates** for the complete frame of 3D information:
  - *azimuth* (θ), *elevation* (ϕ), *distance* (ρ)

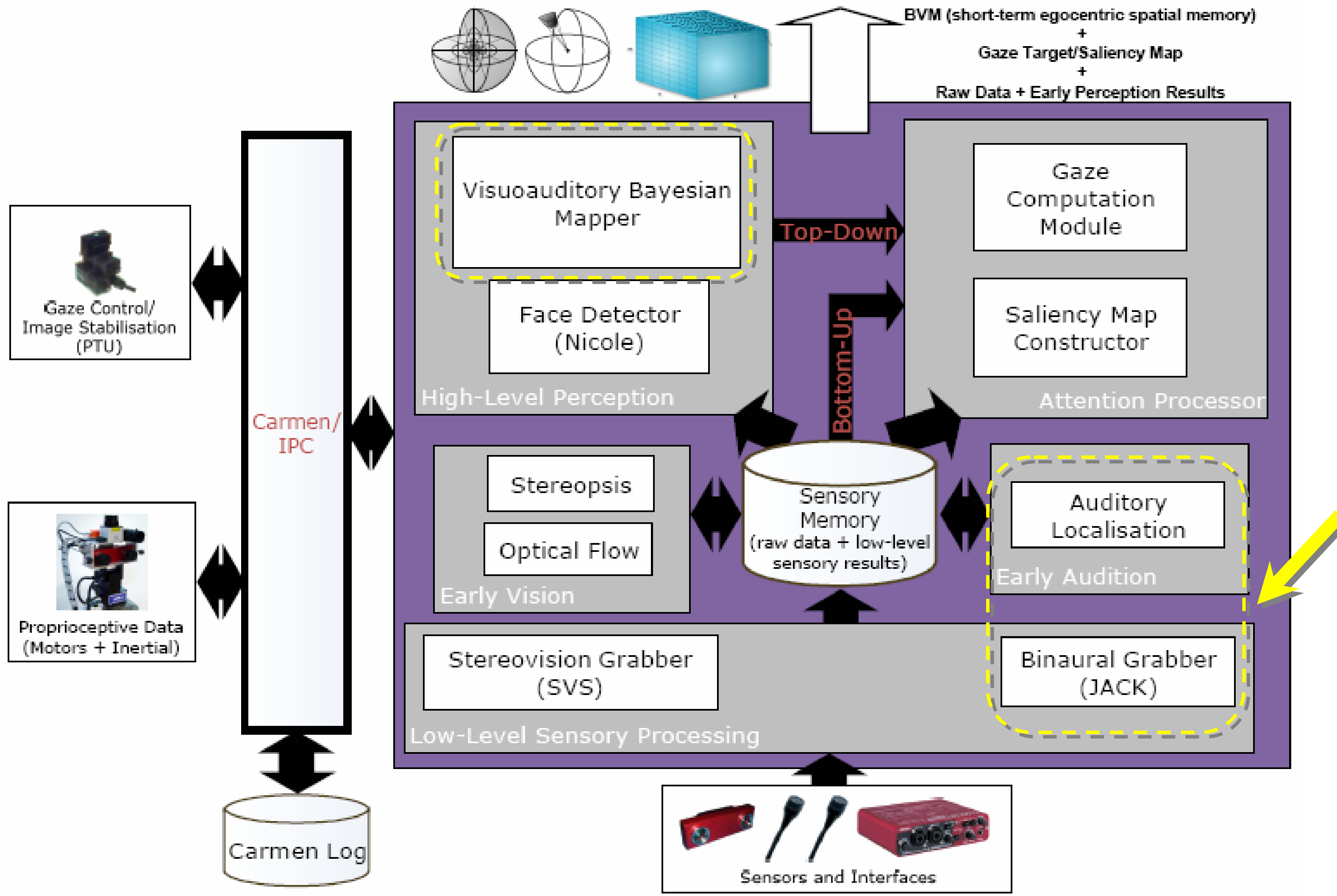
IMPEP 1 (Integrated Multimodal Perception Experimental Platform)

