FP7-ICT-2013-C TWO!EARS Project 618075

Deliverable D4.1, part A

Executive summary

Jens Blauert, Thomas Walther *

November 26, 2014

* The TWO!EARS project (`http://www.twoears.eu`) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 618075.

Project acronym:	TWO!EARS
Project full title:	Reading the world with TWO!EARS

Work package:	4
Document number:	D4.1, part A
Document title:	Executive summary
Version:	1

Delivery date:	30th November 2014
Dissemination level:	Restricted
Nature:	Report

Editor(s):	Thomas Walther, RUB
Author(s):	Jens Blauert, Thomas Walther
Reviewer(s):	Klaus Obermayer

1 Executive summary

As stated in the project proposal, TWO!EARS is going to challenge current thinking in auditory modeling by replacing common paradigms in this field with a systemic approach, whereby human listeners are regarded as multi-modal agents that develop their concept of the world by exploratory interaction. The goal of TWO!EARS is to develop an intelligent, computational model of active auditory perception and experience in a multi-modal context. The resulting system framework will form a structural link from binaural perception to judgment and action, realized by interleaved signal-driven (bottom-up) and hypothesis-driven (top-down) feedback processing within an innovative expert-system architecture.

The main objective of WP4 in this context is to set up a framework that endows the TWO!EARS system with active-listening capabilities. To capture bottom-up data processing as well as the top-down mechanisms as required for active listening tasks, the proposed framework must host suitable feedback loops. It is the general task of WP4 to design an appropriate system architecture for this requirement, investigate meaningful feedback paths, implement them, and finally evaluate them with respect to their functionalities. Input from modalities other than the auditory one will also be considered as a source of feedback information, particularly, position, direction of speed of head-&-torso movements (proprioceptive and sensorimotor input), further, identified optical objects (visual input).

WP4 is split into five tasks. This deliverable mainly entails our advance on the key task for the current project period, namely, task 4.1. As this task consists of several sub-tasks, the remainder of our report is organized as follows.

- **D4.1, part B** A consolidated literature survey regarding "active listening" has been compiled, summarizing available information about the mechanics and functions of feedback in the auditory system. Also, multi-modal approaches have been reviewed and evaluated with regard to their value for TWO!EARS. Herein, particular focus has been put on the two project-specific proof-of-concept applications, that is search-&-rescue and quality of experience.
- **D4.1, part C** Possible and meaningful feedback loops have been identified that might be set up within the actual TWO!EARS model framework, resulting in an updated list of entry ports for feedback and a list of sources for feedback information.

- **D4.1, part D** A supporting document that describes the planned or realized integration of feedback-related concepts and mechanisms in the current TWO!EARS architecture. Though such supporting information is not a mandatory component of D4.1, it was added to provide additional explanations regarding key achievements in this project period.

Significant results in the current project period

The following enumeration highlights the most significant results as have been achieved in the past twelve months and links them to the tasks addressed by WP4. Although the focus was clearly put on the completion of task 4.1, progress was made as well in the formulation and implementation of feedback loops (task 4.2), the realization of cross-modal input (task 4.3), and feedback-related testing and labeling (task 4.5).

- **Task 4.1 – Aspects of active listening** The key achievements in this task have been the creation of a consolidated literature survey on "active listening" and adapted and enhanced feedback port/information lists. These lists provides an update of the possible range of entry ports for feedback and enumerate sources for feedback information – again with specific focus on the proposed system architecture and the relevant scenarios, namely search-&-rescue and quality of experience. During the first twelve months of the project, we stepped far beyond mere analysis of possible feedback paths by setting up a virtual test environment, the *Bochum Experimental Feedback Testbed* (BEFT), to enable some initial testing regarding more complex feedback mechanisms, such as active exploration.

- **Task 4.2 – Implementation of feedback loops** Starting in the second half of the first project phase, two feedback mechanisms could already be realized in the current TWO!EARS system. First, a basic source disambiguation approach has been set up that allows to turn the robot's head in order to verify hypothetical positions of sound sources in a given scenario. Variations in the location estimates of the hypothetical sound stimuli were evaluated in order to eliminate false positives. Note that the observed variations are caused by the head motions of the robot. Thus, the proposed technique offers a proof-of-concept for the advantages of feedback-enhanced scene analysis. Second, a further interesting feedback mechanism has been realized in the current project period, namely, the so-called *dynamic weighting* approach. Herein, the standard head-turn reflex is augmented with a reflective component that allows the robot to focus its attention on important scenario elements, neglecting observed stimuli that are not relevant for the actual task. In the further course of the project, the initial BEFT/MATLAB® implementation of dynamic weighting will be ported, in cooperation with WP5, to operate in the MORSE simulator driven by the TWO!EARS system – see D4.1, part D).

- **Task 4.3 – Cross-modal input** For operations on cross-modal input, the Two!Ears system has been augmented with the MORSE robot simulator. This virtual-reality component allows us to perform initial tests on simulated visual input. To that end, images from the virtual cameras of a robot, as simulated in MORSE, are captured via a "computer-vision-specific" node of the ROS system – compare D4.1, part D. The images are then processed using routines from the OpenCV computer vision library in order to detect faces in videos that have been artificially inserted into the virtual environment. First results indicate good detection performance – also for multiple sources – at acceptable computational costs. On this basis, we intend to step into audio-visual, that is, multi-modal feature analysis in cooperation with WP3.

- **Task 4.5 – Testing & labeling** Already early in the project it was realized that, in order to be able to test active exploration, visual processing techniques, and multi-modal feedback strategies, the Two!Ears architecture needs to be amended by a vision-related virtual-reality component. This was the motivation to develop the *Bochum Experimental Feedback Testbed* (BEFT) that allows to test complex feedback ideas and enables access to multi-modal, here visual, input data. In cooperation with WP1, category and environmental labels defined for BEFT have selectively been transferred into XML scene files as are used by the auralization component of Two!Ears. This practice not only causes partial data fusion in both systems, but also paves the way for the generation of full-scale multi-modal scenario descriptions. Note that the MORSE simulator has inherited from BEFT by now (see D4.1, part D), and allows to emulate complex environments for feedback testing. In the further course of the project, MORSE will be employed to set up feedback-related scenarios that significantly exceed the basic scenes as currently defined, thus stepping into active exploration and complex multi-modal-scenario analysis.

FP7-ICT-2013-C TWO!EARS Project 618075

Deliverable D4.1, Part B

Consolidated Literature Survey

WP4 *

November 16, 2014

* The TWO!EARS project (http://www.twoears.eu) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 618075.

Project acronym:	TWO!EARS
Project full title:	Reading the world with TWO!EARS

Work package:	4
Document number:	D4.1, Part B
Document title:	Consolidated Literature Survey
Version:	1

Delivery date:	30th November 2014
Actual publication date:	–
Dissemination level:	Restricted
Nature:	Report

Editor:	Blauert, Jens, RUB
Authors:	Argentieri, Sylvain
	Blauert, Jens
	Brown, Guy
	Bustamante, Gabriel
	Cohen-Lhyver, Benjamin
	Danès, Patrick
	Gas, Bruno
	Kim, Ryan
	Kohlrausch, Armin
	Ma, Ning
	Walther, Thomas
Reviewer:	Obermayer, Klaus, TUB

Contents

1 Introduction

Modeling active listening implies various feedback mechanisms. Consequently it is a unique feature of the Two!Ears project to include feedback loops into their model of binaural listening. Incorporation of feedback loops into engineering models of sensory perception and cognition reaches out to the edge of current knowledge and has, to our best knowledge, never been tried before in a comprehensive way with regard to audition.

There is strong evidence for numerous feedback loops in biological auditory system – Fig. 1.1 (Schofield, 2009). Some physiologist even claim that the efferent fibers outnumber the afferent ones by far, for instance, Shamma (2013).

Nevertheless, knowledge regarding their functional relevance is rather sparse (He and Yu, 2009). Two!Ears thus, although using physiological findings as a source of inspiration, will take a strictly operational, that is, engineering approach.

Fig. 1.2 provides a list of specific functional improvements that the consortium hopes to achieve by the inclusion of feedback into the Two!Ears system.

Specific feedback activity will be triggered when the signal- and/or symbol-processing stages render output that shows a lack of confidence – such as those listed in Fig. 1.3. Feedback will be activated in these cases in order to modify the processing algorithms and/or to provide additional information to the system.

Fig. 1.4 discusses suitable means for the cases that the lack of confidence stems from too high variances at the signal level, logical inconsistencies of the symbolic level or implausibilities at the level of meaning.

In a second step, it is evaluated from where further up in the model control information can be obtained in signal or symbolic form – as is suitable to improve the performance of the model. Thereby it has to be kept in mind that functional improvements of the model are task-specific and can only be assessed with regard to the actual purpose of the model. Fig. 1.6, taken from page 16 of the Two!Ears proposal, lists ideas as to from where useful control information can be obtained and what functional improvements can be expected by applying it.

It is the purpose of the current document to provide an overview of relevant literature with respect to feedback in the auditory system. In this context, cross-modal, in particular,

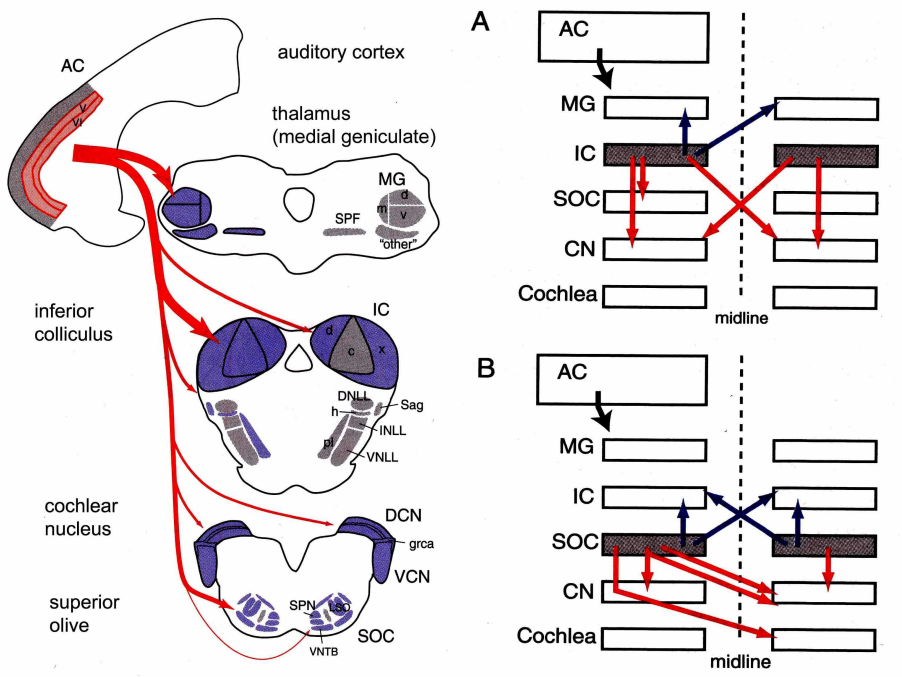


Figure 1.1: Physiological evidence of feedback loops in the auditory system – synopsis of three figures, taken from Schofield (2009)

- Turning the acoustic sensors into optimal position (*turn-to reflex*)
- Cognition-controlled exploration of the environment by means of active head-&-torso movements
- Increasing signal-to-noise ratio by specific enhancement of spectral and temporal selectivity *("sharpening the ears")*
- Activation of specific signal-processing procedures, such as *echo-cancelling, de-reverberation, precedence-effect* preprocessing, reconsideration to solve ambiguities
- Paying attention to specific signal features to deliver additional information as required by the cognitive stage
- Improvement of object recognition, auditory grouping, aural-stream segregation, aural-scene analysis
- Improvement of scene understanding, assignment of meaning, quality judgments, *attention focusing* on specific events

Figure 1.2: Positive performance effects expected from feedback

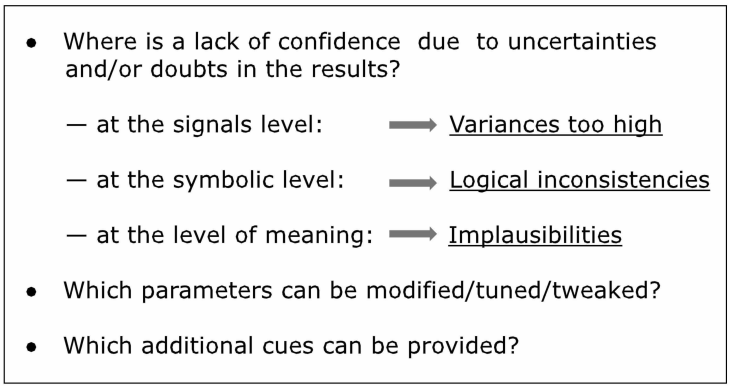


Figure 1.3: Reasons to initiate feedback

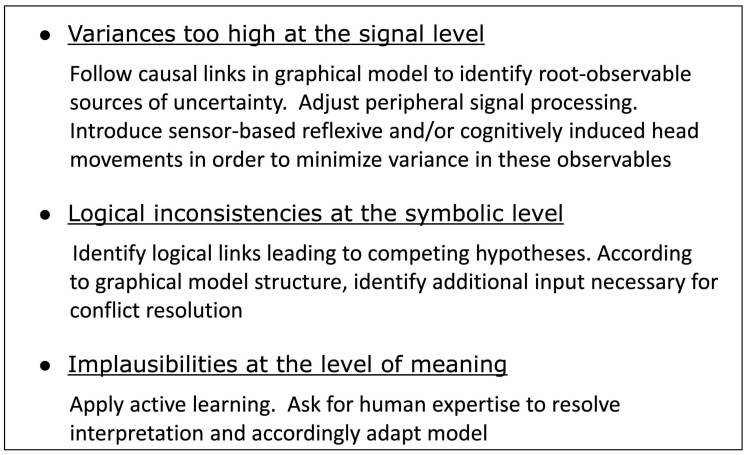


Figure 1.4: Road-map for feedback initialization and actions

visual and proprioceptive cues will be considered where appropriate. It will be taken into account that the output of the auditory system manifest itself not only as auditory percepts, such as auditory events and auditory scenes, but also as auditorily induced head and body movements. Further, the listening persons may describe their auditory events not only in a neutral, narrative way, but also in weighted form, for example, as quality judgments.

The following chapters deal with four important aspects of auditory feedback that are of particular relevance in the context of the Two!Ears modeling sytem. The collection will be amended as the project progresses.

Source of feedback signals and/or symbolic feedback information	Expected functional improvements
– Binaural-processing stage (brainstem level) and visual cues – Cognitive stage (experts, blackboard, scheduler)	– Turning the acoustic sensors into optimal position (turn-to reflex) – Advanced movements of the head-&-torso platform (exploring the environment)
– Brainstem level (superior olivary complex, MSO, LSO) – Cognitive stage (pre-segmentation, experts, blackboard, scheduler)	– Increasing the signal-to-noise ratio, increasing spectral and temporal selectivity – Paying attention to specific signal features to deliver specific additional information as required by the cognitive stage
– Binaural-activity mapping stage – Cognitive stage (pre-segmentation, experts, blackboard, scheduler)	– Activation of specific (computationally more expensive) signal-processing procedures, such as echo cancelling, de-reverberation, precedence-effect preprocessing, re-evaluation (reconsideration) to solve ambiguities
– Optical and sensorimotor sensors of the head-&-torso platform – Cognitive stage (pre-segmentation, experts, blackboard, scheduler)	– Optimal positioning of the head-&-torso platform (task-specific) – Improvement of object recognition, auditory grouping, aural stream segregation, aural-scene analysis, attention focusing
– External knowledge sources including cross-modal information	– Improvement of scene understanding, assignment of meaning, quality judgements, attention focusing

Figure 1.5: Examples of entry ports for feedback in the Two!Ears system and possible actions induced – taken from the proposal

Potential entry port for feedback	Possible control actions
Head-&-torso on movable cart	– 6-DOF rotational and translatory sensor movements
Peripheral modules (cochlea models)	– Adjustment of filter bandwidths and shapes, focusing on specific spectral regions, adjustment of operation points and dynamic ranges of operation
Monaural and binaural-processing stages (brainstem-level modules, MSO, LSO)	– Adjustment of time-windows, time constants and spectral regions. Task-specific employment of additional processing steps, e.g. lateral and contra-lateral inhibition, precedence preprocessing, de-reverberation
Binaural-activity-mapping stage (midbrain level, IC)	– Setting time constants for contra-lateral inhibition, providing masks for dedicated analyses of binaural-activity maps, focusing on specific spectral regions, adjustment of operation points and dynamic ranges – Provision of non-auditory sensory data, e.g. from vision, proprioception, sensorimotor cues
Cognitive stages (pre-segmentation, blackboard, experts, scheduler)	– Provision of external knowledge, e.g., salient features, object-building schemata, rule-systems; knowledge of the situational history, communicative intention of sound sources, task-specific expert knowledge, internal references – Provision of non-auditory knowledge, e.g. from visual scene analyses

Figure 1.6: Points of origin of feedback to be delivered to the feedback-entry ports of the Two!Ears system – taken from the proposal

2 Exploratory movements of the robot

Nichts ist unergründlicher als das System der Motivation unseres Handelns [1]

Georg Christopher Lichtenberg
The Mirror of the Soul

This chapter lays the foundations of active exploration in autonomous robots – as described in Chap. 5 – by listing some of the major brain structures involved in the low-level mechanisms leading to the comprehension of the environment. This comprehension is essential for robotic agents to make relevant decisions quickly – especially in *search and rescue* (S&R) scenarios. Since animals, humans and rodent in particular, show some impressive skills on localization and navigation in unknown environments, understanding the neural mechanisms responsible for such performances is a key step for further powerful and efficient implementations on robotic platforms.

The first part of this chapter, *Biological considerations*, is organized as follow. First, the very-low-level *neural* and, let's say, *neuronal*, structures that contribute to the emergence of an internal, robust, and adaptive model of the environment will be described. Secondly, the way streams of information from the sensors, that is, from eyes and ears, are processed will be discussed through an innovativing approach called *reverse-hierarchy theory*. Thirdly, the importance of self-implication in the exploration of the environment will be introduced by a section about *head movements*. Head movements are indeed the most common movements involved in active exploration.

The second part, *Bioinspired exploration models*, lists three major contributions to the environment-exploration problem in robotics, those three relying on strong biological foundations as detailed on the first part. The work of Cuperlier on transition maps, the RatSlam and PsiKharpax projects, based on the Animat approach, are the three contributions included in this part.

Finally, the third part, *Motivations for exploration*, adresses the question of what leads a human to explore its environment. This notion of motivation is a key to the active-exploration paradigm. Indeed, once a part of the environment, for instance, a room with a door, has been perceived, processed, learned and understood, the question of *What*

1 *Nothing is more unfathomable that the system of motivations behind our actions*

next? arises. How to implement the psychological/ecological/behavioral need, present in humans, to explore behind the door? This section will thus present some of the major studies and experiments on simulated or real robotic platforms that take into account these notions of motivation – especially, *intrinsic motivation, curiosity, surprise, hunger*, and *occupancy*.

2.1 Biological considerations

This section lists the five brain structures and their neural substrates that are directly involved in the comprehension of the environment[2]. First, the anatomy and the functions of the hippocampus are detailed. Then, a focus on the neuronal cells responsible for *localization*, *navigation*, and *mapping* mechanisms is made – namely, the *place cells*, *transition cells*, *head-direction cells*, and *grid cells*. Even if these mechanisms are low-level, they rely on the continuous perceptual stream that comes mainly from visual and auditory sensors. But in which fashion is this continuous perceptual stream processed? Here, the *reverse-hierarchy theory* is an innovative approach, based on biological and psychophysic evidence, which enlightens the transverse mechanisms involved in the emergence of perceptual objects. These objects, as part of the environment, are essential to its proper comprehension. Finally, a most striking evidence of how inherently active the exploration process is, are

2
List of abbreviations used in this chapter

A 1 ... primary auditory cortex
AHV ... angular head-velocity cell
CA ... cornu ammonis
DG ... dentate gyrus
EC ... entorhinal cortex
GC ... grid cell
HD ... head-direction cell
HRTF ... head-related transfer function
HS ... hyppocampus
LV ... local view
MEC ... medial entorhinal cortex
NA ... nucleus accumbens
PC ... place cell
PF ... profrontal cortex
PI ... path intergration
PS ... product space
RHT ... reverse-hierarchy theory
S&R ... search and rescue
SLAM ... simultaneous localization and mapping
TC ... transition cell

the *head movements*. The importance and the mechanisms of these movements will be described from biological and psychoacoustics points of views – particularly, for the case of audition.

2.1.1 Neural substrates

Navigation and its planification need memorization abilities of the already explored environment and/or already taken paths, in combination with prediction of new possible paths, given data are already acquired from the environment and/or results of actions in/on the environment are available. The hippocampus (HS) – see Figs. 2.1 and 2.2 – together with its interactions with the entorhinal cortex (EC), prefrontal cortex (PF) and nucleus accumbens (NA) are the brain structures that mostly handle these processes – for reviews see Brown and Sharp (1995), Ragozzino *et al.* (1999), Lever *et al.* (2002), Nicola *et al.* (2004), Hok *et al.* (2005), Taha *et al.* (2007) and Cuperlier *et al.* (2007). In particular, the dentate gyrus (DG) is thought to be involved in the exploration of new environments (Saab *et al.*, 2009).

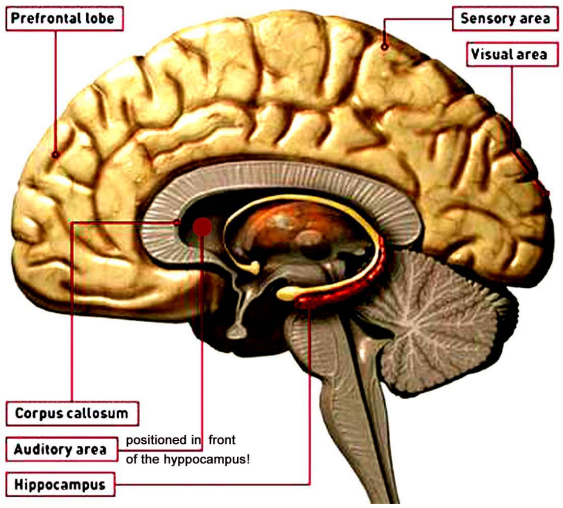


Figure 2.1: Hippocampus in human

Place cells

Redish (2001) has highlighted the existence of *place cells* (PCs) that fire only when the animal is in a particular place within the environment. Note that there is no link between spatial-environment topology and place-cells topology in the HS. Two PCs located next to each other other can code spatial information of two far places. Their

activity is also modulated by the cells that are sensible to head rotation – see Sec. 2.1.1. Electrophysiological studies showed that lesions of the HS lead to deficits in navigation. According to O'Keefe and Nadel (1978), Gaussier *et al.* (2002), McNaughton *et al.* (2006), HS generate a cognitive map – see Sec. 2.2.2 for examples of implemented cognitive maps in robots, that is, maps activated by the cortex.