Platform description

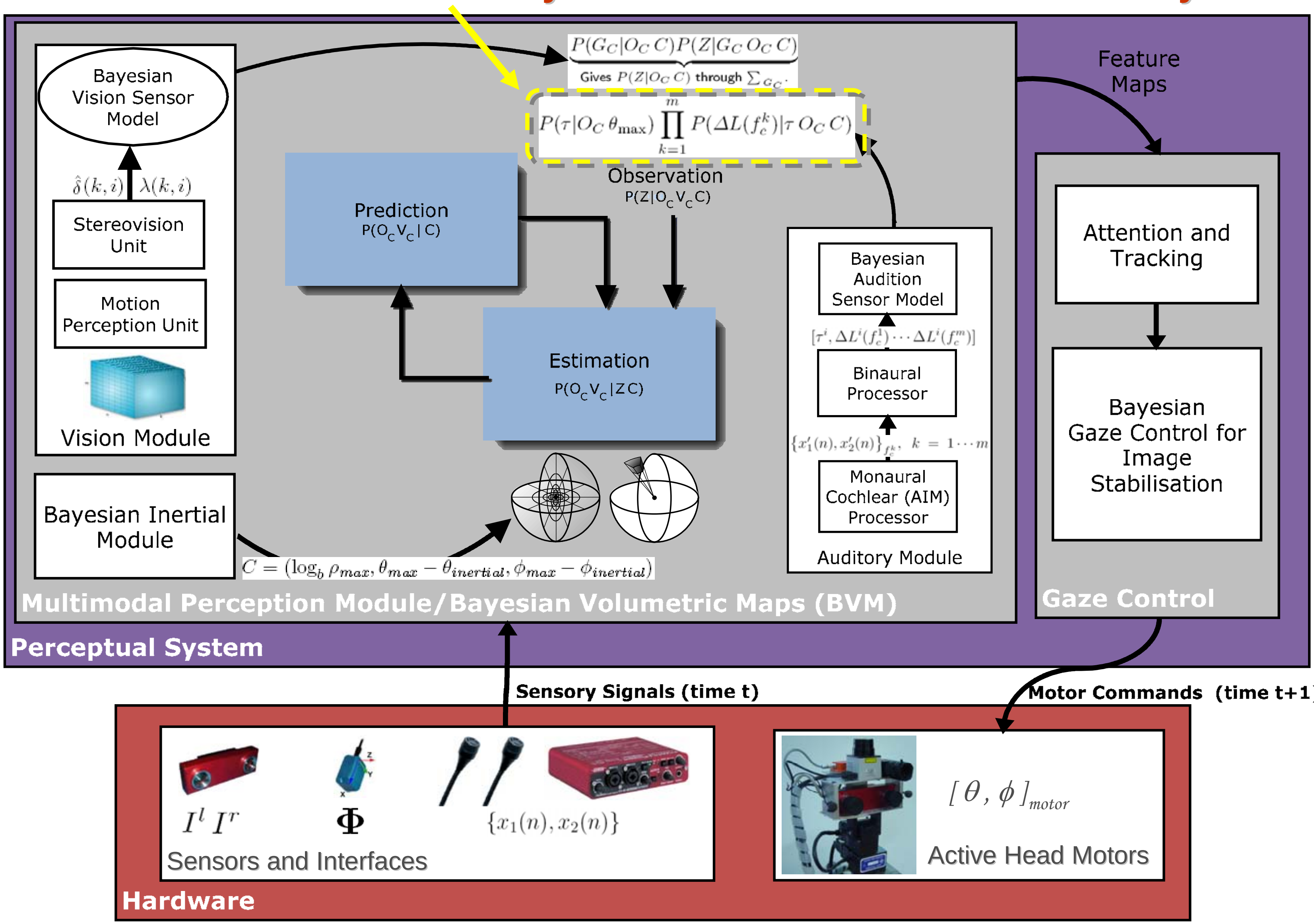




IMPEP - Outputs (*to Multimodal Integration System*)