In unknown environment, PCs are quickly recruited and are stable across time (Jeffery and Hayman, 2004). Surprisingly, PCs are independent of modality, are also active in dark environment (Markus *et al.*, 1994, Quirk *et al.*, 2008, Muller and Kubie, 1987) and are present even in blind people. For instance, smell can lead to the elaboration of cognitive spatial maps in the HS. One of the particularity of PCs is that one PC can fire in two distinct environments (Kubie and Ranck, 1983). Thus, this notion of location is not directly linked to the environment but rather on how it is perceived and conceptualized. Moreover, place cells are not exclusively excited by spatial cues. Hampson *et al.* (1993) showed that these cells can fire in response to salient events in a temporal sequence. Further, Young *et al.* (1994) and Wood *et al.* (1999) showed that stimuli such as texture or odor can induce firing of PC's.

Transition cells

The use of PCs only, even with a sensory-motor association, such as linking a PC and a movement, is insufficient in complex environments or when contradictory motivations arise, as several actions can be performed from the same place, that is, the same PC. This is why Cuperlier *et al.* (2006) used *transition cells* (TCs) as neurons to code the spatiotemporal transitions occuring between two PC's at time t in EC and time $t-1$ in DG. Since only one direction can be associated with a transition, TCs are more relevant and efficient to represent sensori-motor association. For instance, a TC would code the movement used to go from a place A to a place B, thus creating the TC AB.

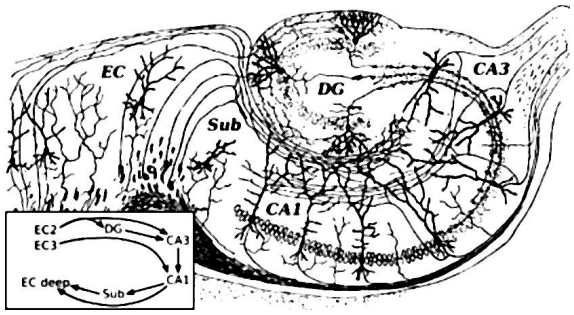


Figure 2.2: Hippocampus structure. EC ... entorhinal cortex, DG ... dentate gyrus, CA ... cornu ammonis, Sub ... subiculum

In this work, exploration of an unknown environment is formalized as a two-steps mechanism as follows.

(1) Planning periods, triggered by the motivational information the robot gets

(2) Exploration periods, triggered by a detection signal, generated while a new transition is learned. Once the robot is able to predict transitions from the current place, planning restarts

See Sec. 2.2.1 for the details of implemented algorithms and results of performed exploration tasks.

Head-direction cells

Head-direction cells (HD) are cells that fire with regard to a particular directional heading of an animal, independently from its location or ongoing behavior. These cells are located in several brain areas such as the presubiculum, the anterior dorsal thalamic nucleus, the lateral mammillary nuclei, the retrospenial cortex, the entorhinal cortex, the lateral dorsal thalamus, the dorsal striatum, and the medial precentral cortex. This widely spread presence of HD cells within several different brain areas shows that information on the direction of the head is required in several different processes. A striking result of Blair and Sharp (1995), Taube and Muller (1998) is that the peak firing rate is reached between 25 ms and 75 ms – depending on the location of the HD cells – *before* the head has reached the cell's optimal firing direction. One of the hypothesis to partially explain this prediction ability is that an efferent copy or corollary discharge is sent to the HD cells, thus allowing them to be aware of future motor commands.

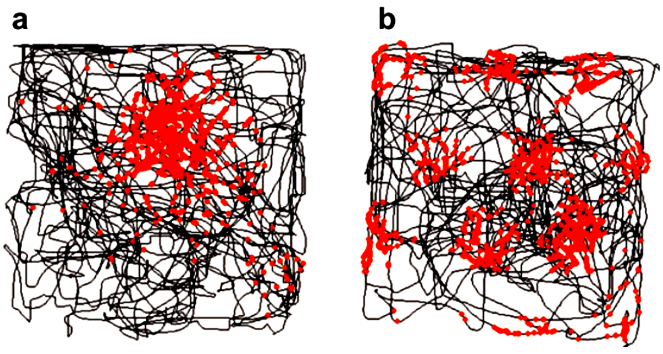


Figure 2.3: Place cell in the hippocampus, (a), and grid cell in the medial entorhinal cortex (MEC), (b). Spike locations (red) are superimposed on the animal's trajectory in the recording enclosure (black). Whereas most place cells have a single firing location, the firing fields of a grid cell form a periodic triangular matrix, tiling the entire environment available to the animal – from Moser *et al.* (2008)

Grid cells

Grid cells (GC) are placed in the EC just below HS – compare Fig. 2.3. Direct perforant path-projections go from CA1 to EC. GC's exhibit sharply tuned spatial firing with multiple firing fields. The fields of each neuron form a periodic triangular array, called *grid*, that covers the environment the animal has already explored (Hafting *et al.*, 2005) – Fig. 2.3. This particular organization is thus thought to be a possible metric system for spatial organization.

According to Moser *et al.* (2008), grids are characterized by

- *Spacing*, that is, the distance between fields
- *Orientation*, that is, the tilt relative to an external reference axis
- *Phase*, that is, the xy displacement relative to an external reference point

Grid cells are thought to be some kind of metric system for spatial navigation, since they collectively signal changes of position while exploring the environment (Hafting *et al.*, 2005). This metric system shares similar properties with the allocentric map of the hippocampus as proposed by O'Keefe and Nadel (1978).

Basal-ganglia/thalamus-cortex loop

Animals often have to face the possibility of doing two antinomic actions, such as drinking and eating. However, since the two actions of eating and drinking both use the oral system, it is basically impossible to execute them simultaneously. Thus, a choice has to be made between them. In human brains, basal ganglia are one of the structures responsible of action selection (Cools, 1980, Mink and Thach, 1993, Kropotov and Etlinger, 1999). This group of substructures, namely, striatum, pallidum and sub-thalamic nucleus – each of them including further substructures – enables actions by disinhibition of the target structures of actions (Chevalier *et al.*, 1985, Deniau and Chevalier, 1985). Prescott *et al.* (1999) and Redgrave *et al.* (1999) proposed a unifying hypothesis of action selection performed by the basal ganglia that relates anatomical and physiological studies. They also proposed a computation model of the basal ganglia – compare Gurney *et al.* (2001a).

Conclusion

The hippocampus and the entorhinal cortex play a very significant role in localization and navigation through dedicated, powerful and adaptive neural substructures – Fig. 2.4.

However, several other areas are involved in these processes also, such as the presubiculum, and the anterior thalamus for the head-direction cells. Further involved are the parietal cortex, for spatial representation and navigation, and the striatum for navigation. The complexity of all the connections existing in all these structures and substructures, coupled with highly specialized neurons that fire under very particular conditions that can be modulated by cognitive processes such as attention and/or goal-driven tasks, has led the work of the robotic community to primarily focus on the hippocampus and the entorhinal cortex – compare Sec. 2.2 for a brief review of bio-inspired implemented models and attempts to model the activity of PCs and GCs. Some studies try to extend these models by integrating the basal-ganglia structure and, particularly, its loop to thalamus and cortex.

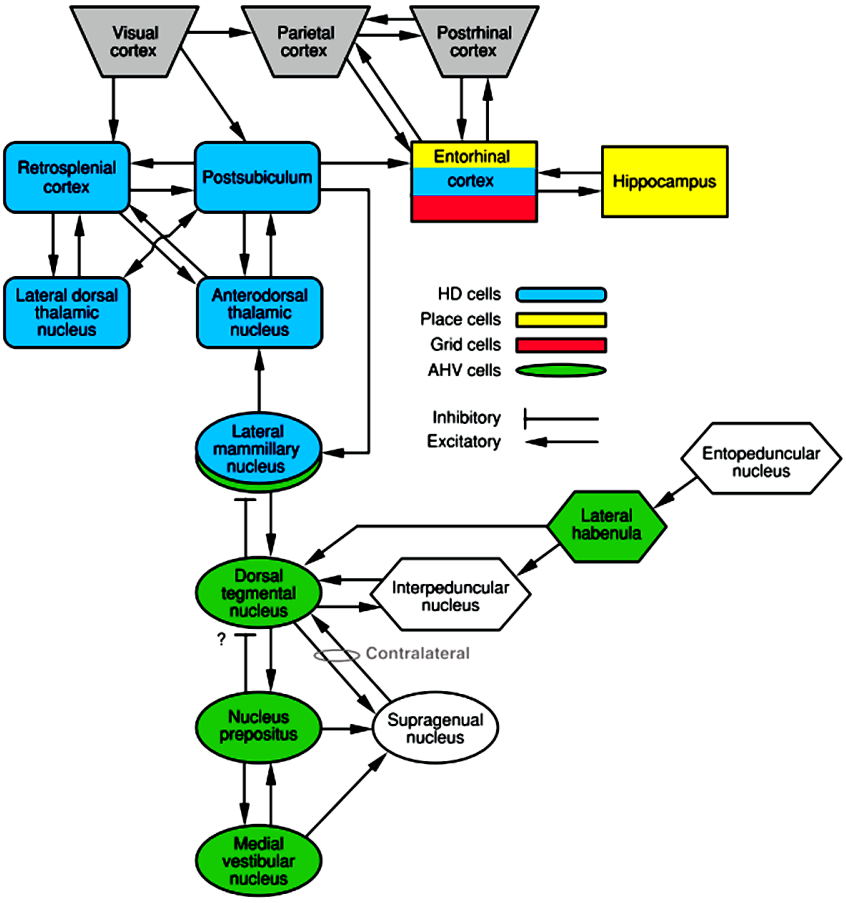


Figure 2.4: Circuit diagram showing principle connections of areas containing head-directions cells, place cells, grid cells, and angular head-velocity cells. Color key indicates the types of neuronal correlates identified for cells in that brain area – from Taube (2007)

2.1.2 The reverse-hierarchy theory

On the one hand we know how the brain codes localization, navigation and mapping. On the other hand, all these phenomena rely on the perception of the environment. Thus, the current section deals with how the perceptual stream, especially as originating from visual and auditory sensors, leads to the formation of the concepts of *objects* in the perceptual world.

The previous list of neural substrates for self-localization, navigation and mapping constitutes the physiological roots of exploratory strategies. However, the computations made by these cells rely on a previous step, namely, of object recognition. This holds for objects considered at a low-level stage, such as psychophysical attributes of a stimulus, or for objects at high-level stage, such as fully recognized multimodal objects. Exploratory strategies are thus highly dependent on the knowledge and the legibility of information that is present in the environment. In particular, discriminating two or more sources of information is a key step in adequate information processing and in the establishment of exploratory strategies. Recently, Ahissar and Hochstein (2004) and Nelken and Ahissar (2006) proposed an innovative theory regarding bottom-up and top-down communication between the sensors. These are, eyes, ears and the low-level structures that are placed just after them, on the one hand, and the higher-level computing areas, namely, the visual/auditory cortices, on the other hand – in ambiguous situations. The new theory, named *reverse-hierarchy theory* (RHT) – see Fig. 2.5 – is supported by neurological studies of Nelken. RHT postulates that a "parsing decision is first based on the highest available level of visual representation" (Shamma, 2008).

With respect to the visual system, this theory states that the speed at which information goes from low-level to high-level areas depends on the ability to discriminate the different sources of information. In a complex discrimination task, only the low-level features of the stimuli will enable the discrimination of two or more concurrent perceptive objects. Yet, when there is no ambiguity, the low-level features are not necessarily exploited to process and understand the incoming stimuli. Consequently, the relevant information proceeds faster to high-level areas. For instance, if one sees a glass fall, the noise of its breaking will be (i) highly predictable and (ii) easily recognizable. Thus, RHT postulates that the processing of such a sound will be fastened by skipping some of the low-level processing steps, for example, ILD and ITD. This evaluation can be skipped because the position of the glass can be predicted due to the availability of visual information. However, if the sound heard is not congruent with the prediction, the RHT postulates that low-level features will be recruited to solve the ambiguity. Feedback loops play a role in this context.

The RHT concept has been applied to the auditory system as well (Nelken and Ahissar, 2006, Nahum *et al.*, 2008) – see Fig. 2.6. Unlike many perceptive processes in the visual

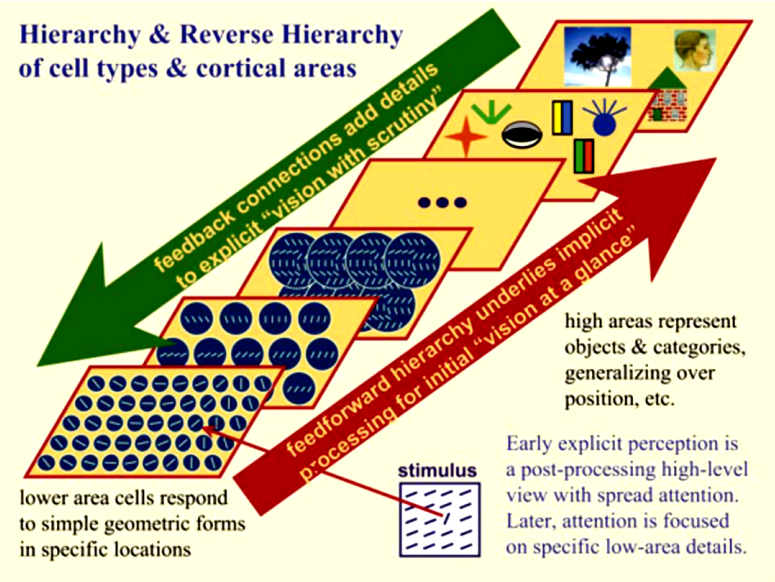


Figure 2.5: Reverse-hierarchy theory in the visual system (Ahissar and Hochstein, 2004, Nelken and Ahissar, 2006, Shamma, 2008)

system, the auditory objects are not considered as static objects but rather as continuous streams of information that generate dynamic representations of objects – compare Shamma (2008) for a short review. Each stream has thus its own perceptual identity. According to Shamma, "... the rules and interactions between the stream percepts and the low-level cues that group the elements of a stream and distinguish it from its counterparts, for example, pitch, timbre, and binaural cues, have been delineated over the years under the umbrella of auditory scene analysis " (Shamma, 2008). The *reverse-hierarchy theory* is thus very useful because this phenomenon is indeed a first step to assess attention processes by their ability to determine whether an information is ambiguous or not. Application of RHT can thus substantially fasten comprehension of the environment and lower the sensitivity regarding potentially irrelevant or non-informative stimuli.

This raises several questions about the nature of an object, among others the following.

- At which neural processing stage is an object recognized as such? Or, in other words, when does the concept of object arise in the perceptual stream?
- How many sensory cues are needed to form an internal representation of an object that is robust and reliable enough to make a decision?

In visual psychophysics, a conflict of sensory-information processing exists between the

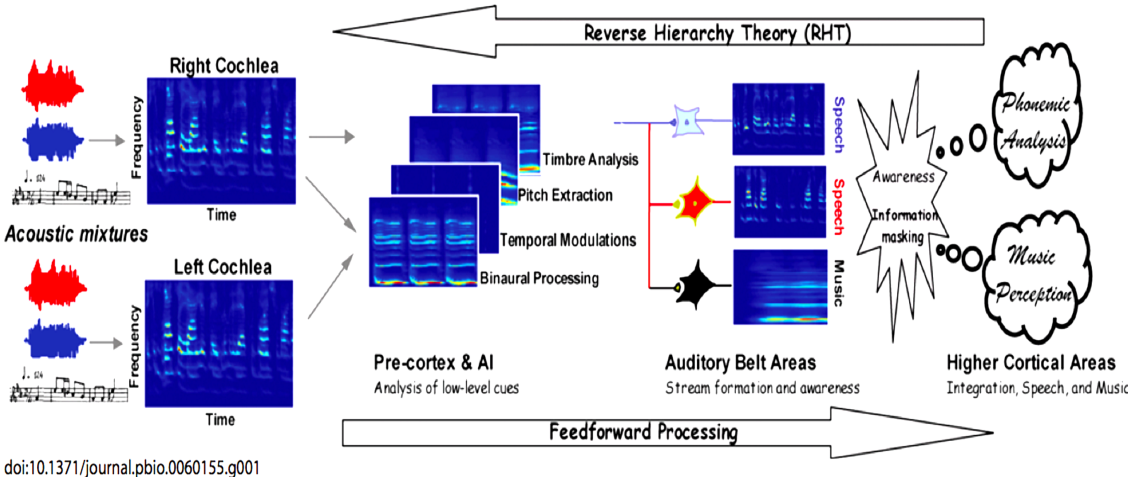


Figure 2.6: Schematic of bottom-up feed-forward flow of auditory analysis and top-down cognitive influences (RHT) that give rise to auditory perception and awareness. From left to right: Natural acoustic scenes usually contain mixtures of multiple speakers (red and blue signals) and music. Low-level cues embedded in the cochlear spectrograms from the right and left ears are analyzed and combined in several pre-cortical and primary auditory cortical (A1) stages. Neural correlates of consciously perceived streams of speech and music emerge in the auditory belt areas beyond A1. In complex realistic scenes, ambiguities as, for instance, "informationally masked" speech and musical streams are resolved through top-down influences as described by RHT – plot from Shamma (2008)

local and the global aspects of a stimuli. It is assumed that "the properties of the parts are determined by the laws of structure of the whole" (Westheimer, 1999). This notion of object is often thought as being an exclusively high-level notion. Indeed, the object emerge from the convergence of several different information coming from many different sensors and from memory – thus indicating the involvement of high-level areas such the associative cortex. However, there is evidence for the existence of *proto-objects*, especially for such audio objects that already emerge in low-level areas (Rodemann *et al.*, 2009) – see Fig. 2.7.

2.1.3 Head movements and listening

This section offers an overview of literature on the specificies of head movements and their benefits to listening. For a short summary of earlier relevant work see also (Blauert, 1974, 2nd ed.1997).

Young (1931) provided one of the earliest pieces of research to investigate the role of head movements on sound perception. He conducted listening experiments where a pair of "hard-rubber sound-receiving trumpets" as "artificial pinnae" were introduced and connected

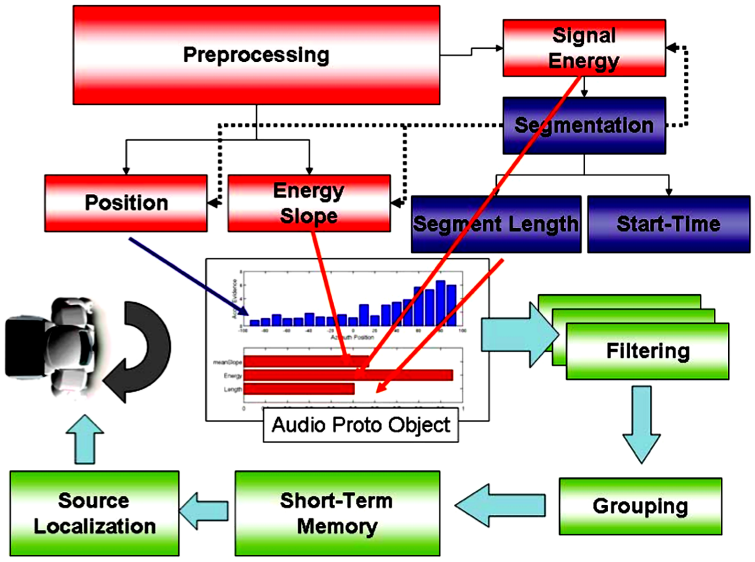


Figure 2.7: In audition, proto-objects arises early and speed up and enhance processing of auditory input (Rodemann *et al.*, 2009) – see Sec. 2.1.2

to the ears through rubber tubes and ear pieces. By this arrangement he tried to create a condition in which head and body movements could not provide any effect that might be considered as localization cues. When the signals were presented at various positions around the artificial pinnae, which were separated from the actual ears by a thick wall, the results of apparent source positions showed general inaccuracy in localization except for right-left discrimination, in particular, there was more chance of front-back confusion, and poor perception of elevation. Therefore he suggested that the lack of binaural cues, which could have been provided by head movement, resulted in a disability of the auditory system to discriminate between *up, down, front, and back*.

Wallach (1938) also conducted early studies in this field, investigating the function of head movements to extract more information from the cone of confusion. Here he defined the angular distance of the sound source from the axis of the ears as the *lateral angle* and argued that the changes in perceived lateral angles, as one moved the head, should provide the information needed to determine the source direction more precisely. For example, according to his hypotheses, the direction of change in the lateral angle of the auditory event should help to resolve front-back confusions, and the rate of change in the lateral angle with respect to the angular displacement of the ear axis, should give information regarding the elevation of the source.

To verify these hypotheses, he attempted to generate perceptual images of sound sources around the listeners by means of a switching loudspeaker array controlled by the listeners'

head movements. Specifically, the switch was set amongst the loudspeakers such that only a certain intended one operated in accordance with the head orientation. By arbitrarily controlling the angular distances of the loudspeakers to be switched for a certain head orientation, it was possible to make the listeners feel that the auditory event occurred at various elevation angles, even though the loudspeakers were arranged in front.

Further in-depth research supporting Wallach's theory includes the study of Thurlow *et al.* (1967). They tried to draw quantitative information on the nature and magnitude of rotational head movements through listening tests where the head movements of the listeners were recorded while they tried to locate the sound source. Ten loudspeakers were distributed at various horizontal and vertical angular positions around the listener in an anechoic chamber. Five of them were driven by low-pass-filtered noise, and the other five by high-pass-filtered noise. Here the types of head movements were grouped into the three categories, nameley, *rotation, pivot* and *tip*. By the term *rotation* they meant the rotation in azimuth only. *Pivoting* was defined as rolling the head to the left or right without changing the facing direction – as was referred to as *tilting* by Wallach above. *Tipping* indicated facing upwards or downwards, which is equivalent to nodding. Figure 2.8 describes the directions of these three movement types more clearly.

From the recorded results, they made a list of how frequently single or combined head movements happened. Some tendencies were found for both the high- and low-frequency signals. Namely, rotation was the most frequent of all single movement types among the three, and the combination of rotation and tipping was the most frequent of all – including single and combined movements. In general, combinations of movements including rotation happened more frequently than the others. Therefore they concluded that rotation plays the most significant role in sound-localizing "attempts". The accuracy of localization was not evaluated in their experiments. Additionally, it was found that the average maximal angle of rotation movement was larger than that of the others. Especially, the maximal rotation was found to be considerably greater for the low-frequency signals than for the high-frequency signals, which seems to imply the greater difficulty in localizing sound sources of low frequencies.