Contextualisation of Auditory Sensor Model within the IMPEP System

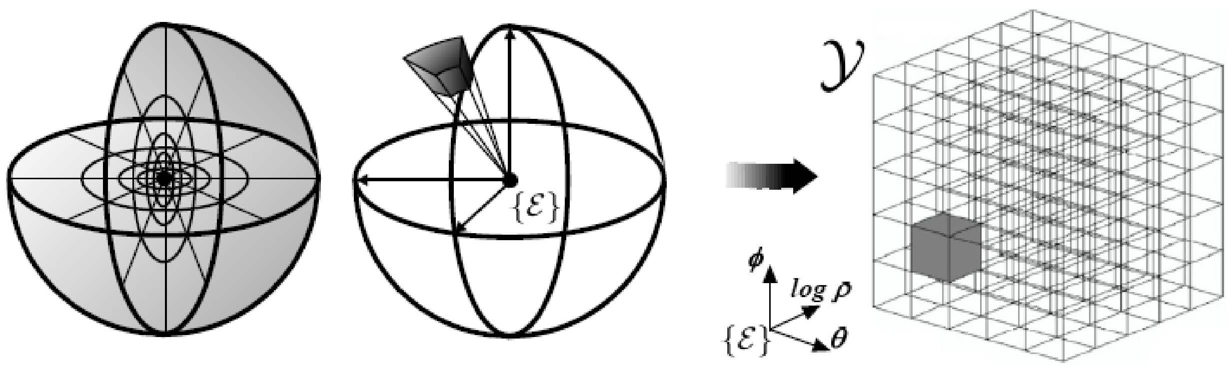


BACS
BAYESIAN APPROACH TO COGNITIVE SYSTEMS

ISR

MRLAB
Mobile Robotics Laboratory
FCT-UC
ISR – Coimbra

BVM - Bayesian Volumetric Map



- An egocentric, log-spherical spatial memory map has been devised as a framework for multimodal sensor fusion, named the Bayesian Volumetric Map (BVM)
- This map stores the independent probabilistic states of occupancy and velocity for each cell in a volumetric grid with log-spherical configuration, defined by the domain $\mathcal{Y}$:

$$\mathcal{Y} \equiv \left] \log_b \rho_{Min} ; \log_b \rho_{Max} \right] \times \left] \theta_{Min} ; \theta_{Max} \right] \times \left] \phi_{Min} ; \phi_{Max} \right]$$

DASM - Direct Auditory Sensor Model

- Defined as a ***normally distributed*** BVM operator

$$P(Z \mid O_C C) = P(\tau \mid O_C\ \theta_{\max}) \prod_{k=1}^{m} P\left(\Delta L\left(f_c^{\,k}\right) \mid \tau\ O_C\ C\right)$$

C – *Cell Identifier* ($\log_b \rho_{max}$, θ_{max}, ϕ_{max})

Z – *Measurement taken by the audio sensor (generic notation)*

O_C – *Occupancy of cell C (0 = no sound-source; 1 = sound-source)*

τ – *frequency invariant interaural time differences (**ITD**s) - ms*

$\Delta L\left(f_C^{\,k}\right)$ – *frequency dependent interaural level differences (**ILD**s) - dB*

- The direct sensor model is used to update the BVM by ***inverting the direct model*** (DASM) using Bayesian inference so as to estimate the probability of occupancy for each cell

- ***Note:*** the angular and range resolution of the auditory sensor space is typically lower than the BVM

IASM - Inverse Auditory Sensor Model

$$P([O_C=1] \mid ZC) = \frac{P([O_C=1] \mid C)\,P(Z \mid [O_C=1]C)}{\sum_{O_c=0,1} P(O_C \mid C)\,P(Z \mid O_C C)}$$

$$= \frac{P([O_C=1] \mid C)\,P(Z \mid [O_C=1]C)}{P([O_C=0] \mid C)\,P(Z \mid [O_C=0]C) + P([O_C=1] \mid C)\,P(Z \mid [O_C=1]C)}$$

DASM

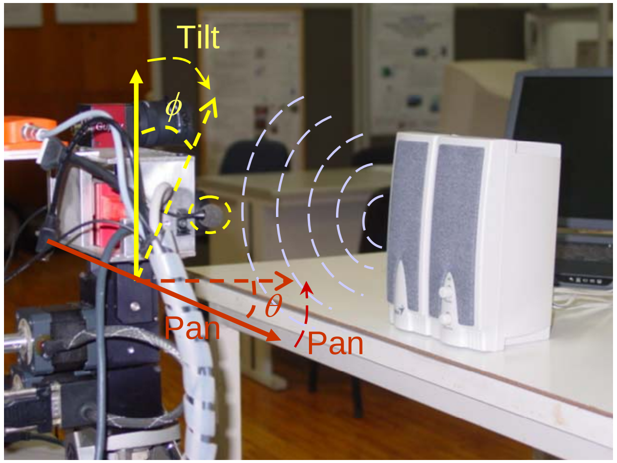
$$P(Z \mid [O_C=0,1]C) \begin{cases} O_C=0 \;\rightarrow\; P(\tau \mid [O_C=0]C)\prod_{k=1}^{m} P\left(\Delta L\left(f_c^k\right) \mid \tau [O_C=0]C\right) \\ O_C=1 \;\rightarrow\; P(\tau \mid [O_C=1]C)\prod_{k=1}^{m} P\left(\Delta L\left(f_c^k\right) \mid \tau [O_C=1]C\right) \end{cases}$$