Further analyses of the directions of the movements showed a tendency of the listeners to make maximal rotation and tip movements toward the source directions. Reversals in the directions of the head movements, toward the starting positions, were also observed. This tendency was seen in all the three types of movements, but the rotation reversals were again the largest of all. This study was meaningful in that the patterns of general head movement were categorized and analyzed.

Another piece of research in the same year by Thurlow and Runge (1967) investigated the contribution of each head-movement pattern to the accuracy of source localization. Here experiments were carried out in a similar setting to the above in an anechoic chamber,

with slightly different number and distribution of the loudspeakers. However, in addition to low- and high-frequency noise, equivalently filtered pulses of short duration were also introduced as source signals. Moreover, the listeners were asked to localize the sources in different listening conditions, where head movement was "induced" by a guiding device. Four motion conditions, *rotate, pivot, rotate–pivot* and *tip* – see Fig. 2.8 – and, finally, free movement conditions were used, whose effects in terms of localization error were compared with those with no head movement allowed.

It was found that the horizontal localization errors were significantly reduced by the induced head movement including rotation, regardless of the signal type. Reduction in vertical localization errors, however, was not as significant. Additional discussions of the results concluded that free movement did not show such a great improvement in localization compared with simple induced rotation, and that for high-frequency noise head movement was not necessary for estimating the elevation. Although the experiments involved only a few fixed positions of sound sources and a few restricted movement conditions, this work established the guidelines for future research, especially by ascertaining the importance of rotation amongst all possible head movement types.

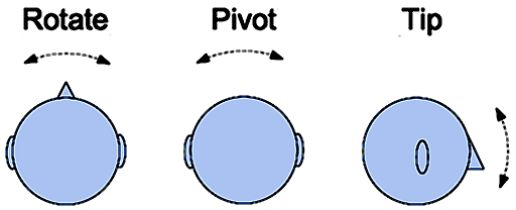


Figure 2.8: Three of the four motion conditions used in the Thurlow *et al.* (1967) study

More recently, Perrett and Noble (1997a,b) carried out experiments to test Wallach's ideas of the role of head movement in front-back discrimination and elevation detection. The loudspeakers were distributed on the horizontal plane, the lateral vertical plane, or the median vertical plane around the listeners. They were asked to locate the sound source in various conditions including *no motion, free movement* and *rotation only.* Broadband, low-pass and high-pass noise were used, and the case of "distorted pinna function" was introduced additionally, where the listeners wore open plastic tubes inserted into the ear canals to bypass the filtering effects of the pinnae.

It was observed that front-back confusion, which was prevalent in motionless conditions, was almost eliminated in rotation conditions. The elevation judgement was also more accurate with rotation allowed, especially for the sources in the upper hemisphere. However, for the high-pass signals above 2 kHz, the normal-motionless condition showed almost the same accuracy. Additionally, though the distorted-motionless condition gave the worst results in general, the distorted-rotation condition was not very effective for signals above 2 kHz. With the sources below the horizon the effect of rotation was not significant. From

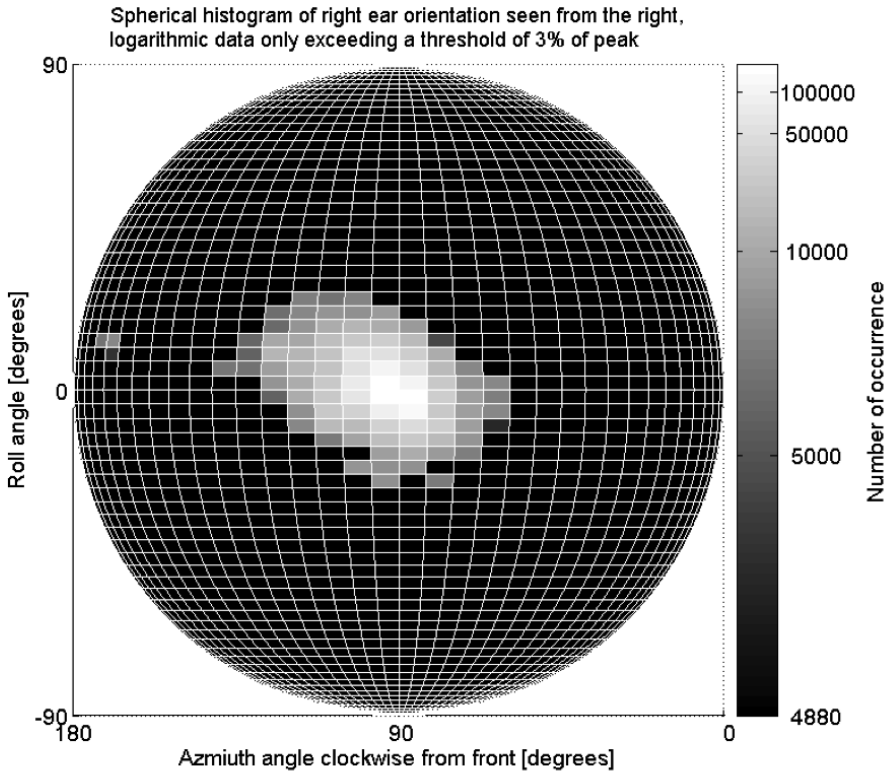


Figure 2.9: Spherical histogram of right-ear orientation seen from the right (Kim *et al.*, 2013)

these results the authors concluded that acoustic energy below 2 kHz was required for head rotation to be effective in vertical localization, and that the benefit from rotation was the greatest when the source was oriented in the frontal vertical plane. These gave support not only to Wallach's argument, but also to Thurlow and Runge's about the role of rotation in sound localization and its effectiveness with regard to various conditions such as source orientation and frequency characteristics.

Wightman and Kistler (1999), in succession, also carried out similar experiments to explore further Wallach's ideas. In this case, in addition to real sound sources located about the listener as in the previous experiments, virtual sources were introduced. The latter were produced by means of head tracking and individualized head-related transfer functions (HRTFs), and presented over headphones. White Gaussian noise was used as sound signal. First, the listeners were asked to localize the stationary sources in the three conditions, *no movement, free movement* and *orienting movement*, while attempting to face the apparent source position. Then the sources, both real and virtual, were set to move during the experiments, and the listeners were asked to locate the starting position without head movement.

It was observed from the results that front-back confusions were substantially reduced when free movement was possible, but that it was not as effective for elevation judgements. Both observations were in accordance with the findings of previous studies. In addition, the analysis of head-movement trajectories showed a tendency of the listeners to orient toward the expected source location when allowed to move freely. It was also suggested that the necessary information for localization could be extracted not only by the listener's movement but also by the source movement as long as the listener had control of the direction of the source movement.

Inpired by these previous findings, Kim *et al.* (2013) conducted investigations into the nature of head movements in various listening situations. Listeners were asked to judge various acoustical attributes of stimuli processed from anechoic recordings reproduced through multiple loudspeakers distributed in a listening room. The evaluated auditory attributes did not only include the location of the auditory event, but also the apparent source width, the sense of envelopment, and timbre. During the experiment, the head movements of the listeners were tracked and recorded.

It was found that the listeners moved their heads to larger extents while judging source width or envelopment than while judging source direction or timbre. In addition, it was found that the subjects faced toward the sources – in fact, not only for localization but also for source-width and envelopment judgments. The ear trace from the head-tracking data showed that the rotational head movement was confined mainly around the initial orientation, that is, facing directly forwards, and that the pattern of ear positions follows a "sloped path", which is higher towards the rear and lower towards the front – see Fig. 2.9.

Additional experiments were then conducted in more natural listening activities, including listening to a live concert, playing video games, and watching movies, to the end of testing whether the above findings are representative of actual listening behavior. The head-tracking data showed a distribution that was similar to, and fitted within the range of values found in the first experiment.

The key findings from these studies related to head movements in listening activities can be summarized as follows.

- Head movements do occur in usual listening activities, especially to a larger extent when spatial properties of auditory scene are being evaluated
- It is generally seen that rotational head movement helps source localization, and that amongst the three dimensions of rotation the rotation in azimuth occurs to of the the largest extent and is most helpful
- There is a tendency of listeners to face towards the source when evaluating spatial properties

2.1.4 Conclusion

The understanding of the biological structures and processes that are responsible for a comprehension of the environment is crucial in order to be able to engineer an efficient computational model as a basis of work. That is why the native brain structures – such as the hippocampus – and the particular organization of information processing, made by PC's, TC's, GC's and the basal ganglia-thalamus-cortex loop are very inspiring clues for autonomous robotic implementations in an active exploration paradigm, especially in unknown environments and for S&R scenarios.

Let us call to mind that active exploration strategies rely on two main processes,

- The available information in the environment
- The motivation that drives exploration

The two sections above about RHT and head movements in listening highlighted two major mechanisms involved in collecting information from unknown environments and consequent processing of the the continuous perceptual stream in an efficient way. But before addressing the question of motivation – as is discussed in Sec. 2.3 below – three major robotic exploration paradigms will be presented in the next section, directly inspired from the biological considerations previously described in this chapter.

2.2 Bio-inspired exploration models

Two main strategies are used by the animals to navigate in an environment, namely,

- *Landmark navigation*, where the animal infers its position and orientation by detecting surrounding landmarks, and thus using allothetic cues acquired by its sensory modalities – such as vision, audition, olfaction and touch
- *Path integration*, where the animal knows its starting position and orientation and thereafter estimates them by using idiothetic cues – that is, internal information, such as motor efferent copy, proprioceptive and vestibular information

Yet, exploring unknown environments for robots involves management of three different tasks, namely *mapping*, *localization* and *motion control* (Makarenko *et al.*, 2002) – Fig. 2.10.

Various techniques and algorithms have been created to accomplish efficient environment exploration. They can be grouped into two general categories, differing by the goal they try to reach. The two categories are

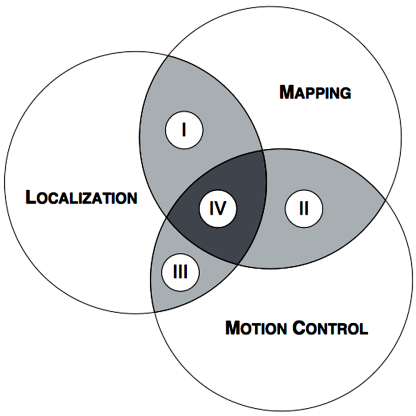


Figure 2.10: Three tasks that a robot should accomplish when exploring unknown environments efficiently. The highlighted regions of integration in robotics are *SLAM*, (I), classic exploration, (II), active localization, (III), and integrated exploration, (IV) – from Makarenko *et al.* (2002)

- *Quick exploration*, that is, techniques that aim at minimizing the time needed to explore the whole environment and, thus, to compute the motion commands that will optimally reach this goal
- *Highest amount of information*, that is, techniques that aim at reducing the uncertainty of the environment by choosing the next observing point, thus maximizing new incoming information

Most of the computational models make an intensive use of landmarks. These parts of the environment, such as objects, persons, sceneries, and so on, can be used by the robot to create an internal map and to locate itself. They are, for example, used in order to solve the loop-closing problem[3].

In S&R scenarios, the robot has to dynamically and actively create its own map of an unknown environment while being able to localize itself in it – computed thanks to *simultaneous-localization-and-mapping algorithms* (SLAM) . Thus, the robot has to solve two main problems, that is, (i) the selection of a new target – in other words, making a choice between searching for victims or trying to extend the map and, (ii) the choice of a path to reach the goal, which means to define possible paths and selecting one of them that will probably fulfill several conditions that are under strong constraints, such as danger and time (Wirth and Pellenz, 2007).

3 The loop-closing problem concerns the ability of a robot to recognize a place in the environment where it has already been before (Stachniss *et al.*, 2004)

The following subsections will list non-exhaustively some of the major studies on navigation and planning in robotic platforms for exploration tasks in both known an unknown environments. All these models are bio-inspired. In order to stay concise, only equations of interest have been noted.

2.2.1 The transition-maps models of *Cuperlier*

Cuperlier *et al.* (2007) and Cuperlier *et al.* (2006) developed a bio-inspired model based on the hippocampus and the prefrontal cortex. In their model, spatiotemporal transitions are explicitly coded by *transition cells* (TC's) thus creating a sensori-motor association between a place in the environment and a movement.

In this model, the *What* and *Where*[4] provided by the visual system are merged into a matrix of neurons called a *product space* (PS) – see Fig. 2.11. The *What* is composed by landmark units through perirhinal cortex neurons and the *Where* is composed by azimuth information through parahippocampal neurons. The transition-map model uses a neural network inspired by the entorhinal cortex (EC) to compute the learning of activity patterns on PS.

The following paragraphs shortly describes some of the mathematical formalization of Cuperlier's model. The variable X^A will denote the activity, X, of the neuron, A. W^{A-B} will denote the weight of the link between neurons A and B. Let's recall that DG is the dentate gyrus, EC is the entorhinal cortex, CA is the cornus ammoni.

PC's are neurons that code for locations in the PS. Given that all the PC's are interconnected with each other, with weights assigned to each connection, the activity of the jth PC results from the computation between the current local view and the learned view writes as

$$X_j^{EC_s}(t) = \frac{1}{W_j}\left(\sum_{kl}^{N_{PS}} W_{j,kl}^{PS-EC_S} \cdot X_{kl}^{PS}(t)\right), \tag{2.1}$$

where $W_j = \sum_{kl}^{N_{PS}} W_{j,kl}^{PS-EC_s}$, and $W_{j,kl}$ is the weight of the link from pixel k, l to the jth PC. N_{PS} is the number of cells in the *product space* (PS), which is a direct function of the number of landmarks and azimuth cells – see Fig. 2.11 and Fig. 2.12.

Thus, if the robot is at the exact position where the PC has been learned, its activity will

4 The *What* and *Where* system is a proposed organization of the visual system which claims that there are two different streams processing visual information, namely, a dorsal stream – going to parietal cortex – and a ventral stream – going to temporal cortex. The former is involved in object recognition, that is the *What*, whereas the latter one is involved in spatial vision, that is the *Where*.

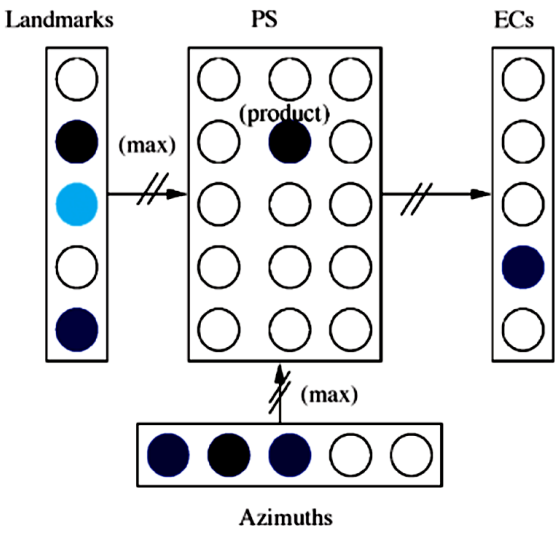


Figure 2.11: The *product space* is the resulting matrix of neurons that code the *What* and *Where* systems. Whereas several neurons of the product space can be activated when the observed pattern matches the previously learned pattern, it is only the winner (in blue) that is shown on EC neurons (Cuperlier *et al.*, 2007)

be maximal. If the robot moves away from this PC, the activity of this PC will slowly decrease. The activity of the neurons of DG, that is, the neurons that store the previous location, is computed as follows.

$$X_i^{DG}(t) = X_i^{ECs}(t-1) \tag{2.2}$$

Transition cells are the cells that integrate informations from DG neurons and EC neurons, thus forming a two-dimensional matrix of CA neurons. The activity of the neuron (i, j) of CA comes out as

$$X_{ij}^{CA}(t) = [\sum_{k=1}^{N}(W_{ij,k}^{DG-CA} \cdot X_k^{DG}(t)) + W_{ij,i}^{EC_s-CA} \cdot X_i^{EC_s}(t) - \theta]^+ , \tag{2.3}$$

where θ is a threshold for weights and $[]^+$ is upper bound of a decimal..

The learning equation that allows increasing weights between DG and CA, thus enabling the prediction of all transitions based on the current location, computes as

$$W_{ij,i}^{DG-CA} = \begin{cases} \frac{X_i^{DG}(t)}{\sum_k^{N^{DG}}(X_k^{DG}(t))} & \text{after learning} \\ \text{small random value inferior to } \theta & \text{before learning} \end{cases} \tag{2.4}$$

Figure 2.13 shows the model including PC's, TC's, PS, the cognitive maps, the motor transitions and the motor commands resulting from all the previous computations.

2.2.2 The ANIMAT approach

ANIMATS are artifical animals, be they simulated or physical robots. The term comes from the contraction of animal–materials. Meyer (1996) formalized the ANIMAT approach motivated by the wish to implement a robotic system that mimics the behavior of animals when placed in unpredictable and potentially dangerous environment. Indeed, animals seem to be able to explore their environment in a highly robust and apparently simple way. Moreover, they are able to adapt easily to a new environment and are quite at ease in an unknown environment. Nowadays, exploration robots are mostly based on (i) supervised learning, which leads to good performances in learned environment but to a poor one in unknown environment and, (ii) on unsupervised learning, which requires severe constraints and strong *a priori* definitions about what could be the environment.

The way the animals act when being confronted with such situations is thus very interesting to the end of understanding how exploratory movements can be used in concrete and severly constrained situations. Two major projects using the ANIMAT paradigm will be presented in the following, namely, the RATSLAM approach and the PSIKHARPAX project.

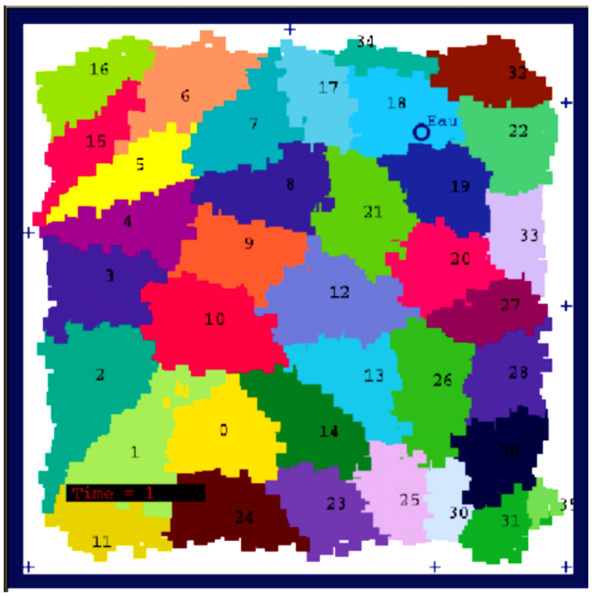


Figure 2.12: In the simulated environment of Cuperlier *et al.* (2007) each coloured region represents the field of a place cell. The crosses are landmarks

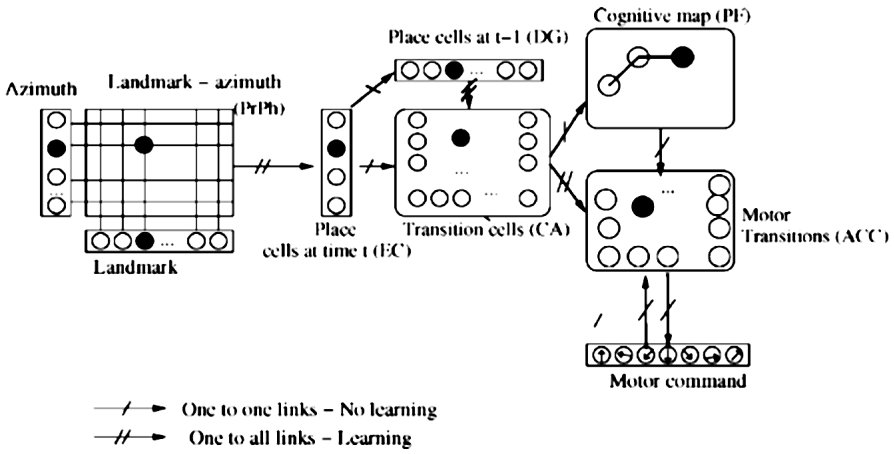


Figure 2.13: The model of Cuperlier *et al.* (2007), aiming at modelling the interactions between the hippocampus and the prefrontal cortex in navigation and planning strategies

These two recent models are highly innovative and are both inspired by the functioning of the rat, since rats as rodents in general, show very good abilities to explore unknown environment.

RATSLAM

Milford *et al.* (2004) have proposed a new approach to the SLAM paradigm, called RatSlam and being directly inspired by computational models of the rodent hippocampus. Whereas a lot of models compute the head direction, θ, and its position, x, y, with two distinct attractor networks (Arleo and Gerstner, 2000), the idea of RatSlam is to implement the eigen-pose of the robot, that is, its own location and orientation as one competitive attractor network represented as x, y, θ. Indeed, one of the main drawback of two-networks models is the inability to compute multiple beliefs in pose for any period of time. Pose cells do not necessarily map on to the Cartesian space that they encode. In fact, one place can be coded by multiple pose cells – a *discontinuity*. Further, one pose cells can code for multiple places – a *collision*. The external sense of vision, providing allothetic information, brings the local view (LV), whereas the internal sensing provides idiothetic information and enables path integration (PI). LV & PI together thus constitute *pose cells* (P) – Fig. 2.14. On the basis of both LV & PI inputs, activity packets will then be generated in the pose network. These packets will inhibit the pose cells that are far from them while exciting the ones that are close to them. The winning packet is computed in order to identify the pose estimate with the highest probability – compare Fig. 2.15.

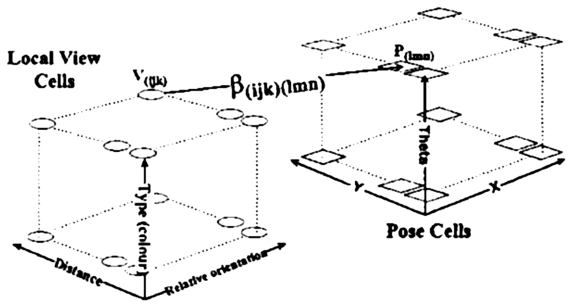


Figure 2.14: Illustration of the local-view network and the pose-cell network. Units in the local view become associated with units in the pose cells via learnt weighted connections of the two networks (Milford *et al.*, 2004)

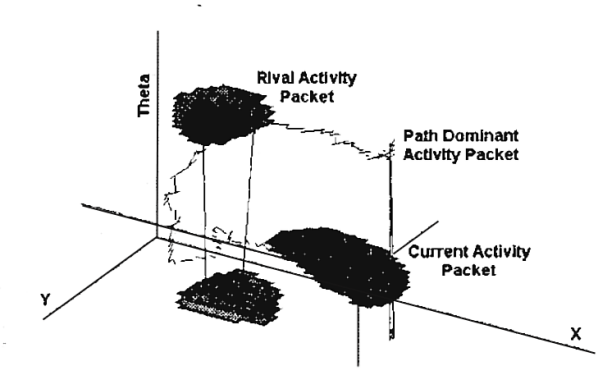


Figure 2.15: Snapshot of pose cell activity during an experiment. Note that the current activity packet is smeared, indicating that it is moving. The rival activity packet will not win unless it receives reinforcement from further visual input – from Milford *et al.* (2004)

Pose-cell coding is an innovative representation of the environment and of the robot itself in it. It captures the benefits of grid-based, topological and landmark-based representations. This allows the robot to not being rigidly constrained to the classical Cartesian grid.

PSIKHARPAX

The PsiKharpax project has been initiated by Meyer *et al.* (2005) and aims at designing biomimietic sensors and neural control architectures that will provide the robot with the capabilitiy of autonomy and adaptation. PsiKharpax is equipped with the following allothetic sensors: two eyes, two ears (cochleas) and sixty-four vibrisses. Further, as idiothetic sensors, it has a vestibular system computing linear and angular accelerations of

the head, an odometry system monitoring the length and direction of the displacements, and an energy monitor – Fig. 2.16. Moreover, low-level reflexes are implemented such as "keep looking to an object even when its head is moving", or "avoid an obstacle detected by the whiskers and/or by its visual or auditory systems".

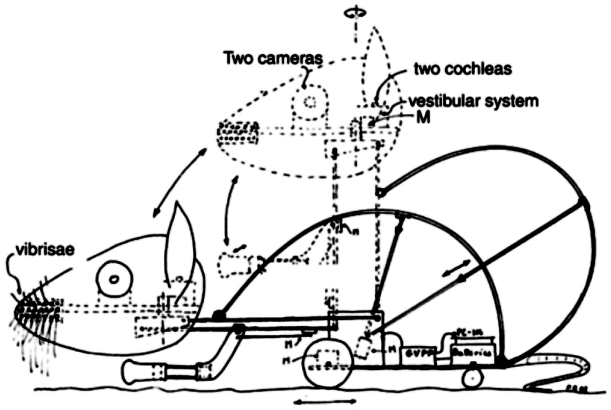


Figure 2.16: Design of PsiKharpax (Meyer *et al.*, 2005)

The model of navigation of the PsiKharpax is based on a multiple-hypothesis tracking strategy, all hypothesis being updated in parallel and thus providing a dense topological map. The actions of the PsiKharpax are computed on the basis of the *Gurney–Prescott–Redgrave* model (GPR) (Gurney *et al.*, 2001a,b). GPR aims at modeling the activity of the nuclei of the basal ganglia, based on the *basal-ganglia–thalamus–cortex loop* – Sec. 2.1.1. It assumes that each discrete motor action is coded in segregated channels in the nuclei of basal ganglia. These channels are inhibited by default. The inputs to these channels are called *saliences*. They consider both internal and external perceptions when evaluating whether each action is relevant regarding the robot's needs – see Fig. 2.17. Moreover, a positive feedback loop with the thalamus introduces persistence to these assessments. In the end, that action will be selected which has been the least inhibited one.

Two additional loops have been implemented in PsiKharpax, namely, a *ventral loop* that selects locomotor action and a *dorsal loop* that selects non locomotor actions, both also being modeled by a GPR system. The interconnections between these two loops prevents the robot from doing a locomotor action and a non locomotor action at the same time. In fact, the dorsal loop sends some excitatory inputs to the ventral loop, thus raising the inhibition level of all the locomotor actions.

Different directional profiles have been set up, nameley, a *planning profile*, a *homing profile*, and an *exploration profile* – Fig. 2.18. Each profile is determined by the position in the environment where the robot wants to go. If the robot is motivated to go to two broad

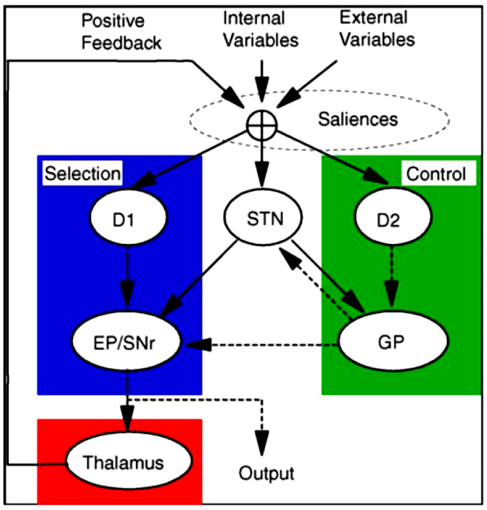


Figure 2.17: A single channel within the basal ganglia in the GPR model. Solid arrows represent excitatory connections, dotted arrows inhibitory ones – from Meyer *et al.* (2005)

directions corresponding to two sources recorded in its map, then the planning profile is activated. If it is motivated to return to already explored regions of the environment, thus lowering the disorientation variable of these regions, the homing profile is activated. Finally, if the robot is motivated to go to unexplored regions, it pursues the exploration profile.

According to Meyer *et al.* (2005), PsiKharpax is characterised by

> "Being able to integrate the past (through its recorded map), the present (through its sensors) and the future (through its planning capacities), will represent an embodied example of a motivationally autonomous animat whose control complexity may well challenge the possibilities of external control and, hence, its capacities to withstand any imposed autonomy"

2.2.3 Conclusion

All the models described in this section use different concepts that have been used to design robots that are able to efficiently explore the environment, especially when still unknown to them. These models rely on biological considerations about exploration and self-localization. This holds in particular for the hippocampus, as this is the major brain structure that allows animals to perform exploration tasks – compare the place cells of the RatSlam approach, the transitions cells in the Cuperlier's model, and the Basal-Ganglia–Thalamus–Cortex loop of the PsiKharpax model.

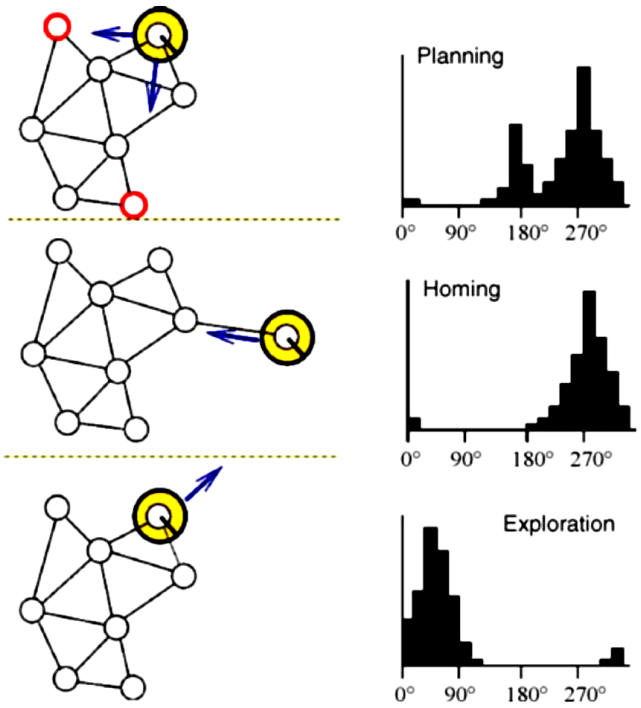


Figure 2.18: Three direction profiles (right) that call upon the current map of the environment (left) (Meyer *et al.*, 2005)

Going back to the two main processes of active exploration as discussed in Sec. 2.1.4, the following two issues are critical.

- The available information in the environment
- The motivation that drives exploration

Now that the first issue has been described and illustrated by examples from robotic implementation, the question of motivation will be addressed. The next section will describe what *motivation* is (a) from a psychophysics point of view and (b) in an active-exploration-in-robots context. From the work of Berlyne in the mid-sixties to the latest implementations of Oudeyer during the last years, this section will show examples of how to mathematically formalize high-level concepts such as *intrinsic motivation* or *curiosity*.

2.3 Motivation for exploration

From decades, *motivation* was considered as an essential mechanism that very well explain spontaneous exploratory behavior in humans, particularly in infants (Berlyne, 1965). Exploration is influenced by different features of the incoming stimuli, such as intensity, color, pitch, and also by the biological notion of reward vs. punishment. However, to be sure, further high-level parameters are determinant in exploration, for instance, *novelty*, *change*, *surprise*, *curiosity*, *incongruency*, *ambiguity* or *indistinctibility*. Since the last twenty years, several kind of motivations have been conceptualized to the point that they have been implemented in several simulated or real robotic platforms. Motivations can be different from what is rendered from classical goal generators, such as the wish of complete coverage of an unknown environment or finding the closest best observation point. Motivation is just above that, namely, a goal generator that will bring to the robot a notion of *reward*. This notion of reward will intrinsically be satisfying for the robot.

Ryan and Deci (2000), based on a definition by Berlyne (1950), described intrinsic motivation as " ... the doing of an activity for its inherent satisfaction rather than for some separable consequence." Thus, intrinsic motivation can be defined as a self-satisfying activity, whereas extrinsic motivation is the motivation to accomplish a task for an external outcome. Recently, the work in computational modelling of robots and robotic implementation of exploratory robots, as put forward by Oudeyer and Baranes, have been guided by this notion of intrinsic motivation (Oudeyer and Kaplan, 2008, Baranes and Oudeyer, 2009, 2010). Exploratory activity guided by intrinsic motivation is not homeostatic, that means that the desire of exploring an environment is not caused by a single case of need, for instance, reducing the biological perturbation caused by the appearance of a new stimulus, but rather is a desire constant over time. This desire seems to be guided by very different kind of needs that are function of intrinsic goals for the organism. From Ryan's definition of intrinsic motivation, Oudeyer proposes a computational definition claiming that

> "An experienced situation ... is intrinsically motivating for an autonomous entity if its interest depends primarily on the collation or comparison of information from different stimuli and independently of their semantics, whether they be physical or imaginary stimuli – that is, measured by physical sensors or by internal (software) sensors perceived in the present or in the past ... "

Here, information is understood from an information-theoretic perspective, that is, by the intrinsic mathematical structure of the stimuli, independently of its meaning. Consequently, a system that takes intrinsic motivation into account should integrate a mechanism for evaluating how a situation evokes a state of *surprise, complexity, challenge* and/or *novelty* in the robot and, further, for measuring an associated reward. Maximising these

measures can lead to the elaboration of suitable autonomous and active-exploratory processes.

Intrinsic motivation regroups several kinds of motivation, such as the *uncertainty motivation*, the *information-gain motivation*, or the *empowerment motivation*.

- *Uncertainty motivation* is defined as the attraction for novel stimuli. Thus, for every observed event, $e^k \in E$, in the ensemble E, a reward, $r(e^k)$, will be generated that is inversely proportional to its probability, $P(e^k, t)$, of observation at time t. Huang and Weng (2002)) formalized this as follow.

$$r_{UM}(e^k, t) = C \cdot (1 - P(e^k, t)), \tag{2.5}$$

  where C is a constant.

- *Information gain motivation* can be defined as the "pleasure of learning" and guides robots to minimize the level of uncertainty of their knowledge of the environment (Roy *et al.*, 2001), and be formalized as

$$r_{IGM}(e^k, t) = C \cdot (H(E, t) - H(E, t + 1)), \tag{2.6}$$

  where $H(E) = -\sum_{e^k \in E} P(e^k) ln(P(e^k))$ denotes the entropy characterizing the shape of the distribution function for discretized spaces.

- *Empowerment motivation* leads to a behavior that encourages the acquisition of the maximal amount of information by the sensors of the robot. Thus, the robot will try to find the sequence of actions that produces the maximal flow of information (Capdepuy *et al.*, 2007), formalized as follows.

$$\begin{aligned} &r_{EM}(A_t, A_{t+1}, \ldots, A_{t+n+1} \rightarrow S_{t+n}) = \\ &= \max_{p(\overrightarrow{a})} I(A_t, A_{t+1}, \ldots, A_{t+n+1}, S_{t+n}) \,, \end{aligned} \tag{2.7}$$

  where $p(\overrightarrow{a})$ is the probability distribution function of the action sequences,nameley,

$$\overrightarrow{a} = (a_t, a_{t+1}, \ldots, a_{t+n+1}) \,,$$

  and I is the mutual information, that is, the information shared by the different actions $\overrightarrow{a}$.

The following subsection describes some selected novel approaches for generating goals for successfully and efficient exploration of unknown environment.

2.3.1 The SAGG-RIAC algorithm

Baranes and Oudeyer (2010) proposed a *self-adaptive–goal-generating–robust–intelligent–adaptive–curiosity algorithm* (SAGG-RIAC). This algorithm represents an enhanced version of the RIAC algorithm (Baranes and Oudeyer, 2009), which in turn, is an evolution of IAC algorithm (Barto *et al.*, 2004). SAGG-RIAC has been developed in a *competence- based active-motor-learning framework* (Oudeyer and Kaplan, 2008). It is working on two different levels and acting at different time scales as follows – compare Fig. 2.20.

(a) At a lower time scale that "considers the goal-directed active choice and active exploration of lower-level actions to be taken to reach the goals selected at the higher level, and depending on local measures about the evolution of the quality of learnt inverse and/or forward models"

(b) At a higher time scale that "considers the active self-generation and self-selection of goals, depending on a feedback defined using the level of achievement of previously generated goal" (Oudeyer and Kaplan, 2008)

The conception of the goal-directed exploration and learning mechanism (lower time scale) includes an inverse and/or forward model, generated during the exploration and available for a later reuse. Further, a learning feedback that drives the choice of new actions in the active exploration task. The goal-self-generation and goal-self-selection process (higher time scale) is based on the "competence improvement in given subregions of the space where goals are chosen" (Baranes and Oudeyer, 2010). This notion of competence forms the intrinsic motivation of the robot. This is indeed what makes this approach innovative and interesting.

Competence is measured by the $\gamma_{s'_g}$ criterion, whereby, for a given goal-reaching attempt, competence is defined as follows.

$$\gamma_{s'_g} = \begin{cases} \min_C & \text{if } C(s'_g, s'_f, \rho) \leq min_C \\ C(s'_g, s'_f, \rho) & \text{if } \min_C < C(s'_g, s'_f, \rho) \leq \epsilon_C < 0 \\ 0 & \text{otherwise}, \end{cases} \tag{2.8}$$

with s'_f being the state reached when the goal-reaching attempt has terminated, s'_g is the actual goal of this reaching attempt, and ρ is some constraint. C is the cost function measuring these conditions[5] ϵ_C is a tolerance factor and $\min_C$ is a limiting factor representing the minimum competence value. A $\gamma_{s'_g}$ close to 0 thus indicates that the

5 This cost function is always negative (Baranes and Oudeyer, 2009) such that the lower $C(s'_g, s'_f, \rho)$ comes out, the more inefficient a reaching attempt will be considered as

system is competent to reach the goal, s'_g, in the light of the constraints, ρ. Feedback based on this notion of competence is dependent on the monitoring of the progress of local competences[6]. The goal-self-generation and goal-self-selection process proceeds in two steps.

(1) Split the space, S', where goals are chosen into subspaces, according to heuristics that allows to maximally distinguish areas according to their levels of interests

(2) Select the subspaces where future goals will be chosen from

Goals are chosen into a mix of three different modes, that is,

Mode (1) A random goal is chosen inside a region selected according to its interest value[7]

Mode (2) A random goal inside the whole space is chosen

Mode (3) The algorithm performs a random experiment inside the region where the mean competence is the lowest

Each mode can be selected with a given probability. Typically, *Mode (1)* is chosen with 70% , *Mode (2)* with 20% and *Mode (3)* with 10% probability. The global pseudo-code as generated by this heuristic is depicted in Fig. 2.19.

2.3.2 Curiosity, surprise and hunger

Among the intrinsic kinds of motivation that lead to the generation of an exploratory strategy, *curiosity*, *surprise* and *hunger* are thought to be of primary relevance. *Curiosity* can be defined as the desire of acquiring information on new objects or objects that have uncertain features that seem of interest (Berlyne, 1950) – the interest being intrinsic or extrinsic. This lack of information on *a priori* objects of interest stimulates an exploratory will. Considering *surprise* as a *motivational motor* is slightly different from considering it as an *emotion*. The distinction has to be made between a strategy that aims to search for perceptive events that will cause the *surprise feeling* on the one hand, and the biological reaction provoked by a new and/or incongruent and/or unknown stimulus, on the other one. At last, *hunger* is simply the need to find an energy source.

Schmidhuber (1991) has implemented an artificial agent with the notion of *curiosity*. The computational model aims at provoking situations for which "... it learned to expect to learn something about the environment"(Schmidhuber, 1991). The idea is to learn

6 See Baranes and Oudeyer (2010) for the mathematical formalization

7 A formal definition for this model can be found in Baranes and Oudeyer (2010)

Algorithm 1 Pseudo-Code of the SAGG-RIAC Algorithm

input: **M**: empty model of the robot; ρ: constraints;
input: g_{Max} : maximal number of elements of a region
input: thresholds: ε_C; ε_{max}; *timeout*
input: rest position (s_{rest}, s'_{rest}); arm reset value: r
input: starting position (s_{start}, s'_{start})
loop
  $(s_{start}, s'_{start}) = (s_{rest}, s'_{rest})$ every r reaching attempts
  High Level of Active Learning : Part 1
  Goal Self-Generation
  Selection of a region R_n and a goal s'_g using the $mode(m)$, with probability p_m
  Reaching Attempt: Low Level of Active Learning:
  Let (s_c, s'_c) represent the current configuration of the system
  while $C(s'_g, s'_c, \rho) \leq \varepsilon_C$ & *timeout* not exceeded **do**
    Reaching Phase:
    Compute a desired $\Delta s'_i$ that minimizes $|C(s'_g, s'_c + \Delta s'_i, \rho)|$
    Perform the action a_i computed using $\mathbf{M}^{-1}$, and given the desired $\Delta s'_i$
    Get the resulting performed $\widetilde{\Delta s'_i}$ and update **M** with the element $(s_c, s'_c, a_i, \widetilde{\Delta s'_i})$
    if $\|\widetilde{\Delta s'_i} - \Delta s'_i\| > \varepsilon_{max}$ **then**
      Local Exploration Phase (e.g. as in SSA)
    end if
  end while
  High Level of Active Learning : Part 2
  Interest Update:
  Compute the competence $\gamma_{s'_g}$
  Update R_n by adding s'_g and competence $\gamma_{s'_g}$ inside
  Update the interest value $interest(R_n)$
  Split R_n if $|R_n| > g_{max}$
end loop

Figure 2.19: Pseudo-code for the SAGG-RIAC heuristic – from Baranes and Oudeyer (2010), see Sec. 2.3.1

estimating the effects of further learning. Given an adaptive discrete time predictor, M, the model will rely on two main modules, that is, (i) *adaptive confidence* and, (ii) *adaptive curiosity*. The confidence module (C) aims to determine how confident the agent can be in the environment inputs. Given a noisy learning world, M will still make some errors. Thus, C will decrease the error rate by generating an output that provides information about how reliable M's predictions can be expected to be.

In their work of the last fifteen years, Macedo and Cardoso have particularly studied the role and the importance of these kinds of motivation in robotic systems for the exploration of unknown environment (Macedo, 2004, Macedo and Cardoso, 2004, 2005). Macedo and Cardoso (2005) have implemented a motivation module, taking into account the three variables *curiosity, surprise* and *hunger* – Fig. 2.21. This module generate a weighting of the potential goals that the system can try to reach. These goals are then computed by the deliberative/decision-making module.

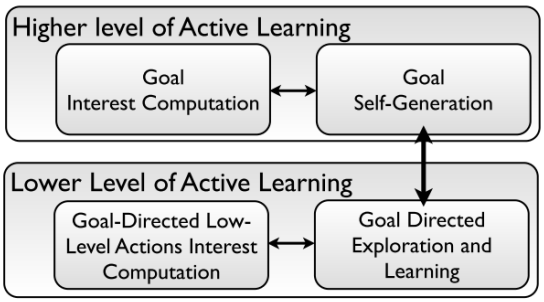


Figure 2.20: Global Architecture of the SAGG-RIAC algorithm. The structure is composed of two parts defining two levels of active learning: a higher which considers the active self-generation and self-selection of goals, and a lower, which considers the goal-directed active choice and active exploration of lower-level actions, to reach the goals selected at the higher level – from Baranes and Oudeyer (2010), compare Sec. 2.3.1

Let E_g be an event, with $g \in \{1, 2, ..., m\}$ among a set of m mutually exclusive events, $E = \{E_1, E_2, ..., E_m\}$, and let E_h, $h \in \{1, 2, ..., m\}$ be the event with highest probability of the set, E. Let S be *surprise*. *Surprise* elicited by an event *after* it has occurred is computed as follow.

$$S(E_g) = \log_2 .(1 + P(E_h) - P(E_g)) \tag{2.9}$$

According to this equation, there is always at least one, E_h, the event with the maximum probability, $P(E_h)$, which is entirely unsurprising. In order to predict *beforehand* the surprise felt by the agent from a scenario, s, the following equation is used.

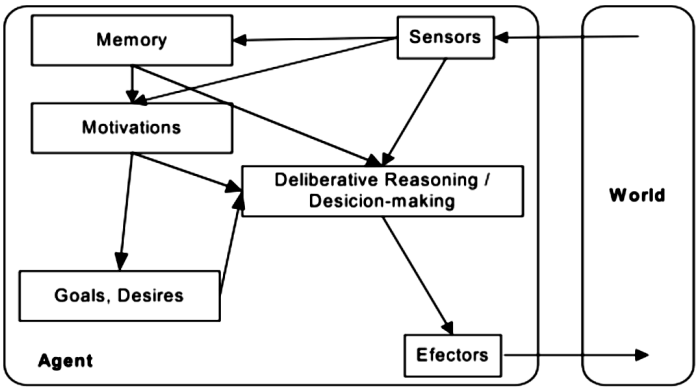


Figure 2.21: Agent's architecture of Macedo and Cardoso (2005). This agent has a motivation module that takes into account different components of motivation, such as *surprise*, *curiosity* and/or *hunger*. A deliberative module computes a weighted motivation provided by the motivation module and generates a new goal

$$E[S(s)] = \sum_{i=1}^{n} P(E_i) \times log_2(1 + P(E_h) - P(E_i)) \tag{2.10}$$

Surprise is also directly linked to *unexpectedness* (Macedo, 2004) and *uncertainty*. Thus, an object is considered as consisting of the following different components, namely, "... the cells of the analogical description, the propositions of the propositional description, and the function" (Macedo and Cardoso, 2005). Each of these pieces is considered as a possible scenario and can be uncertain or not. To compute the amount of surprise, both the pieces of an object, X, with no uncertainty, X_C, and the pieces with uncertainty, X_U, are used. The amount of surprise is then computed as follows.

$$\begin{aligned} S(X) = S(X_C) + E[S(X_U)] = \\ = \sum_{E_g \in X_C} \log_2 .(1 + P(E_h) - P(E_g)) + \\ + \sum_{S \in X_U} \sum_{i=1}^{m_S} P(E_i) \times log_2(1 + P(E_h) - P(E_i)) \end{aligned} \tag{2.11}$$

Curiosity is basically the attraction caused by an unknown object and the desire to learn what this object is (Berlyne, 1950). Curiosity is also directly linked to *novelty* or *uncertainty*. While novelty implies that new information is present, uncertainty represents information that is probably to be acquired. As successful aquisition of information leads to a decrease in uncertainty. The theory of information then stipulates that it can be computed thanks to the entropy[8]. Let C be the *curiosity*, N the *novelty*, U the *uncertainty*, and H the entropy of an object. Curiosity is then determined as

$$C(X) = N(X) + U(X) = \ = \min_k \left(\sum_{i=0}^{|S|} HD(X_i, AgtMem_{ki}) \right) + H(X)\,, \tag{2.12}$$

where HD is the Hamming Distance, $HD(X_i, AgtMem_{ki}) = 1$ in case that the ith component of X, X_i matches $AgtMem_{ki}$, that is, the ith component of the kth object in the memory of the agent.