Objective

- The auditory calibration's purpose is to characterise the normal distributions of the ***Direct Auditory Sensor Model (DASM)***
- This will allow the full localisation of sound-sources in three-dimensional space:
  - Azimuth *(θ)*
  - Elevation *(ϕ)*
  - Distance *(ρ)*
- **Final aim:** To feed the BVM framework with an accurate direct audition sensor model so as to allow multimodal perception of 3D structure and motion.

Experimental plan

- ***Goal****:* to capture binaural readings using the stereo microphones of the IMPEP Active Perception Head for each cell in the auditory sensor space of a broadband noise sound-source (1s)

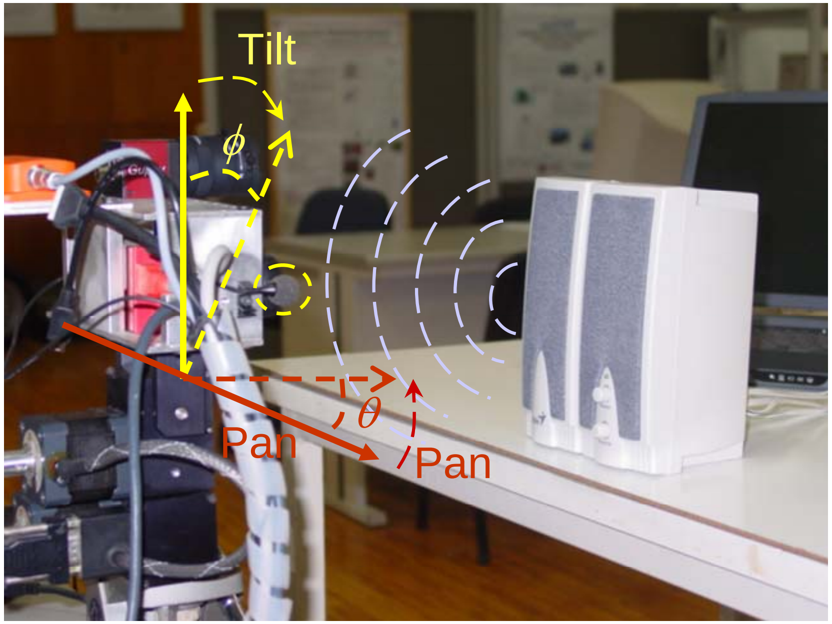


- To cover the complete auditory sensor space, the sound-source must be positioned at the centre of each grid cell on the BVM

- ***Methodology****:* sound-source is positioned at a specific distance ***(*** ρ ***)*** from the IMPEP head, directly facing the front of the Pan & Tilt Unit (PTU), and the corresponding relative rotation is performed by the PTU;
  The different distances (ρ) between the sound-source and the IMPEP head are obtained manually.

- To avoid redundancies $\Rightarrow$ ***only one quadrant*** is used due to:
  - *Symmetry from* ***front-back confusion phenomenon***
  - *Left-Right anti-symmetry (ITD = - ITD and ILD = - ILD)*

Experimental plan (example)



Resolution (Pan / Tilt):

- 2 degree of azimuth (θ) – performed by the **Pan** motor – **360º / 2º = 180 cells (θ)**
- 10 degrees of elevation (ϕ) – performed by the **Tilt** motor – **180º / 10º = 18 cells (ϕ)**

- Acquisition for **N_d = 2** different distances ⇒ d : d_1 ~ 0.32m ; d_2 ~ 3.2 m **= 2 cells (ρ)**

Experimental plan (example contd.)