As for surprise, the object is considered as composed of of components. The uncertain components are used to compute novelty, N, while the components with no uncertainty are used to compute the entropy, H, of the object. Novelty means to be unknown to

8 After Shannon (1948), entropy, H, of a discrete random variable, X, with n elements. called *the source*, is defined as $H(X) = -\sum_{i=1}^{n} P_i log(P_i)$, with the probability P_i of the ith element to occur

the agent. It is thus necessary to have access to the memories – $AgtMem$ in Eq. 2.12 – of the already known objects by the agent. A comparison between every object in the memories of the agent and the perceived object is made. The propositional and analogical descriptions are graph-based. A superposition of the objects in memory and the perceived object leads to a match function by counting the minimal number of changes of nodes and edges to transform one graph into the other. The entropy term, H, of Eq. 2.12 is computed as

$$\begin{aligned} H(X) &= H(X_A) + H(X_P) + H(X_F) = \\ &= \sum_{i=1}^{m} p^i \log_2(\frac{1}{p^i}) + \sum_{z=1}^{l} \sum_{j=1}^{r_Z} p_j^z \times \log_2(\frac{1}{p_j^z}) + \\ &\quad + \sum_{K=1}^{n} p_k \times \log_2(\frac{1}{p_k}) + \\ &\quad + (1 - p^i) \log_2(\frac{1}{1 - p^i}) \,, \end{aligned} \tag{2.13}$$

where X_A and X_P are the analogical and propositional description of the physical structure of the object, X and X_F, and its function, respectively. Three simulated versions of environment, populated with entities, have been used to test the exploratory ability of the agent, given a single kind of motivation, that is, a mixture of *surprise, curiosity* and *hunger*. At an average of 60% of the entities were similar to each other.

2.3.3 The occupancy grids of *Wirth*

Wirth and Pellenz (2007) have developed an algorithm based on *occupancy grids* (Elfes, 1989) to determine the next interesting "frontier" between *known* and *unknown* parts of the environment (Yamauchi, 1997). The model uses the *path transform* (Zelinsky, 1988, 1991). This is an extention of the *distance transform* (Jarvis and C., 1986) and includes an *obstacle transform*, Ω, in order to compute the cost to reach the target cell, c_g, of the occupancy grid while avoiding obstacles – see Fig. 2.22.

Let $\chi_c^{c_g}$ be the set of all possible paths from c to c_g, and let $l(C)$ be the length of the path, C. Further, let $c_{danger}(c_i)$ be the cost function for the discomfort of entering cell, c_i, – see below – and α be a positive weighting factor determining how far the path is free to stay away from obstacles. The path transform, Φ, of a cell, c, to reach the target, c_g, is then defined as

$$\Phi(c, c_g) = \min_{C \in \chi_c^{c_g}} \left(l(C) + \alpha \sum_{c_i \in C} c_{danger}(c_i) \right) \tag{2.14}$$

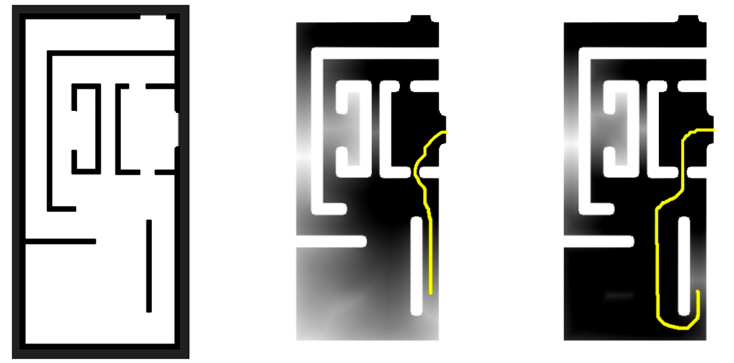


Figure 2.22: Occupancy grids, (a), and results of the exploration transform for different values of α, (b) and (c)

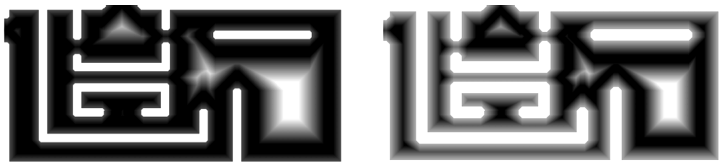


Figure 2.23: Discomfort costs for different values of d_{min} and d_{opt}. (a) $d_{min} = 3, d_{opt} = 20$. The robot will be allowed to use narrow pathways that are potentially damaging for it but allow to detect and use landmarks. (b) $d_{min} = 5, d_{opt} = 30$. The robot will not be allowed to use narrow pathways but will stay also away from potential landmarks (Wirth and Pellenz, 2007)

Discomfort costs provide information about the quality of a path – Fig. 2.23. Indeed, if a path is following a very narrow corridor, such that the robot could damage itself, consequently, this path must not be chosen. The model developed here uses coastal navigation[9] to explore the environment. Thus, a compromise has to be found between

- The safety of the path, by staying at a distance, d_{opt}, from the obstacles
- The ability to find landmarks that are close enough, that is, at a distance, d_{min}

The *discomfort-cost function*, c_{danger}, is defined as noted below, inspired by a definition of

9 Coastal navigation is a kind of robotic navigation in which the robot follows the wall of the room that is being explored. This kind of navigation is often used in indoor navigation but shows limits in outdoor environments

Zelinsky (1994) and adapted by the authors.

$$c_{danger}(c_i) = \begin{cases} \infty, & \text{if } d < d_{min} \\ (d_{opt} - d)^2, & \text{else}\,, \end{cases} \tag{2.15}$$

where d_{min} is a constant that determines the minimum distance to obstacles and mainly depends on the size of the robot. The steepest gradient is used to go to the target cell. Wirth et al. have extended the path transform definition – see Eq. 2.14 – in order to compute the cost of a path going to a close frontier rather than going to a target cell. Let F be the set of all possible frontiers. With the same variables and constants definitions of Eq. 2.14, the exploration transform of Wirth is defined by

$$\Psi(c) = \min_{c_g \in F} \left(\min_{C \in \chi_c^{c_g}} \left(l(C) + \alpha \sum_{c_i \in C} c_{danger}(c_i) \right) \right) . \tag{2.16}$$

Finally, the model includes a *navigation grid* that enables the robot to detect potential victims – a thermal sensor, ThS, has been used here with a field of view of 180° and a two-meter range. This navigation grid is defined as

$$\text{navGrid}(c_i) = \begin{cases} \text{free}, & \text{if occGrid}(c_i) = \text{free} \bigwedge ThS(c_i) \\ \text{occupied}, & \text{if occGrid}(c_i) = \text{occupied} \\ \text{unknown}, & \text{else}\,, \end{cases} \tag{2.17}$$

where *occGrid* is the occupancy grid and $free$ is the free space in the scanned environment. Compare (Wirth and Pellenz, 2007) for further details about the mathematical formalization. This navigation grid allows the robot to compute the areas that the thermal sensor has scanned and that have been mapped by the laser scanner. Thus, this grid is used for the exploration transform in S&R scenarios. Moreover, the navigation grid can be adapted to any multi-sensor data.

2.3.4 Conclusion

The results of the Macedo & Cardoso experiments on a simulated robot – Sec. 2.3.2 and Fig. 2.24 – show that *hunger* is a most powerful motivational motor to perform an exhaustive and fast exploration of an unknown environment. However, in a scenario where a time constraint is imposed, a strategy taking into account both *hunger* and *curiosity* is better. *Surprise* seems to be of less importance for efficient exploration. However, in the case of S&R scenarios, hunger can not be considered as a primary motivation to explore. Indeed, at the beginning of a rescue task, the exploration of the environment is far more important than the basic energy needs of the robot. Nevertheless, the question of the

energy needs can be raised for rescue missions that last for hours or even for days. The robot has also to be aware of its energy level and to know autonomously how to deal with it. This can involve hunger as a new short-term motivation to find energy sources in order to not to be forced to go back and lose precious time.

Among all that, curiosity seems to be a primary and efficient motivation for exploration. As concerns surprise, even though it is not a motivation that is sufficient and efficient enough for an exhaustive exploration of an unknown environment, it is obvious that surprise, as a biological reaction to an unknown/unpredictable event, is one of the premises of curiosity. Indeed, this kind of motivation only exists when one faces a new object/situation/scenario, and that this object/situation/scenario/event is of interest. In the case where this motivation would be gainful for the execution of the current task, this interest is no more curiosity. It is, indeed, an extrinsic and/or extern motivation, what means that it will of advantage for an external agent. On the other hand, when the interest is gainful only for the agent, it is an intrinsic motivation, devoid of any external goal.

The work of Oudeyer and Kaplan (2008), Baranes and Oudeyer (2010) on intrinsic motivations, based on psychophysics studies by Berlyne (1950, 1965), shows how important these notions of motivation are. Many different paradigms have been developed by the robotic community, such as *empowerement motivation*, *information-gain motivation*, or *uncertainty motivation* These motivations are a key step in order to build autonomous robots that really understand the environment they are exploring.

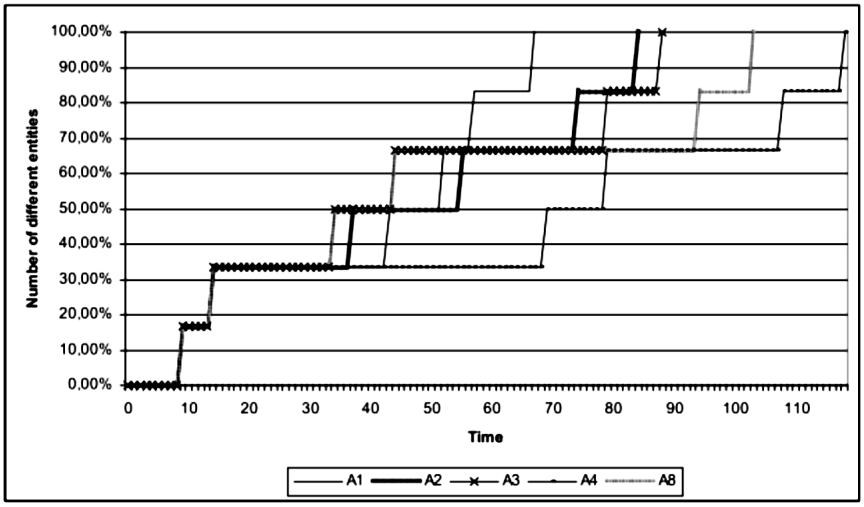


Figure 2.24: Results of the simulated agent of Macedo and Cardoso (2005), taking into account three components of motivation for the exploration of an unknown environment, namely, *hunger, curiosity* and *surprise*

2.4 Conclusions in view of the goals of the current project

This chapter covered in a transverse manner the *active-exploration paradigm* in robotics, from neural substrates and different brain structures as the biological roots to self-localization, navigation and environment mapping, but also from the psychophysics of perceptual streams, both visual and auditory, to the creative and innovative computational models as have been developed in the last three decades.

Moreover, this chapter gives concrete clues on how important feedbacks are in an active-exploration paradigms. Section 2.1 describes all the links that exists between perception – considered as perceptual streams. see Sec. 2.1.2 – and self-localization, navigation and mapping, see Sec. 2.1.1. *Action and perception* are two mechanisms that are strongly entangled, since one cannot exist without the other. Thus, continuous communication between sensors and localization processors is the key for valid, efficient and adaptive modeling of the environment. Section 2.2 provides examples of robotic implementations of these biological considerations for active-exploration tasks. The algorithms presented in this section show the inportance of incessant adaptation of the links that exist between the different modeled brain structures, such as dentate gyrus, entorhinal cortex, and cornus ammoni.

The models are not stationary but evolve constantly according to changes in the environment. This is what makes these models powerful and relevant. Finally, Sect. 2.3 provides clues on the processing stage that is, on the one hand, just above the active exploration algorithms – in terms of complexity and on their intrinsically high-level nature – and, on the other hand, just below them – since these algorithms rely on motivation, such as by *curiosity*. This motivation is essential to build autonomous robots with more than just exploratory abilities. Indeed, the need to have a higher-level need which drives exploration and acts as a continuous feedback between perception and exploration, is a key step in order to build a bridge between low-level abilities of exploration and the robot's high-level consciousness of its environment.

Most of the existing studies/models/algorithms rely exclusively on visual information so far. Auditory processing has been less studied by biologists and thus, by the roboticians and engineers that aim at being bio-inspired. In exploring their environment, humans rely primarily on vision. This is a fact proven by numerous electrophysiological, psychophysical and behavioral studies. But instead of considering audition as a parallel sense to vision, one should consider it as a complementary sense, often working in the background and providing information that either confirms or infirms visual information. Two main advantages of audition are, (i) it has a 360° field of listening, and (ii) auditory features of a multimodal object may be perceived while the other features can not – for instance, due to an obstacle, because of the distance or in darkness or fog. Further, humans communicate primarily via the auditory channel, that is, via speech. All this makes audition a major

source of information.

However, one of the main difficultiesin the processing of acoustic objects is (i) the dependance of the hearing conditions, that is, the characteristics of environment and listeners and, (ii) the intrinsically dynamical characteristics of these objects. Acoustic features are dependent on the characteristics of the environment, such as indoor/outdoor, reverberance, obstacles, and these are constantly changing over time. For instance, an accelerating car is still visually a car but the sound of the motor is changing at high rates and depends on several variables, such as the driver, the age of the car, the quality of the road, the distance from the car, etc. Thus, inferring robust auditory objects in real-time and without a long and multiply-constrained supervised learning phase is an important challenge for the acoustic, robotic and engineering communities.. Yet, is no longer bio-inspired and shows its limits as soon as the environment and the scenarios become more and more complex.

Within the Two!Ears project, audition is considered as a major modality in active-exploration tasks but also in the relevant comprehension of the environment. Taking into account all the biological phenomena coupled with some of the algorithms and models described in this contribution is essential to be able to provide innovative robotic models that succeed in integrating one of the major phenomenan in animal perception. Multimodality, for instance, is a field that deserves intensive attention in this context. Concretely, according to the bibliographical review in this literature survey, we have identified a number axes of research that will specifically be tackled in the Two!Ears framework. The following list explicitly denotes some more important ones.

- The *reverse-hierarchy theory*, which claims that a feedback loop exists in the continuous perceptual stream between low- and high-levels of perceptual information processing – see Sec. 2.1.2)
- The *transition cells*, which are triggered by vision but by audition as well – see Sec. 2.1.1 and 2.2). In the model of Cuperlier, only vision is taken into account so far
- *Motivation*, with components such as *uncertainty motivation* or *curiosity*, having the potency of providing our system with a substantial gain in exploration speed and efficiency due to its ability to adapt a strategy of adaptive motivation – see Sec. 2.3.1 and 2.3.2.

3 Attention-driven feedback

3.1 General remarks

Attention is a fundamental issue of sensory and cognitive psychology and concerns any attempt to model the perception, recognition and interpretation of objects that form our environment. Common definitions of the phenomenon of attention read as follows.

- "...the act of carefully thinking about, listening to, or watching someone or something. Also: the act or state of applying the mind to something. And also: a condition of readiness for such attention involving especially a selective narrowing or focusing of consciousness and receptivity" (Merriam-Webster Dictionary, accessed 2014.05.21)
- "...the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things" (Wikipedia, accessed 2013.11.02)
- "...is that state of consciousness where the contents of it show particular clearness and, in terms of their sequence, regularity and order" – translated from an older German encyclopedia (Brockhaus Konversationslexikon, 1869)

But note also the following, slightly different definitions.

- "...notice taken of someone or something. Also: the regarding of someone or something as interesting or important" – Oxford Dictionary (accessed 2014.05.21)

From a philosophical point of view attention remains to be a challenging issue. A famous, though rather general early view on it stems from John Locke (Locke, 1689). He characterized attention simply as a particular "mode of thought". Further brilliant minds, such as René Descartes and George Berkeley, have dealt with the phenomenon as well, but it is only since the 20^{th} century, that attention is conceived as concentration process in information processing. For a concise overview of relevant philosophical viewpoints and theories see the Stanford Encyclopedia of Philosopy (accessed 2014.05.21). Attention is often seen in close relationship with *consciousness*, but this aspect is obviously less relevant in the context of the Two!Ears project. Thus, we shall omit it here from further discussion.

By the way, the notion of attention being evoked by "causes" that actively change perceptual and cognitive functions from otherwise non-attending states to attending ones, is epistomologically imprecise. What is actually observed is that biological organisms, while being alive, enter into states that can be described as more or less attending. This is simply the biological course of life. Its description does not require the assumption of any additional mysterious "causes".

For Two!Ears, the two most relevant aspect of attention are

(a) The term attention denotes states of a perceptual/cognitive systems in which the performance with regard to given tasks is enhanced. For our specific model system this task is, in general terms, "Read the World with Two!Ears"

(b) Attention requires feedback from "higher" to "lower" computational stages of the system

The necessity of top-down control to enable attention, that is, *feedback*, has been pointed out by numerous authors lately, for example, by Reynolds and Desimone (2000), Bundesen and Habekost (2008), Beck and Kastner (2009), Engelke *et al.* (2011). However, there are obviously no substantial implementations of this principle on technological auditory-analysis and -recognition systems so far, with the expeption of modern hearing instruments, some of which automatically adapt their processing "programs" to the current environmental situation – yet, not in the sense of attending to specific auditory objects, since perception is not part of the functionality of hearing instrument. However, this may actually be so in the not-too-far future.

In vision, research on attention has been pursued more actively than in audition. However, many of the findings from there can be generalized to audition as well. Thereby, in terms of technological applicability, there is wide agreement that the most interesting point about attention-related processes is that they *reduce the complexity of scene analysis* by focusing on the most relevant features and/or object(s) at any particular point in time – compare Engelke *et al.* (2011).

3.2 Concepts and findings relevant for engineering models of listening

Most of the empirical research on attention has been performed under the auspices of psychology. Overviews can be found in Nobre and Kastner (2014), Carrasco (2011), Jones and Yee (1993) and in the chapters by Spence and Santangelo (2010) and Hafter *et al.* (2008) – the latter two dealing particularly with auditory attention.

In psychology, a number of conceptual models exist of how attention may come about. Amongst those are the *early-selection model* for instance, Broadbent (1958), which assumes that the selection of the auditory stream to be concentrated on happens at the lower (more peripheral) stages of the auditory system. In contrast, the *late-selection model*, see Deutsch and Deutsch (1963), supposes that the selection takes place in the "higher" (cognitive) stages. There are models which form a mixture of the two, compare, for example, Johnston and McCann (2006). More recent models take notice of empirical findings as to which concurrent auditory streams are not completely switched off but rather processed in parallel, and will only finally be attenuated by assigning appropriate weights to them – Murray (1970), compare also the *work-load model* of Lavie (2005). This assumption makes sense in view of the massive parallel processing that takes place in the brain with its more than 18 billion synapses! As to the question of how the weights are assigned, a number of further hypotheses have been put forward, such as the *binding model* or *feature-integration model* – see Treisman and Gelade (1980), Treisman (2003), and the *coherence model* – Hirst *et al.* (1980). Summerfield and Egner (2003) propose a *Bayesian approach.*

When looking at these conceptual models in terms of technological applicability, they turn out to be of limited use. The reason is that in technological modeling one has to mimic the functions of attention processes by means of current software technology and implement them on currently available hardware. Unfortunately, the full potency of human brains is thus not available. Nevertheless, engineering models are still useful, even for basic research, as they enable the testing of psychological hypotheses for functional feasibility.

Attention relies on cues, that is, on information provided to the system to be used in the adjustments to specific states as are biologically adequate in given situations. In this sense, the specific nature of the cues is a characteristic determinant of any attentional activity.

The following distinction can be made between two different types of attention,

- *Reflexive attention*... is triggered by primitive perceptual cues without cognitive processing, that is, in a reflexive way. The cues that trigger this kind of attention have been called *exogenous.* They provoke a fast attraction to objects and/or their locations within a comparatively short reaction time. This happens regardless of a specific task. Also, reflexive attention does not need and can most likely not be modified by training. It has been supposed that exogenous cues cannot be ignored. The cues that trigger reflexive attention decay within less than a second and then, consequently, become ineffective. Some authors argue that this kind of attention is performed by sole bottom-up processing with no feedback involved (Engelke *et al.*, 2011), but even reflexes may contain spontaneous feedback, such as the *stapedius reflex*, which modifies the middle-ear transfer characteristics triggered by loud sounds.

- *Reflective attention...* requires reflection in the mind, that is, it puts a cognitive load on the modeling system. The attention will then be guided with the goal of supporting a given task. The cues that trigger such a behaviour have been termed *endogenous*. It comes without surprise that this kind of attention needs longer to react. Also the cues that initiate reflective attention may be memorized almost infinitely and remain effective as well. This long-time effectiveness of endogenous cues can be understood by taking into account that attention may not only focus on sensory objects, but also on emotions – feelings – and thoughts – ideas, concepts.

As a rule, the reaction times for reflexive attention are shorter than those for reflective attention (Spence and Driver, 1994). Yet, it should to be stressed at this point that the categorization of attention processes into reflexive and reflective ones is only an operational one – one category mainly calling for short-term, the other one more for long-term memory. Also, the two types are certainly not completely independent as has, for example, been demonstrated by Santangelo and Spence (2007). The following provides more details.

3.2.1 Reflexive attention

This kind of attention is most likely to occur when sudden changes in the environment happen, such as surprising spatial movements of objects or sudden appearance of unknown sounds, changes in loudness, frequency, timbre, etc. In vision comparable salient features are sudden changes in brightness, colour, motion, orientation and size, among others (Wolfe and Horowitz, 2004). In the context of TWO!EARS, the *turn-to reflex* as well as the *olivocochlear reflex* – see Chapts. 2 and 4 of this document – are candidates for predominantly but not exclusively reflexive attention. The *precedence effect* – for reviews see Blauert (1974, 2nd ed.1997), Blauert and Braasch (2005), and *dereverberation* – see Tsilfidis *et al.* (2013) – also contain reflexive components besides reflective ones.

In many cases reflexive attention just consists in rescanning the environment, that is, resetting the current processes and start them anew, since obviously something has popped up in the environment – what calls for immediate reorientation. In these cases reflexive supersedes reflective attention. Exogenous cues that trigger reflexive attention do not necessarily point to the location in the perceptual world from which they originate and, in these cases, are not informative and thus invalid. Invalid cues may slow down reaction in further trials (Posner and Cohen, 1984). They may even leave the listener unoriented for a while, due to an internal inhibition to attend to the cued perceptual location. [1] Itti and Baldi (2009) proposed a saliency-based attention model and later considered *Bayesian surprise* as a trigger of reflexive attention.

1 Not attending means to be unfocused and, consequently, unoriented!

3.2.2 Reflective attention

Reflective attention is planned and goal directed. It can be modified by training and always involves top-down processing, namely, feedback from the cognitive stages of the system. Consequently, many cognitive eventualities may have to be considered, such as emotions, current actions, case history, domain knowledge, social state. The following briefly reviews selected aspects of this kind of attention as may be relevant for the Two!Ears.

Attention in detection tasks with sinusoids or very-narrow-band signals Signal detection, in general, concerns the ability to extract signals from a background. Most explanations regarding this problem take into account that auditory system performs a spectral decomposition of the incoming ear signals. The band-pass filters used by the auditory system for this purpose are called *auditory filters* or *critical bands* (Fletcher, 1940, Zwicker, 1961), for a review see Moore (1997). Listeners can detect narrow-band signals easier when they fall into only one of these bands. Detection is improved substantially when the listeners have been informed by a test signal of what to be expected, that is, are enabled to adapt to the task (Greene, 1960, Scharf, 1970). However, detection performance decreases when the listeners are presented with signals at unexpected frequency regions and, thus, adaptation is obstructed (Green and Swets, 1966). Uncertainty can also be introduced with the same effect by modifying the test-signal durations randomly (Wright and Dai, 1994).

Attention in detection tasks with harmonic complexes It has often been assumed – particularly in the context of the *olivocochlear reflex*, compare Chap. 4 – that uncertainty regarding the expected test signals may widen the bandwidth of the ear filters. Data derived by Schlauch and Hafter (1991) with complex tones of different complexity support this idea, but the actual effect is smaller than expected. Yet, the listeners may also respond to signal components outside the specific auditory filter on which the system is currently concentrating. However, although these components may be detected, they may nevertheless not been taken notice of in the further course of processing (Dai *et al.*, 1991, Scharf *et al.*, 1987).

However, this may change with the experimental paradigm. In effect, including the output of further auditory bands can be conceived as a widening the spectral range that the system is screening. The interesting question then is the following: At which processing stage is the decision taken as to what to include for the further processing besides the output of the most prominent auditory filter. Hafter *et al.* (2008) hypothesize that this decision will be taken when a salient feature becomes detectable, such as a virtual pitch. Attention, then, no longer requires focusing on primitive features. More complex or

even abstract ones, which test-signal components evoke in common, may also serve as triggers.

Attention with regard to music signals Attention does not only aim at signal detection, but largely any thinkable feature of sounds can be its subject – no matter whether the sounds are partly masked or clearly audible. However, attentional effort tends to increase with substantial noise being superimposed.

Musical signals have characteristic complex structures, both in the temporal and spectral domains. Cues that attract attention to specific attributes of such signals are consequently also spectrally and/or temporally specific. It seems obvious that such cues can guide attention to particular features of musical sound, such as rhythm, pitch, loudness, timbre, or even more abstracts features like the melody, the sound quality, or even the emotions communicated. Actually, when listeners are assigned the task of judging on *sound quality*, they may pay attention to various different features, depending on their actual assessment tasks. When judging on primitive psycho-acoustic features like loudness, roughness or pitch, they take on an analytic (discretic) listening mode, where they try to selectively identify individual primitive features and disregard auditory objects as entities. However, when judging on the quality of auditory scenes, they go into a holistic (syncretic) listening mode, where the auditory object as entities and the relations between them are of prominent interest. These two modes of of listening are also observed with regard to attention, see, for instance, van Noorden (1975, 1977). Jones and Yee (1993) distinguish between *integrative* and *selective attending* in this regard. Whereas, when the auditory events function as signs to communicate meaning and aural-communication quality is the topic, the assessors attend to the meanings that are communicated. The quality of sounds as information carrier is then what counts, and not how the sounds actually "sound" (Jekosch, 2005, Raake and Blauert, 2013, Blauert and Jekosch, 2012).

Music has specific internal structures like a syntax, and meanings are assigned to its elements. One has to learn the "language" of a musical style to appreciate the related music. Thus, attending to music has very much in common with listening to spoken language – compare next paragraph. A simple way of initiating attention to musical segments is modification of spectral or temporal features, such as the timbre of instruments, the musical key, or the rhythm. Yet, to understand and model the kind of attention that seizes the mind of listeners in excellent musical performances is still beyond the reach of current engineering models. Interestingly, listeners, when fully concentrating on one feature, may completely blind out others, such as the actual musical piece while concentrating on the acoustics of the concert hall. This effect is know as *attentional amnesia* (Wolfe, 1999, Gregg and Samuel, 2008, Shinn-Cunningham, 2008).

Attention with regard speech signals Spoken language is the prominent medium of human inter-individual communication. Accordingly, speech signals are of principal interest in the context of auditory attention, particularly, regarding *selective attention*. Selective attention to speech is a specific focused state of the sensory system and mind that results in improved understandability of speech – in engineering terms,*intelligibility* – under adverse conditions, for instance, in the presence of noise, reverberation, linear and nonlinear distortions, concurrent speech, but also with crossmodal distractors such as visual or tactile ones. As to speech perception there is plentiful literature. See Moore *et al.* (2010) as an introduction – this book includes a review on listening in the presence of other sounds (Darwin, 2007/2010), see also Assmann and Summerfield (2004). The particularities of speech signals and sounds are, for example, discussed in Lotto and Sullivan (2008) and Diehl *et al.* (2004).

A broad discussion, which is also of relevance for modern communication technology, was started by the famous work of Cherry (1953, 1954), who coined the term *cocktail-party effect* for the phenomenon as to which it is possible in a babble as generated by a flock of concurrent talkers, to concentrate on the speech of one of them and disregard the others. To further investigate into the details of this effect, Cherry played different speech samples to listeners' right and left ears, that is, *dichotic* presentation while the listeners were instructed to repeat synchronously what the talker in one ear was saying – *speech shadowing*. The assessors were able to shadow the attended, relevant talker. As to the unattended, irrelevant one in the other ear, they could, after the experiment, hardly remember anything of what he/she had said . At best, they could report whether it was a male or female speaker, or whether the speaker was replaced in between by another speaker, or just by a tone. In other words, they had only detected sub-semantic features. Yet, there are exceptions: For instance, listeners detected their own name when it was repeatedly mentioned by the irrelevant talker (Moray, 1959). Obviously, primitive features and higher-level features are treated differently, and at all levels attention is involved. The question as to which extent the respective features are processed in succession, and thus may give rise to bottleneck issues, or processed in parallel, and thus may give rise to issues of work-memory size, has not yet been solved in detail.

It is a particular characteristic of speech signals that they can undergo severe distortions at the physical and, consequently, at the level of primitive perceptual features, without becoming unintelligible. For instance, speech signals reconstructed from only the zero-crossing of the original are fully intelligible. Also, speech superimposed by continuous noise of the same power level, is still very well understood. Further, speech is extremely robust with regard to interruptions, such as by impulsive noises or channel drop-outs (Warren *et al.*, 1972). At the higher-level features, this robustness is even more striking. For phonetic, phonologic, syntactic, grammatical, and semantic items such as allophones, phonemes, morphemes, syntax, grammar, prosody and meaning, we always see an immense redundancy – which is obviously a, if not the, reason for the remarkable robustness

of spoken language. If uncertainties are left in the process of speech recognition and understanding, they can often be solved by inference from the cognitive system on the basis of linguistic and domain knowledge – a capability also known as *combinatorial competence*.

Since speech recognition is no core concern of TWO!EARS, this issue will not be discussed here in more detail. Yet, two items are due to be mentioned at this point,

- How are the "higher-level" features derived from the primitive ones – a process which is performed almost instantly most of the time and, as a rule, unconsciously?
- How does the system actually attend to a specific talker, that is, which features does the system concentrate on when tracing a talker and trying to understand him/her?

The answer to the first question is given by the assumption of specific routines that reside in the central nervous system and perform specific tasks very much like *apps* in a smart-phone – see Bregman (1993). They will be dealt with in the following section. A common answer to the second question is the assumption that the auditory system and the mind *"glimpse"* at the incoming information and pick those segments of speech features, be it high and/or low level, for further processing that show low amounts of statistical uncertainty (Miller and Licklider, 1950, Bregman, 1991, Cooke, 2003). These are then, consequently, the items for the system to attend to.

Some remarks on auditory grouping, *Gestalt* rules and stream segregation In the process of forming *objects* and *auditory scenes* composed of such objects, routines are assumes that identify and fuse those primitive features that define an object – usually physical attributes that, in each case, originate from an individual sound source. This process is also called *grouping*. Two categories of grouping routines are usually distinguished (Bregman, 1991, 1993, Warren, 1982),

(a) *Primitive grouping routines* These are so-to-say "hardwired", that is, they cannot be modified. They operate very quickly and unconsciously. There is no way of manipulating or modifying them

(b) *Learned grouping routines* These routines are called *schemata* – a term from cognitive psychology. They develop and optimize themselves in the course of experience with the environment. They can thus be modified by training and may also repond to external instruction

Primitive grouping works according to rules that *Gestalt* psychology has discovered, namely, among others, the rules of *proximity, similarity, common fate, closure, simplicity, habit* and *persistence* – compare, for example, von Ehrenfels (re-edited 1990), for a review see Jekosch (2005). Schemata-based grouping represent cognitive processes that, although

more complex than primitive grouping, nevertheless run in fast and routine ways. They are typically task specific in that they represent efficient ways of grouping to the end of receiving plausible output, that is, output that shows low uncertainty in the respective environments.

Apparently, the properties of the two grouping-routine categories resemble the categories of attention as introduced in Sec. 3.2 so much that it suggest itself to conceive them as two sides of the same coin in terms of underlying principles. Thus, it holds also that the two grouping-routine categories are confluent, that is, are without a stringent border between them.

Similar routines as used to form auditory scenes are also applied to analyze auditory scenes – compare the literature regarding *computational auditory scene analysis* (CASA), for example, Rosenthat and Okuno (1998), Cooke (1993). Here perceptual objects or auditory scenes are identified, and the auditory streams that form them are segregated on the basis of attributes such as *onset times, duration, harmonic structures, interaural arrival-time and level differences*, and *AM and FM features*.

As to the consequences of grouping and segregation for the design of feedback in the Two!Ears system, the challenge is to identify the specific tasks to be solved, select approriate grouping algorithms, eventually adapt them or even develop new ones. The tasks will be set externally or assigned by expert modules within the system.

3.3 Realization of attention processes in robotics

One aspect of attention processes in biological organisms is that they can be seen as an "economical" way of dealing with the vast amount of information that is continuously available in the surrounding world (Fabre-Thorpe, 2003) and to achieve real-time processing despite limited computational capacities (Schauerte and Stiefelhagen, 2013). In order to analyze and understand this amazing biological ability, a range of attention models evolved over the past years, including the *feature integration theory* of Treisman and Gelade (1980) and the *guided search model* as proposed by Wolfe (2007). For a more detailed overview of attention models see, for example, Frintrop *et al.* (2010) or Begum and Karray (2011). Computational realizations of such psychophysical models of attention eventually led to artificial attention mechanisms like Vocus (Frintrop, 2006) or the *neuromorphic-vision toolkit* by Laurent (2014).

Following Frintrop (2006) in that computational attention systems usually intend to improve technical systems, it comes without surprise that artificial attention is of profound interest for robotics. Although visual attention dominated this field so far, auditory attention mechanisms became more and more important recently, particularly, in robotic

systems that rely on multimodal input. In the following overview, recent approaches in visual and auditory robotic attention are subsumed, whereby a focus is layed on elements that may become of interest in the context of the Two!Ears project.

Most robotic attention systems based on vision combine several *saliency maps* (Frintrop *et al.*, 2010), such as for color, shape, texture and motion (Trifa *et al.*, 2007) to the end of arriving at a *master map* (Zaheer Aziz *et al.*, 2006) that eventually guides the attention of the machine. As such a process may become computationally intense, Zaheer Aziz *et al.* (2006) proposed to work on image regions instead of plain pixels, whereby fast region-growing methods extract salient image blocks from a given input stream, such as region-based features like *symmetry* and *eccentricity*. This allows for significant acceleration of downstream attention-focusing. Though Zaheer Aziz *et al.* (2006) concentrated on vision, their ideas might well be transferred into the acoustic domain. By treating, for instance, the binaural activity maps as two-dimensional "images", it were possible to apply region-based feature extraction and subsequent block-wise processing to accelerate the identification of salient auditory events within given activity patterns.

Ruesch *et al.* (2008) demonstrated the power of selective attention in a humanoid robotic framework, based on an iCub robot (Metta *et al.*, 2008). Realizing the use of *Salienyovert* attention in humans (Frintrop *et al.*, 2010), they proposed to endow the *iCub* with the ability to perform *active vision*, which can be seen as an technical equivalent ofr overt attention (Frintrop *et al.*, 2010). *Saliency maps* (Frintrop *et al.*, 2010) for visual and acoustic input are generated independently. The evaluated visual features include *intensity, color hue, directional features* and *motion* (Ruesch *et al.*, 2008). Saliency maps formulated from acoustic features include the position of potential sound sources regarding their azimuth via interaural time differences, and their elevation via spectral notches (Hörnstein *et al.*, 2006). Saliency maps of both modalities are then unified by projecting them to an *ego-sphere*, which is head-centered and fixed with respect to the robot's (Ruesch *et al.*, 2008). In conjunction with a *dynamic inhibition-of-return mechanism*, the unified saliency maps allow the iCub to demonstrate a "rich attentional behavior" and to autonomously explore multimodal stimuli in moderately complex environments (Ruesch *et al.*, 2008).

Considering that purely visual attention systems fall short of reacting to *salient* events outside the visual field of view, Kuehn *et al.* (2012) learned from *Bayesian surprise* techniques (Itti and Baldi, 2009) and introduced a concept of *auditory surprise* (Schauerte *et al.*, 2011). Thereupon unexpected sound events are identified and corresponding sound sources are localized using the SRP-PHAT approach of Machmer and Moragues (2009). Cue fusion then takes place on the basis of a Gaussian-mixture model that integrates visual and auditory information in the sensor space. The proposed mechanism tries to generate exploration strategies to reduce the amount of necessary ego-motion for saving energy and, also, wear-and-tear in the robotic device (Kuehn *et al.*, 2012).

Okuno *et al.* (2001), among other experts, emphasized the importance of audition in modern robotics. They created a multimodal control framework to guide a humanoid SIG robot (Kitano *et al.*, 2000) in service and assistance tasks, based on a distributed architecture where vision, audition, motor control and speech synthesis are realized as single modules that communicate with a "cognitive processing unit", a SIG server, that addresses tasks like *association*, that is, multimodal stream formation, and attention focusing (Okuno *et al.*, 2001). Distributed communication is realized via a high speed TCP/IP network that ensures high systemic flexibility. The system's audition component employs phase and intensity differences of the signals captured by two microphones to perform sound localization via an *active-audition* method proposed by Nakadai *et al.* (2000a). Note that Okuno *et al.* (2001) explicitly deals with *ego-noise* suppression. Following Nakadai *et al.* (2000a) they employed two auxiliary microphones inside the robot's case in order to record and eventually compensate for motor noise.

Claiming that multimodal perceptual abilities would greatly enhance the ability of a robot to interact with humans, Trifa *et al.* (2007) discussed bottom-up sound-source-localization techniques with regard to their precision and their usefulness in multimodal, bottom-up/top-down attention-focusing frameworks. To that end, Trifa tested the *generalized cross-correlation* with and without phase transform, the *information-theoretic delay criterion* of Moddemeijer (1988) and *cochlear filtering* based on *gamma-tone-filter banks*. While being "handicapped by the lack of dynamic top-down frequency band selection" (Trifa *et al.*, 2007), the cochlear-filtering approach nevertheless seems to outperform other approaches with regard to reliable source localization in combined bottom-up/top-down frameworks. Further, cochlear filtering is suitable for straightforward integration of multimodal information, particularly visual cues. This could be relevant to the *sharpening-the-ears* aspect of Two!Ears, as establishing mechanisms for sophisticated frequency-band selection/attenuation is a typical feedback task that could benefit from the insights of Trifa *et al.* (2007).

Realizing that top-down factors play the dominant role in attentional competition, Yu *et al.* (2010) clearly emphasized the importance of cognitive feedback in robotic attention. In top-down processing, they proposed to rely on *task-relevant feature(s)* (Yu *et al.*, 2010) that encode only the important elements of a certain category or task. Note that these relevant-feature representations might also be seen as *concepts* for a certain category (Murphy, 2004). The top-down mechanisms of Yu then load down such conceptual scaffolds from an emulated *long-term memory* into a short-term *working memory* to actually perform salient-pattern matching. This idea of employing two different memory modalities matches biology (Bear *et al.*, 2007) and seems to be quite interesting in the Two!Ears project context. Further, the *concept formulation* paradigm, as pursued by Yu *et al.* (2010), is closely related to ideas found in *organic computing* (Würtz, 2008) and could also prove valuable in Two!Ears. Note that the system proposed by Yu *et al.* (2010) does not only realize sophisticated artificial-attention mechanism, but also learns novel objects from

presented images by applying these. From the perspective of our project, the weakly supervised learning routines employed in this procedure could become of interest for acoustic object learning and attention focusing.

With the aim of fusing audio and video processing from a biologically motivated point of view, Ravulakollu *et al.* (2011) proposed to set up an artificial *superior colliculus*, SC, which is an area of the human mid brain that is thought, together with the *inferior colliculus*, IC, to be responsible for early multimodal integration of auditory and visual stimuli processing (Ravulakollu *et al.*, 2011). Note that the emulated SC employs quite standard neural-network techniques for fusing information from multiple input modes. Herein, plain sound-source localization based on temporal differences is combined with visual LED-marker detection in order to detect a source's azimuthal position. While this strategy might be too simplistic to be incorporated in the Two!Ears framework, the idea of learning from biology is definitely worthy of further evaluation.

Walther and Cohen-Lhyver (2014) employed *dynamic weighting* methods to guide a PR2 robot's (WillowGarage2014, 2014) attention in search and rescue tasks. To that end, a virtual environment, the *Bochum experimental feedback testbed* (BEFT) was created where an artificial robot seeks for avatar "victims". The rescuing device is equipped with a baseline expert system scheduled via *Petri nets* (Murata, 1989), a separate task planner, and a path synthesis module based on energy-minimization techniques. Simulated acoustic and visual stimuli in the BEFT framework are artificially degraded to mimic adverse natural-scenario conditions and sensor issues. The dynamic weighting module is designed to operate on these degraded features. By continuously monitoring a given scene for incoming sounds, it evaluates the *congruency* of a given stimulus (Walther and Cohen-Lhyver, 2014). High weights are assigned to novel stimuli that seem to be incongruent with the current environmental model and appear "interesting". Lower weights are assigned to objects that have already been explored or are of less interest for the actual task of the robot. In this way, the robot can make a decision, based on the weight distribution, on which of the objects shall receive attention in the next frames. Given a novel, interesting sound signal with low congruency, the machine might instantaneously turn to this stimulus or might also suppress the *turn-to reflex* in cases were the received acoustical input is congruent and thus less interesting. Though the actual connection between BEFT and the dynamic weighting module is still under construction, it seems that the feedback characteristics of the dynamic-weighting module are highly interesting within Two!Ears and should definitely be pursued to better understand the reflective components inherent to the well-known turn-to reflex in human beings.

Emphasizing that "for a robot to show intelligent and interactive behavior in the presence of humans, it is important that both verbal and non-verbal behaviors of humans, such as facial expressions and body language that accompany speech, be detected", Yan *et al.* (2013) propose to employ artificial attention in order to boost the capabilities of robot-

based telepresence systems. Their system is based on audio-visual-cue fusion to locate speakers in a conference room. The attention of the robots guided by evaluating visual cues from face detection and very basic lip-reading algorithms together with standard audio-based source localization techniques. Note that the proposed system has to be initialized manually and requires arrays of more than two microphones. Nevertheless, Two!Ears could well learn from ideas of Yan *et al.* (2013). Their robot generates an *environmental model* that encodes the position of any speaker the system could see and hear during manual initialization. This model, also referred to as the *short-term memory* of the robot (Yan *et al.*, 2013), is updated on-the-fly during autonomous system operation and enables the machine to reduce the search space and confirm the precise locations of each conference participant. Note that the proposed approach naturally relies on the *turn-to reflex* – refer to Chaps. 2 and 5. In the proposed aproach, if the azimuth of an initiating speech signal does not coincide with the current optical focus, the acoustic stimulus takes lead and causes the robot to direct its optical sensors towards the expected speaker position. Vision is then used to verify the robot's hypothesis and to enhance the acoustic azimuth estimate.

While being based on microphone arrays instead of the ear signals of an artificial head such as KEMAR, and by using a realtively small NAO humanoid (Aldebaran, 2014) instead a near-human-sized PR2, the *embodied audition for robot* (EARS), project is interesting (EARS, 2014). Due to the fact that that *human-robot interaction* (HRI), is a complex and hence largely unsolved problem, particularly when faced with multiple persons who may simultaneously require the robot's attention, the EARS-project tries to mimic human capabilities in source localization, tracking and environment mapping, focussing of arrays, echo cancellation, blind speech dereverberation, noise reduction and interference suppression (Evers *et al.*, 2014). Further, it incorporates multimodal information to bias attention focusing and establishes audio-visual environmental mapping in order to track the position of observed speakers. While the first idea, multimodality, will definitely be realized within the feedback scope of Two!Ears, our project could well learn from the latter method in order to set up a purposeful world model that naturally integrates auditory and visual hypotheses. With the above, it is easily seen that Two!Ears and EARS are closely related, and many of the techniques researched in EARS will also find application in Two!Ears, making a close monitoring of potential synergies between both projects mandatory.