- N_d – *number of distances*
- N_m – *numbers of measurements in each cell*

To avoid redundancies ⇒ ***only one quadrant*** is used

Because:

- *Symmetry from* ***front-back confusion phenomenon***
- *Left-Right anti-symmetry (ITD = - ITD e ILD = - ILD)*

Auditory sensor space angular ranges simplify to (including PTU spec limitations for elevation):

Azimuth (θ) : 90º / 2º = **45 cells**

Elevation (ϕ) : (- 30º to 30º) / 10º = **6 cells**

Consequently: *Azimuth meas.* *Elevation meas.*

- $[N_d \times (45 \times 6)] \times N_m = 2 \times 270 \times N_m =$ $\mathbf{540 \times N_m}$ sets of measurements

  (*N_m measurements = 20 stimulus*) in each place to perform a statistical description

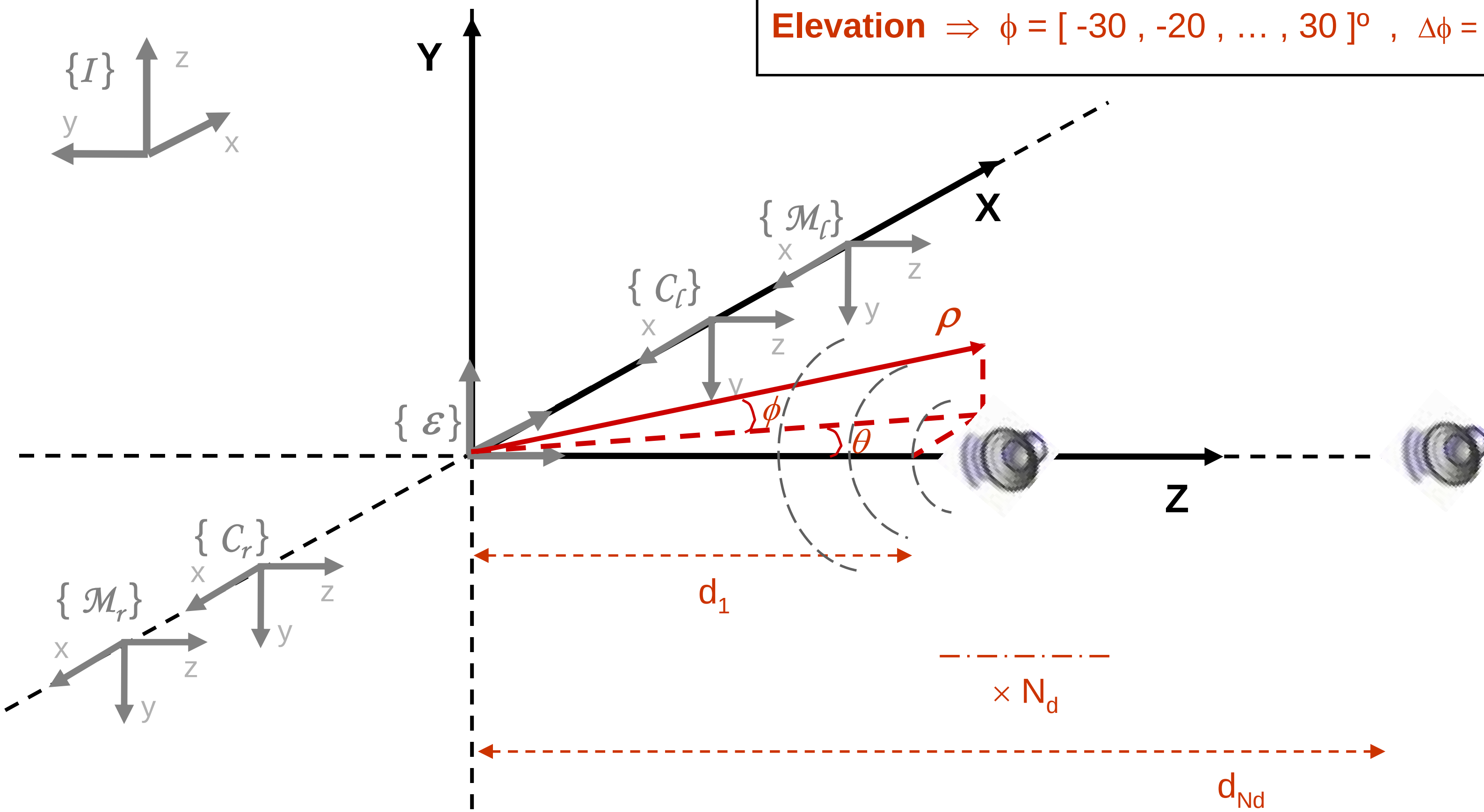
Becoming:

- $540 \times 20 = 10800$
- if each measurement takes 1s plus 1s of pause (play / record), calibration for each distance (i.e. the calibration process is conveniently divided into N_d sessions) will take: $10800 \times 2s / 2$ = ***3 h***