3.4 Conclusions in view of the goals of the current project

Attention of biological systems is a specific state of their perceptual organs and their central-nervous systems – often with involvement of the motor systems – in which a concentration on an attended issues can be observed, usually in relation to a task that has

been communicated to the system.

The relevant literature, as briefly reviewed in the current chapter, provides information on specific cues which may trigger specific attention, and on the perceptual and cognitive consequences of particular states of attention – here, predominantly regarding the auditory system. What is still scarce, though, are more detailed descriptions of the actual physiologic processes that take place when particular states of attention are entered. There is wide agreement, though, that the system at large must embody bottom-up (signal-driven) as well as top-down (hypothesis-driven) processes that are interleaved via numerous feedback loops. Thus, the complete system can be conceived as a *cybernetic* organism. Unfortunately, as is well known, the analysis and control of such systems is not a trivial task.

Engineering models of listening, like the one currently being developed in the Two!Ears project, attempt to design computer algorithms to mimic processes of attention with the limited resources of today's technology. If this approach, hopefully, shows success, the results may – in turn – assist psychology and physiology in the testing, or even verifying, some of their hypotheses regarding the processes of attention.

From approaches in the information technologies and in robotics, where systems with attention algorithms have already been conceptualized and implemented, the following useful hints can be obtained, among others.

- Blockwise processing and, consequently, the selection of adequate block-based feature types will enhance system performance
- Noise which is generated by the system itself, that is, the ego-noise of robots, must be considered and, eventually, compensated for
- Weakly-supervised learning routines, such as desribed by Yu *et al.* (2010), may be advantageous for attention-focusing procedures
- When multimodal information is available to the system, it should be integrated and used
- In-depth exploration of human cognitive functions and the underlying biological "hardware" will help to enhance sound and vision processing by adopting concepts from nature

4 Feedback via the olivocochlear system

This chapter reviews the anatomy and functional significance of the olivocochlear system, with a particular emphasis on aspects that are relevant to the goals of the Two!Ears project. Computer models of the olivocochlear function are reviewed, and conclusions are drawn about the likely role of efferent circuits in the Two!Ears software architecture.

4.1 Structure of the olivocochlear system

In order to understand the functional relevance of the olivocochlear system, it is necessary to know something about its anatomy and physiology. The following sections review these topics only briefly; the reader is referred to the comprehensive reviews by Guinan (1996) for more details. More recent reviews by Guinan (2010, 2014)will also be helpful in understanding current topics of study.

4.1.1 Anatomy

The olivocochlear system (OCS)[1] is named as such because its neurons originate in the superior olivary complex (SOC) and project from there to the cochlea. The SOC is an

[1] List of abbreviations used in this chapter

AN ... auditory nerve
CAP ... compound action potential
CAS ... contralateral acoustic stimulation
DPOAE ... distortion-product oto-acoustic emission
DRNL ... dual-response-nonlinear (model of cochlear filtering)
LOC ... lateral olivocochlear system
MOC ... medial olivocochlear system
OAE ... oto-acoustic emission
OCB ... olivocochlear bundle
OCS ... olivocochlear system
SOC ... superior olivary complex
SRT ... speech-reception threshold

area of the brainstem that receives binaural inputs and is dedicated to the processing of acoustic signals.

The neurons of the OCS, which collectively form the olivocochlear bundle (OCB), are divided into two subsystems. The medial olivocochlear system (MOC) originates in the medial part of the SOC and projects mainly to outer hair cells in the cochlea. In contrast, neurons of the lateral olivocochlear system (LOC) originate in the lateral portion of the SOC and mainly project close to inner hair cells and their associated auditory nerve (AN) fibres. The peripheral effects of the LOC are not well understood, and will not be the focus here. Most research on the OCS concerns the effects of the MOC on cochlear function. A schematic view of the MOC pathways is shown in Fig. 4.1.

MOC neurons respond to sound, and thus form the efferent (descending) part of a reflex – the *MOC reflex*. They project both to the ipsilateral and contralateral ear, although in most mammals the majority of MOC fibres project contralaterally. A small proportion of MOC neurons are excited by sound in either ear. Figure 4.1 shows the pathways of the sound-evoked MOC reflexes for one cochlea – shown on the right. The ipsilateral pathways are shown in solid black, whereas the reflex pathway in response to contralateral sound is shown in gray.

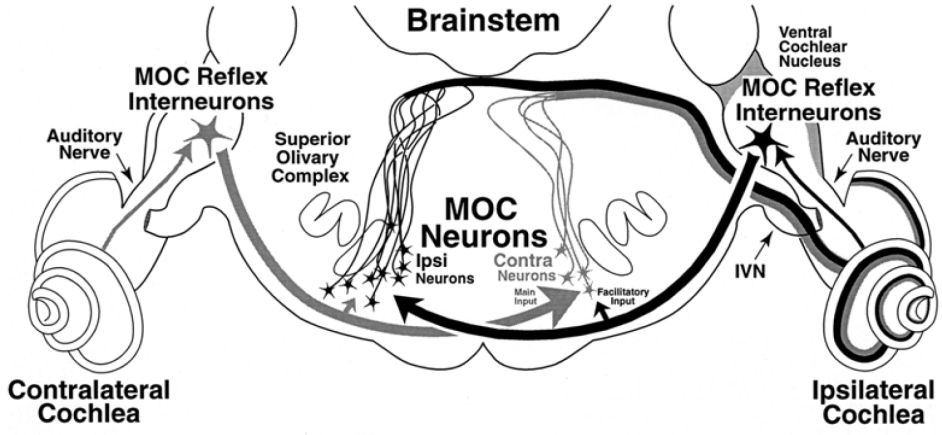


Figure 4.1: Schematic diagram showing the pathways of the medial olivocochlear (MOC) reflexes. The pathways in solid black correspond to those activated by sound delivered to the ipsilateral ear. Pathways in gray are activated by acoustic stimulation of the contralateral ear – from Brown *et al.* (2003)

4.1.2 Physiology

Physiological studies of the OCS usually take one of two approaches. One approach is to sever the fibres of the OCB in order to determine the effects on cochlear function when

efferent feedback is absent. Alternatively, AN responses can be recorded with and without electrical stimulation of the OCB, delivered by electrodes placed on the nerve bundle. Also, in both physiological and psychophysical experiments, the effects of the OCS in one ear can be studied by delivering a MOC-evoking sound to the opposite ear – so-called contralateral acoustic stimulation (CAS).

The effects of MOC activity are generally attributed to two underlying processes that have fast and slow time scales. When the OCB is electrically stimulated, both processes cause a suppression of the auditory nerve compound action potential (CAP), and both are extinguished when the OCB is severed. However, fast processes act over a time scale of 10–100 ms, whereas slow processes occur on a time scale of 10–100 s (Sridhar *et al.*, 1995, Zhao and Dhar, 2011). They also differ in their frequency dependence. In guinea pigs, fast effects are greatest for stimulus frequencies in the range 6–10 kHz, whereas slow effects are maximal in the frequency range of 12–16 kHz. Finally, they differ in their effect on the phase of basilar membrane displacement, namely, fast effects cause phase leads, whereas slow effects cause phase lags (Cooper and Guinan, 2003).

4.2 Functional significance

4.2.1 Role of the MOC in unmasking

It has been thought for some time that the MOC enhances the ability to detect and discriminate signals in noise. Kawase *et al.* (1993) studied the responses of AN fibres to tone bursts in continuous masking noise. When MOC-activating noise was presented contralaterally, the AN response to the masked tone was increased and the response to the ipsilateral noise masker was decreased. The largest anti-masking effects – in cat – were seen for AN fibres with best frequencies between 6–12 kHz. Similarly, Dolan and Nuttall (1988) recorded CAP responses to short tones in noise and silence. When the tones were embedded in noise, the response to the tone was reduced in the CAP response. Yet, masking of the tone could be partially reversed by electrical stimulation of the OCB.

Both of these findings can be explained by the observation that activation of the MOC leads to a shift of the cochlear rate-level function towards higher sound levels (Dolan and Nuttall, 1988). As long as the tone is more intense than the noise, the effect of a shift in the rate-level function will be to reduce the response to the (low-level) noise more than the (higher-level) tone. Hence, the MOC effectively unmasks the tone. Adaptation also plays a role. By suppressing the response to the noise, the MOC reduces adaptation in the AN and thus permits a larger response to the tone when it occurs. Hence, it seems very likely that a functional role of the MOC is to enhance the auditory response to transient stimuli (Kawase *et al.*, 1993).

It should be noted that such an effect of an enhanced auditory response to transient stimuli requires that the tone amplitude is clearly higher than the amplitude of the noise within the tuning range of the nerve fibre. In terms of human listening conditions, this translates into conditions with clearly suprathreshold levels of the desired signal (tone, speech) relative to the background sound. At signal detection threshold or, correspondingly, at the *speech-reception threshold* (SRT), this signal-to-noise ratio is slightly negative, and one can conclude that for these conditions, activation of the MOC will not influence human performance. The only exception is the detection threshold for very short signals, as they are used in overshoot conditions, because for very short signals, the signal amplitude at threshold is considerably higher than the corresponding noise amplitude. As will be described in the following section, the basilar-membrane nonlinearity in combination with the MOC activity has been used to model several aspects of the psychophysical overshoot effect.

There is also evidence that the MOC reflex benefits speech perception in noise. Kumar and Vanaja (2004) measured speech identification scores in quiet and with ipsilateral noise maskers in two conditions, in which a contralateral (MOC-evoking) stimulus was either absent or present. It was found that speech identification scores were higher when the contralateral stimulus was present, suggesting that the MOC aids speech recognition in noise. However, the ecological relevance of such studies has been questioned. Mishra and Lutman (2014) found that MOC activity is not related to speech-recognition performance in noise in the absence of contralateral acoustic stimulation (CAS), but also that MOC inhibition was correlated with speech-recognition performance when CAS was applied. It therefore appears that MOC-mediated unmasking of speech is possible, but that "the auditory system does not use this resource in a reflexive manner" (Mishra and Lutman, 2014, p. 5). Instead, it seems likely that top-down processes recruit MOC unmasking mechanisms only in specific listening situations, according to attention and experience. Note that there may be instances where MOC activity will not be helpful. For example, an MOC-mediated shift in the cochlear rate-level function will not help the detection of a target sound if its level is well below the level of the masking sound. Similarly, MOC activity may have a positive or detrimental effect on speech identification depending on the particular speech material and the spectral region in which masking occurs (de Boer *et al.*, 2012).

4.2.2 The overshoot effect

The psychophysical phenomenon termed *overshoot effect* refers to the fact that the detection threshold for a brief tone presented in noise improves as the onset of the tone is delayed after the start of the noise (Zwicker, 1965). Overshoot is maximal at mid-masker levels, and is reduced in individuals with cochlear hearing loss. The MOC might underlie this effect, due to the same unmasking effect discussed in Sec. 4.2.1 above. The noise burst will elicit

efferent suppression via the MOC, which decreases the gain in the cochlear amplifier. As a result, the rate-level function of the auditory periphery shifts to the right, which reduces the response to the low-level noise more than it does to the higher-level tone. Several studies have looked for correlates of this mechanism in otoacoustic emissions (OAEs) with conflicting results – see Guinan (2010) for a review. On balance, it seems likely that the MOC reflex is at least partly responsible for overshoot.

4.2.3 Binaural hearing

It has been suggested that LOC efferents balance the outputs of the two cochlea in order to optimize binaural hearing. Darrow *et al.* (2006) found that unilateral lesion of LOC efferents disrupted the strong interaural correlation in neural excitation level observed in normal ears. They propose that LOC efferents act to correct disparities in excitation level over a long time scale – tens of minutes. For example, slow growth of excitation in one ear might be suppressed by a corresponding growth in inhibitory feedback via the ipsilateral LOC. Such a mechanism would ensure the accuracy of interaural level difference (ILD) computations, which involve a comparison of excitation level in the two ears. However, Larsen and Liberman (2010) have recently presented evidence against this *output-balancing hypothesis*. In animals with an intact LOC, they produced threshold shifts in one ear and found that these were not matched by changes in cochlear function in the opposite ear.

4.2.4 Protection against acoustic trauma

It has been suggested that slow MOC effects protect against acoustic trauma. Physiological data recorded from the guinea pig show that electrical stimulation of the OCB reduces the threshold shift caused by long-term exposure to intense sound (Reiter and Liberman, 1995). However, recent data obtained from humans using a non-invasive technique suggests that slow MOC effects are very small for noise levels of up to 83 dB SPL (Zhao and Dhar, 2011). It is therefore currently unclear whether slow MOC effects protect against acoustic trauma in humans.

4.2.5 Ipsi- and contralateral MOC effects

As visible from Fig. 4.1, the MOC reflex can affect the hair cells in both the ipsi- and the contralateral ear. To distinguish these two reflex loops, the terms ipsi- and contralateral MOC effect are used. In the *ipsilateral* MOC reflex, neural connections in both the ascending (afferent) part, towards the MOC neurons, as well as in the descendent (efferent)

part of the reflex loop, towards the outer hair cells, cross the midline. In the *contralateral* MOC reflex, the afferent connections again cross the midline, while the efferent connections do not cross but connect to the inner ear of the same side. The relative strengths of ipsi- versus contralateral effects on *pontaneous otoacoustic emission* have been studied in Lilaonitkul and Guinan (2009). While for narrowband noise, ipsilateral MOC effects were about twice as strong compared to contralateral effects, the two effects were of similar size for broadband noise elicitors. This combination of ipsi- and contralateral MOC effects does imply an interesting difference in the strength of the MOC reflex between conditions with stimulating only one ear, that is, one channel in headphone presentation, and those where both ears are stimulated. The changes in the peripheral nonlinearity caused by the MOC reflex should be stronger in the latter case – compare the discussion in Langhans and Kohlrausch (1992).

4.2.6 Attention and learning

The MOC reflex has recently been implicated in selective auditory attention. Harkrider and Bowers (2009) found that contralateral suppression of click-evoked OAEs was greatest when listeners were not attending to a particular task. In contrast, MOC inhibition declined – compared to passive listening – when subjects attended either to the ipsilateral clicks or a contralateral suppressing noise. The authors conclude that attention causes a top-down, cortically-mediated release from efferent inhibition at the level of the cochlea. Modulation of MOC activity by attention has also been reported by Maison *et al.* (2001) and de Boer and Thornton (2007).

Subsequent studies have provided further support for this finding. For example, Smith *et al.* (2012) measured the effect of MOC efferents on the cochlear amplifier by measuring the amplitude and rapid adaptation of *distortion-product otoacoustic emissions* (DPOAEs). Rapid adaptation of DPOAEs occurred in active and passive listening conditions, suggesting a medial efferent process that is unaffected by attention. However, the overall DPOAE level was significantly affected by changes in attentional focus. Smith *et al.* (2012) conclude that two MOC mechanisms are at play – one which rapidly suppresses the responses of outer hair cells to sustained or repeated stimulation and another one that increases the salience of attended signals. Interestingly, DPOAE magnitudes were lower for attended rather than non-attended signals, suggesting that the MOC actively suppresses the cochlear responses to *attended* acoustic signals.

4.3 Computer models

There have been a number of attempts to implement computer models of auditory efferent processing and to use these models to assess the likely role of the OCS in hearing – particularly in regard to the perception of speech in noisy conditions.

Ferry and Meddis (2007) describe a model of MOC function which might be described as "open-loop", that is, the amount of efferent suppression in the model is directly fixed by the experimenter rather than derived from the properties of the acoustic input. Their model is based on the *dual-resonance-nonlinear model* (DRNL) of cochlear filtering, which has two parallel pathways, namely, a broadly tuned linear pathway and a more narrowly-tuned nonlinear signal path. The suppressive role of the MOC is modelled by inserting an attenuator at the start of the nonlinear path – as shown in Fig. 4.2. Hence, efferent suppression leads to a reduction in the amount of cochlear nonlinearity. A related computer model was previously proposed by Ghitza *et al.* (2006) in which the cochlear model of Goldstein (1990) was modified in a similar way. A related approach has also been used by Jennings *et al.* (2011), using a modification of the auditory model described by Zilany and Bruce (2006).

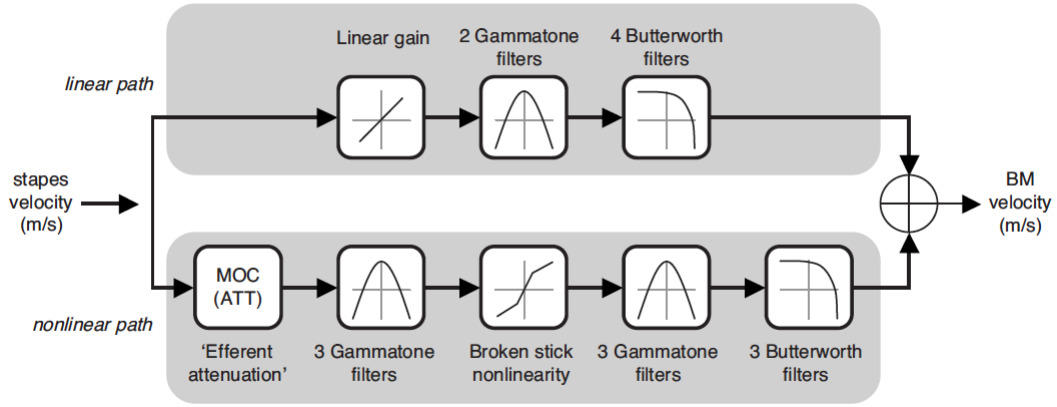


Figure 4.2: Schematic diagram of the DRNL filterbank, modified to include an "efferent-attenuation" stage. The DRNL consists of parallel linear and nonlinear signal paths. The MOC is modeled by an attenuator at the start of the nonlinear path. The degree of efferent activity is determined by the parameter ATT, where larger values of ATT correspond to greater suppression by the MOC – from Brown *et al.* (2010)

The authors show that this simple model is able to account for many aspects of the physiological data, including the observation of Russell and Murugasu (1997) that activation of the MOC reduces pure-tone-evoked displacements of the basilar membrane and leads to a shift of the auditory nerve rate-level function towards higher sound levels. They also show

a good match to shifts in rate-level functions as reported by Guinan and Stankovic (1996) and the data reported by Dolan and Nuttall (1988) in which CAP responses were evoked by brief tones in silence and noise. Regarding the latter, the model correctly predicts that stimulation of the MOC leads to an overall reduction in the observed CAP. Dolan and Nuttall (1988) also found that the CAP response to a tone in noise *increases* when the MOC is stimulated – effectively causing the tone to be unmasked. The computer model also replicates this finding, which can be explained by the effect of MOC stimulation on adaptation in the auditory nerve. Namely, activation of the MOC by the noise preceding the tone causes the response to the noise to be suppressed, therefore reducing adaptation and allowing a greater response to the subsequent tone.

In a follow-up study, Brown *et al.* (2010) studied the implications of this apparent "release from adaptation" for the perception of speech in noise. They used the model of Ferry and Meddis (2007) as the front end to a HMM-based automatic speech recognizer and found that recognition performance was poor in the presence of broadband noise when no efferent suppression was applied. However, recognition performance was greatly improved when the efferent circuit was activated. Furthermore, optimal recognition performance was obtained when the amount of efferent activity, that is, the value of the attenuator ATT in Fig. 4.2 was made proportional to the noise level. Again, Brown *et al.* (2010) explain their results in terms of "release from adaptation" by the MOC. When speech is embedded in a noise background, the noise preceding the speech will cause a shift of the rate-level curve to higher levels, effectively reducing the response to the noise and unmasking the speech.

We previously noted that the model of Ferry and Meddis (2007) is "open-loop", because the amount of efferent suppression is a free parameter that is determined by the experimenter. However, in a complete model, the amount of efferent suppression should vary in response to the recent history of the acoustic input. Such an approach is consistent with experimental data that show that cochlear gain varies continually during psychophysical tests (Jennings *et al.*, 2009). Ghitza and co-workers (Ghitza *et al.*, 2006, Messing *et al.*, 2009) describe a closed-loop model of efferent processing in which the mechanical filtering properties of the cochlea are regulated by feedback based on short-term measurements of the dynamic range of simulated auditory-nerve fibres. They used this to model consonant-confusions made by listeners in noise in a diphone discrimination task, using a simple template-matching speech recognizer. They also found that efferent feedback made a positive contribution to the intelligibility of speech in noise. Similarly, (Clark *et al.*, 2012) report a closed-loop version of the Ferry and Meddis (2007) model, in which an anatomically plausible feedback loop was used to regulate efferent attenuation separately in each peripheral frequency channel. It was found that this arrangement improved speech- recognition performance beyond that reported by Brown *et al.* (2010). Additionally, independent regulation of efferent feedback in each channel was found to be important in unmasking speech embedded in noise with different spectral profiles, such as pink noise and babble.

When translating these results to realistic speech in noise environments, it is important to note that release from masking as found in the modeling work by Brown *et al.* (2010) required that the nonlinear basilar-membrane model, and in particular the MOC feedback loop, first adapted to the noise alone. The following speech stimulus was then presented for a shorter time period, during which the MOC effect remained constant, that is, it was not affected by the presence of the speech stimulus. This experimental condition can thus not directly be compared to ongoing conversation situations in which both the background noise and the desired speech will be present for a period that is long compared to the time constant of the MOC reflex. In such a condition the MOC effect will not only be determined by the noise, but by the total stimulus comprising both noise and speech. Furthermore, as mentioned in Sect. 4.2.1 the differential effect of the MOC on noise and speech can only happen at clearly positive speech-to-noise ratios, but is less likely to happen at the human speech reception threshold where this ratio lies in the range of -5 to 0 dB. And finally, the observed influence of the MOC on speech recognition critically depended on the dynamic range of the nerve fibres included in the model and the match between speech level and this dynamic range.

Computer models of efferent feedback have also been used to investigate the mechanisms underlying the overshoot effect – see Sec. 4.2.2. Jennings *et al.* (2011) adapted the computer model of Zilany and Bruce (2006) to include MOC effects, and found that the model predicted the magnitude and level-dependence of overshoot when efferent feedback was active. Conversely, psychophysical overshoot data could not be matched when MOC feedback was not included in the model. The authors conclude that overshoot is mediated by dynamic range adaptation, which may occur at several levels in the auditory pathway but is predominantly mediated by the MOC reflex. Likewise, Ferry (2008) found that the DRNL-based model shown in Fig. 4.2 also reproduced the overshoot effect when efferent feedback was included.

4.4 Conclusions in view of the goals of the current project

The preceding review suggests a number of factors that should be taken into account in the design and implementation of the Two!Ears software architecture. These are summarized below.