Schematic of the experimental acquisition

Azimuth ⇒ $\theta = [\,0 , 2 , \ldots , 90\,]^{o}$, $\Delta\theta = 2$

Elevation ⇒ $\phi = [\,-30 , -20 , \ldots , 30\,]^{o}$, $\Delta\phi = 10$



Processing of the data given by Auditory Calibration

- Mathematical formulation - *Direct Auditory Sensor Model*

$$P(\tau \mid O_C\ \theta_{\max}) \prod_{k=1}^{m} P\left(\Delta L\left(f_c^{\,k}\right) \mid \tau\ O_C\ C\right)$$

- Measurement definitions (after applying AIM and binaural processing to binaural readings, [1,2,3,4]):

ITD   *M × ILD*

$$C \rightarrow \left\{ \tau\ ,\ \Delta L\left(f_c^{\,1}\right), \ldots, \Delta L\left(f_c^{\,M}\right) \right\}_c^1$$

$$\ldots$$

$$\left\{ \tau\ ,\ \Delta L\left(f_c^{\,1}\right), \ldots, \Delta L\left(f_c^{\,M}\right) \right\}_c^{N_m = 20}$$

M_C Measurements for cell *C*

Measurements for all cells

$$M = M_C \cup M_{\overline{C}}$$

Measurements for cell *C*

Measurements for all other cells except *C*

Processing of the data given by Auditory Calibration

- *Normal probability distribution statistical characterisation process:*

$$P(\tau \mid O_C\ \theta_{\max}) = \begin{cases} P(\tau \mid \underbrace{[O_C = 1]}_{\text{Cell occupied}}\ \theta_{\max}) \Rightarrow \underbrace{\mu_{M_C}}_{\text{Measurements average}}(\tau)\ ,\ \underbrace{\sigma_{M_C}}_{\text{Standard deviation}}(\tau) \\ P(\tau \mid \underbrace{[O_C = 0]}_{\text{Cell not occupied}}\ \theta_{\max}) \Rightarrow \mu_{M_{\bar{C}}}(\tau)\ ,\ \sigma_{M_{\bar{C}}}(\tau) \end{cases}$$

$$P(\Delta L \mid \tau\ O_C\ C) \approx \begin{cases} P(\Delta L \mid \underbrace{[O_C = 1]}_{\text{Cell occupied}}\ C) \Rightarrow \underbrace{\mu_{M_C}}_{\text{Measurements average}}(\Delta L)\ ,\ \underbrace{\sigma_{M_C}}_{\text{Standard deviation}}(\Delta L) \\ P(\Delta L \mid \underbrace{[O_C = 0]}_{\text{Cell not occupied}}\ C) \Rightarrow \mu_{M_{\bar{C}}}(\Delta L)\ ,\ \sigma_{M_{\bar{C}}}(\Delta L) \end{cases}$$

Conclusions

- The auditory calibration's purpose is of characterising the normal distributions of the **DASM** (Direct Auditory Sensor Model) is thus solved efficiently.
- This will allow the full localisation of sound-sources in three-dimensional space:
  - Azimuth *(θ)*
  - Elevation *(ϕ)*
  - Distance *(ρ)*
- within the BVM framework

References:

[1] - Faller, C. and Merimaa, J. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. J. Acoust. Soc. Am., 116:30753089, 2004.

[2] - Akeroyd, M. A. A Binaural Cross-correlogram Toolbox for MATLAB. February 2001.

[3] - Johannesma, P. I. M. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In Symposium on Hearing Theory, pages 5869. IPO, Eindhoven, The Netherlands, 1972.

[4] - Slaney, M. Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work. Technical report, Interval Research Corproation, 1998.

[5] - C.J. Lee, S.D. Wang, A Gabor filter-based approach to fingerprint recognition, in: IEEE Workshop on Signal Processing System, SiPS 99, pp. 371-378, 1999.

[6] - Dankers, A., Barnes, N., and Zelinsky, A. Primate structures in synthetic dynamic active visual saliency. 6th International Conference on Epigenetic Robotics (Epirob 2006), Modeling Cognitive Development in Robotic Systems, Paris, France 2006.

[7] - Itti, L., Koch, C., and Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11): 1254–1259, November 1998.

[8] - Shic F, Scassellati B. A Behavioral Analysis of Computational Models of Visual Attention. International Journal of Computer Vision, 73(2):159-177, Jun. 2007.