- *Role of the MOC reflex in unmasking* There is compelling evidence that the MOC reflex can assist in the detection of simple sounds, such as a tone masked by noise, and also the identification of speech that is masked by noise. However, current studies are converging on the idea that the underlying mechanism is not a simple reflexive process. Rather, it is likely that MOC feedback is directed in a task-dependent manner and is modulated by attention and experience. Current computational

models of efferent feedback assume a reflexive mechanism based on measurements of AN level or dynamic range and are undoubtedly too simplistic. A key issue, then, is to identify the conditions for which MOC feedback is a benefit and those for which it is a hindrance. This issue could be addressed through computer simulations with the Two!Ears software architecture.

- *Effect of MOC activity on interaural cues* The pathways of the MOC influence activity in both the ipsilateral and contralateral ears, where they alter both the magnitude and phase of the cochlear response. In order to reliably measure ILD and ITD cues, it is therefore necessary to balance the effect of MOC activity in the two ears. There are currently conflicting reports in the literature as to how this might be achieved

- *Fast and slow effects* There is some evidence that slow MOC effects, that is, over a time scale of 10–100 s, protect against acoustic trauma. However, it is unclear from current data whether such a mechanism is relevant in human hearing. In any case acoustic trauma is not a great concern for machine hearing systems. It therefore seems appropriate that Two!Ears should focus its efforts on fast MOC effects

- *Attention and learning* As noted above, it seems likely that MOC feedback is not used indiscriminately in a reflexive manner. Rather, it is modulated by attention and experience. This is entirely consistent with the proposed structure of the Two!Ears software architecture and raises three key questions. First, how can the system identify acoustic conditions in which MOC activity is likely to be beneficial? Second, how can MOC feedback be integrated into an attentional system? Finally, how can learning algorithms be integrated with a computational model of the OCS, so that prior experience can influence the extent to which efferent circuits are activated?

5 Feedback at the sensorimotor level

The sensorimotor level constitutes the lowest layer both in the Two!Ears computational architecture and in the deployed robotics architecture. It lies just on the top of instrumentation. It is constituted of perception and/or motion functions which correspond to "hardwired" reflex behaviors and must run under severe time and communication constraints. These functions entail neither decisional nor cognitive ability.

Feedback at the sensorimotor level essentially corresponds to perception/action loops. Motor actions at a given time instant may not come as an instantaneous mapping of the perceived data, but rather as a function of the sensorimotor flow, that is, the history of the perceived data and motor commands, over a time window. Their synthesis may rely on several paradigms, ranging from bioinspired approaches or learning, to control theory.

This chapter attempts to overview the literature related to such feedback, mainly following a control-theory perspective in robotics. It is organized as follows. Section 5.1 proposes a short historical perspective on AI-based reflex actions and their implications on feedback at the sensorimotor level. Then, Sec. 5.2 discusses "situation-based" and "sensor-based" feedback to motion control as well as information-based sensorimotor feedback for *simultaneous localization and mapping* (SLAM), and for the control of sensor parameters.

The use of sensorimotor feedback in robot audition is the subject of Sec. 5.3. Some contributions are outlined, and some methodological elements are given in view of existing successful results obtained with other modalities. This section ends with examples of sensorimotor feedback which are not rooted in control or information theory.

5.1 Introduction

Early research on autonomous robots, that is, robots that can explore their environment to accomplish specific tasks, dates back to the late 1940s. In 1948, the neurophysiologist William Grey studied the bases of simple reflex actions (Grey, 1950, 1951). For his work, he built two *turtles* named "Elsie" and "Elmer", capable of autonomous behavior by moving in reaction to light and sound stress. The underlying idea was to reproduce kind of a

conditioned reflex. The invention of the electronic turtles encouraged many researchers to develop the so-called *artificial life*. Their ability to explore their environment – hence their name *Speculatrix Machina* –, their simple reflex behavior with regard to light depending on its intensity, and their adaptative capability by conditioned reflex, inaugurated the advent of robots and robotics.

Close to this idea, Valentino Braitenberg described in his book *Vehicles: Experiments in Synthetic Psychology* a series of experiments in which extremely simple robots show complex behavior (Braitenberg, 1986). *Braitenberg vehicles* became the first simple examples of *reactive methods*, realized on the basis of a pair of cameras directly connected to a pair of wheels. This strategy therefore involves *reflex actions*, where each perception gives rise to an action. This is a local strategy, effective only in the area of the environment in which the object is visible. According to the inhibitory or excitatory connections relationship and their direct or inverted connection, Fig. 5.1, the resulting behavior can become, for the same stimulus, either an approach behavior of a light target or a removal behavior. More formally speaking, these vehicles simply perform a gradient descent on the intensity of light (Ranó, 2007). However, their behaviors are subject to oscillations. In addition one has to assume that the target, namely, a light, is visible from the whole environment. This is rarely the case in practice. Nevertheless, this model describes the easiest event-driven way to make a movement toward a target, even if it remains difficult to use it in real applications.

An important step was initiated by Rodney Brooks when he proposed the concept of *embodied intelligence*. Within this new approach, perception became the central problem, whereas before it had been considered as a separate or secondary problem. These robots do not use prior models of the outside world, considering that the real world is the one which is continously perceived by the sensors. This was the birth of the approach of *behavior-based robots* (Brooks, 1991). One justification for this kind of modeling was the poor performance of robots in dealing with the real world through symbolic descriptions from traditional artificial intelligence. The goal was neither to produce cognition nor a human-like thinking process, but instead to create agents that could act in an intelligent way, postulating that the bottom-up behavioral approach is the very principle underlying biological intelligence.

Whatever the used framework to represent knowledge, in the end robots must be controlled in closed-loop in order to ensure task fulfillment. After the seventies, feedback control of robots became a research topic of its own. Control schemes of manipulator arms in manufacturing cells first relied on proprioceptive sensors – such as encoders or tachometers – which measure internal variables. Later on, exteroceptive sensors were also used in control algorithms – such as force sensors, laser range finders and cameras – which reflect the interaction with the environment. This step dramatically increased the versatility of manipulator or mobile robots. Actually, it enabled them to perform tasks such as

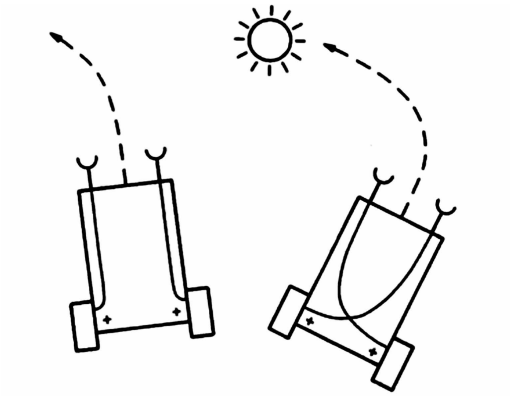


Figure 5.1: The Braitenberg vehicles

positioning with regard to objects and target tracking (Weiss *et al.*, 1987) (Espiau *et al.*, 1992), and manipulation and (Thrun *et al.*, 1998) (Cherubini and Chaumette, 2013) – and thus constituted an important milestone towards autonomy.

A topic related to exteroceptive control is active perception, the aim of which is to control the degrees of freedom and the parameters of sensors in order to improve the quality of the perceptual results (Aloimonos *et al.*, 1988) (Tsotsos *et al.*, 1995). Rather than stabilizing a quantity of interest as in exteroceptive control, the aim of an active observer is to gather information. The benefit over a passive strategy is that it helps turning some difficult or unsolvable problems into well-posed ones.

Most of the above exteroceptive control or active perception algorithms were first designed on the basis of visual sensors or laser range finders. These have been by far the most often used exteroceptive modalities in robotics. In comparison, robot audition is a fairly new topic, which emerged in the 2000's (Nakadai *et al.*, 2000b) and brought to the fore by the need of natural human-robot interaction. Early contributions focused on the analysis of static acoustic scenes from a robot at rest. Original algorithms were proposed for various auditory functions, particularly, sound source localization, speaker or speech recognition, under time and embeddability constraints in realistic acoustic environments entailing noise, spurious sources, reverberations, and so on. Over recent years, the unexpected problems and rich perspectives brought by mobility raised an increasing interest for the field of robot audition within and outside robotics (Nakadai *et al.*, 2000a),(Cooke *et al.*, 2007). For instance, active binaural localization offers the perspective of overcoming limitations in the static context, such as front-back confusion or range non-observability.

5.2 Sensorimotor feedback in robotics

The synthesis of motion inside exteroceptive control or active perception schemes follows a common structure in that, at each sampling time, the following two stages are run in sequence. First, an analysis of the perceptual or sensorimotor flow is performed. Second, on the basis of the extracted information and of the goal to be reached, the motor commands to be sent to the actuators of the robot are synthesized. Two sets of approaches can be distinguished. On the one hand, some approaches entail a localization step, that is, their internal analysis of the sensory or sensorimotor flow leads to a three-dimensional information. This information constitutes the input to the feedback control that delivers the motor commands. Such "situation-based" approaches are often termed *state feedback*, for the extracted 3D information can play the role of the state vector in automatic control. On the other hand, the so-called "sensor-based" approaches state the control problem in the space of the exteroceptive sensors. The input to the controller are then features extracted from the sensory data, from which motor commands are deduced. Since no localization is involved, these strategies are deemed as *output feedback*.

5.2.1 Situation-based motion control

Planned situation-based motions

Within motions issued from an exteroceptive localization, two subsets can be identified. The first category concerns planned, that is, reflective motions (Raffo *et al.*, 2011). Such displacements involve the sequence of the following three stages.

(a) Prior definition of a metric map of the environment

(b) Planning of motions inside this map

(c) Reactive execution of the planned movements

If the environment is not known in advance, Stage (a) consists in *simultaneous localization and mapping* (SLAM) (Durrant-Whyte and Bailey, 2006), that is, the building of the metric map of the environment and the concurrent update of the absolute localization of the robot inside this map. In its simplest form, it takes as input data the records of "landmarks" – characterized from features which can be easily extracted and matched between two consecutive percepts – and the motor commands applied to the robot. It thus consists in an analysis of the sensorimotor flow. Kowledge of the prior dynamics of the robots and a measurement model are needed to infer the hidden situation variables from the spatio-temporal coherence of the assimilated measurements. The underlying

techniques can be stochastic filtering and smoothing (Thrun *et al.*, 2005) (Chiu *et al.*, 2013) or optimization (Mouragnon *et al.*, 2006).

Stage (b) relies on motion-planning techniques (LaValle, 2011). Stage (c) entails the online exteroceptive localization of the robot within the pre-learnt map, as well as the execution of the planned movement using control techniques for stabilization or path/trajectory following (Canudas de Wit *et al.*, 1993).

Dynamic environments are subject to obstacles. When static obstacles that were not included in the map are present during trajectory execution, Stage (c), avoidance strategies must be launched, which can consist in an online suitable deformation of the planned motion. As a consequence, the environment map can be modified so as to include these obstacles. If the obstacles are moving during map building, then SLAM and *moving-object tracking* (SLAMMOT) techniques must be deployed in order to track them and prevent their incorporation in the map. The existence of mobile objects during trajectory execution can also be handled, for instance, by characterising their spatiotemporal behavior and taking this into account in the distortion of the planned trajectory.

Situation-based reflex motions

The second category of motions can be described under the term *situation-based exteroceptive control*. It concerns genuine feedback controllers which take as input the relative pose between a robot – or its end-effector – and a target, issued from an exteroceptive localization. It aims at regulating this signal to a reference value, for instance, in the context of a positioning or target-following task.

An example in this context is the use of vision for robot control. The flexibility brought by this perceptual modality was acknowledged long ago, when robotics was still mostly restricted to industrial applications. For instance, the visual guidance of assembly tasks with a manipulator arm proposed in (Shirai and Inoue, 1973) enabled for the first time to

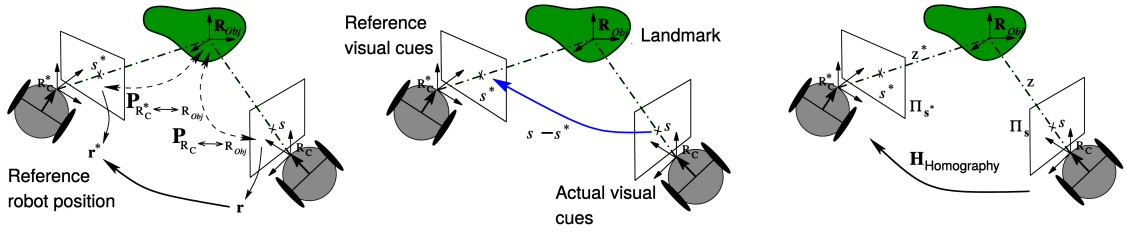


Figure 5.2: Visual based control schemes – extracted from(Folio, 2007), (left), 3D situation-based visual servoing, (middle), and 2D1/2 visual servoing, (right)

consider less contrained working environments, with a tolerance on the placement of the objects to put together. The idea was to define a location/attitude pair to be reached by the end-effector of the robot from a vision-based pose-estimation algorithm, then to run a proprioceptive control of the end-effector to this pose until it reached its steady state value. Due to robot-modeling and calibration errors, this process, called *static look-and-move*, should be repeated several times. A natural extension was made possible several years later, when the visual sensors reached enough performance – Fig. 5.2, left panel. It consists in the continuous visual-based localization of the end effector and in its regulation to a reference value. This strategy is named *dynamic look-and-move*, or *position-based/3D visual based control* – see, for instance, Wilson *et al.* (1996). Perceiving while moving thus raises conventional problems of control, such as stability, performance, robustness to uncertainties and/or perturbations. Depending on the amount of information used, the visual localization can be a static or dynamic process, that is, may involve the current image or the visuo-motor flow on a sliding time window.

5.2.2 Sensor-based motion control

In opposition to situation-based strategies, sensor-based motion comes from a feedback-control problem expressed in the space of the sensors. The idea is to extract features and build a closed-loop system which aims to drive them at some reference configuration. Getting rid of the localization stage is interesting on many viewpoints. Namely, it is no longer necessary to use a model of the sensor, nor of the system which supports it. Further, there is no need to use a model of the target. Also, feature extraction is much faster and less error prone than localization. In the end, the obtained control scheme is naturally less sensitive to unavoidable uncertainties in the various underlying models.

In the case of vision, the idea of building a control system directly in the image plane of one or several cameras dates back to Weiss *et al.* (1987), giving rise to *image-based/feature-based/2D visual servoing*. The visual feedback controller then aims at extracting some features extracted from the image, such as points, lines or moments, reach a reference value corresponding to the fulfillment of the task – Fig. 5.2, middle panel. The reader may refer to the seminal work of Espiau *et al.* (1992) and the tutorials by Hutchinson *et al.* (1996), Chaumette and Hutchinson (2006) and Chaumette and Hutchinson (2007). 3D visual servos are seldom used nowadays. Conversely, applications of 2D visual servos are plenty – from automated harvesting to surgery assistance, for instance. Some hybrid "2D–1/2" schemes combine 2D features and localization variables (Malis and Chaumette, 2000) – Fig. 5.2, right.

Compared to planned displacements, sensor-based motion is local in essence. Nevertheless, strategies do exist for large-scale sensor-based navigation. The principle is to define a

topological map of the environment which describes the spatial visibility relationships of landmarks or even consist in a memory of images. Then, local sensor-based servos to the series of landmarks constituting the task are sequenced (Souères *et al.*, 2003). The main issues here concern how to switch between control laws when the sensor-based goal changes and how to schedule or combine different motions for a multiple degrees-of-freedom structure – typically a humanoid robot – for example, through a "stack of tasks" (Mansard and Chaumette, 2007). Perceptual constraints have to be taken into account like the following. Typically, a switch can happen only when the next visual goal is detected in the current image. An active behavior is proposed in Durand Petiteville *et al.* (2010) so that the camera is controlled both to reach the current visual goal and to look for the next one, while handling occlusions and static obstacles. Robot control using an image memory involves a specific data structuration and a higher complexity (Cherubini and Chaumette, 2013).

5.2.3 Active information-based sensorimotor feedback

Exploratory reflex motion for SLAM

During map building via simultaneous localization and mapping (SLAM) – compare Stage (a) introduced in Sec. 5.2.1 – the robot/sensor can be moved by hand, for instance, by means of a joystick. Nevertheless, it can also be endowed with autonomous motion capability, in order to improve the information held in the measurements about the situation variables to be estimated. This extension obviously leads to feedback from the result of the analysis of the sensorimotor flow to the motor commands.

Considering a stochastic filtering strategy that assimilates the sensory flow and combines it with the motor commands of the robot/sensor, the following has to be dealt with. Maximizing the information about the hidden state variables, that is, the situation, is known as *information-theoretic control* or *information-gathering control*. To conduct this kind of control synthesis, information theory can be applied to quantify the information contained in the posterior probability density function (pdf) coming from data fusion. The concepts of Shannon entropy and mutual information, together with the Bayesian extension of the Fisher information matrix (FIM), as formally defined in Cover and Thomas (1991), have been widely used in the literature, for instance, in Feder *et al.* (1999). When dealing with Gaussian approximations of the posterior pdf, this leads to maximize a "size criterion", such as the determinant or the trace, among others, of the inverse of the posterior covariance matrix. In some studies (Bourgault *et al.*, 2002, Grocholsky *et al.*, 2003) the notion of mutual information is preferred. It measures the amount of information one pdf contains about another pdf, This concept has also been used to characterize the information between the map and future sensor measurements in Julian *et al.* (2013). All these scenarios lead to the resolution of optimal control problems. A case

study can be found in Grocholsky (2006). For interesting general statements of several information-based control problems compare Scardovi (2005).

Active control of sensor parameters

Besides active SLAM methods which aim at controling the robot/sensor motions, some approaches control the sensor parameters – such as the zoom and attitude of a pan-tilt-zoom camera – in order to improve the modeling process along some criteria like map resolution, map accuracy, robot position accuracy. Generally, this is based on the optimization of a utility function (Sanchiz and Fisher, 2000).

An information-theoretic approach can also be applied in a visual context when the aim is not the best move of the robot but instead the set of camera parameters that best improve image processing. To improve object recognition, an active camera control problem – in terms of focal length and pan & tilt angles – can be defined in such a way as to maximize an information criterion (Denzler and Brown, 2002) (Sommerlade and Reid, 2008). Likewise, a technique refered to as *next-best-view planning* is performed by applying information-theoretic control instead of a 3D–reconstruction scenario (Wenhardt *et al.*, 2006).

5.3 Sensorimotor feedback in robot audition

The incorporation of motion in robot audition leads to so-called *active functions*. Corresponding contributions are reviewed in this section. Though they constitute a promising way to improve auditory scene analysis, the literature is still quite scarce, and active localization is mostly concerned with.

Importantly, exploiting the motion of a robot/sensor raises an important problem, namely, the noise coming from the robot itself, also called *ego-noise*, can significantly alter perception. An illustrative extreme has been reported by (Furukawa *et al.*, 2013), where a multirotor UAV, endowed with a microphone array, has to face non stationary ego-noise emitted during its flight and while performing a sound-source localization task. Ego-noise cancellation is still under investigation. Feasible solutions to this problem are based on noise patterns that are collected into ego-noise databases and then substracted from the desired signals according to the actual movement (Ince *et al.*, 2009, 2011).

5.3.1 Situation-based analysis of the sensorimotor flow for active audio-motor localization

Short-term analysis of the binaural stream to extract 3D information about the environment constitutes the well-known field of binaural source localization. As in humans, this kind of a short-term processing leads to the extraction of features such as interaural arrival-time (ITDs) and level differences (ILDs), that is, features that are specific for actual position of the sensors in a given acoustic scene. However, spatial information as determined from these features is ambiguous. Indeed, the interaural transfer function is not structurally identifiable, as a given transfer function can correspond to multiple source positions – for instance, for a spherical head. Actually, the admissible source positions constitute a so-called *cone of confusion*. Consequently, without further prior information, front-back or up-down ambiguities are hard to resolve, and the distance of farfield sources is difficult to estimate. Theoretically, an anthropomorphic head-and-torso simulator, a so-called *artificial head* removes some indeterminations. However, in practice, these still remain effective to a certain extent due to measurement noise and modeling uncertainties.

The assimilation of the extracted short-term spatial cues over time and their combination with the motor commands of the sensors can constitute a way to remove such ambiguities. A solution of this problem in a stochastic-filtering context was proposed in Ward *et al.* (2003) and Asoh *et al.* (2004) – outside robotics. For robot audition with a microphone array, Valin *et al.* (2006) developed a system that is able to localize and track moving sources in the presence of noise and reverberation. To this end, a particle filter is fed with the approximation of the output energy of a delay-and-sum beamformer via a generalized cross-correlation with modified PHAT processor. In the same vein, Marković and Petrović (2010) proposed a particle-filtering strategy for a planar problem entailing various geometries of a four-element freefield microphone array. The originality of the approach lies in the fact that from a generalized cross-correlation applied to each pair of microphones, a likelihood is defined as a von-Mises-distribution mixture to capture the fact that azimuths are circular random variables. However, when considering an audio sensor mounted on a mobile robotics platform, the robot model and its odometry do not seem to be exploited in the definition of prior dynamics.

Another particle-filtering approach was proposed in the binaural active audio-motor context by Lu and Cooke (2010). In this work, the left and right signals are first processed by gammatone filters, then the ITDs are estimated as the argmax of the sum of the cross-correlations of these filters outputs. The likelihood function of the hidden spatial variables constituting state vectors with respect to ITDs is learnt offline. The localization relies on an *auxiliary-sequencial-importance-resampling* (ASIR) particle filter.

The study is conducted in simulation, and illustrates how the localization performance is

influenced by various movements of the binaural head, such as random walk, correlated random walk, and/or displacement towards the source. However, the position and attitude of the head is required, which, in practice, would require its absolute localization by a system synchronized with audio acquisition. Moreover, the selected particle filtering is subject to caution as the considered practical context represents typical risks for failure of the ASIR particle-filter algorithm, due to noisy prior dynamics and/or sharp likelihood modes.

Considering time delays captured by a moving pair of microphones in freefield, a Gaussian-mixture square-root unscented Kalman filter (GM-SRUKF) was proposed in Portello *et al.* (2011) as an alternative to particle filters for single source localization. The GM-SRUKF scheme can be endowed with self-initialization and ensures consistency of the estimates. In other words, it can prevent overoptimistic posterior covariances.

Importantly, the underlying state-space model derives from a careful analysis of the involved rigid-body motions and acoustic propagation. This work was extended to intermittent moving sources and false measurements by Portello *et al.* (2012). A modified version of the GM-SRUKF was then defined in (Portello *et al.*, 2014) to cope with an artificial head. Therein, the closed-form output equation uniting the sensor-to-source situation to the measured time delay is replaced by an unnormalized Gaussian-mixture curve fitting of the source-azimuth likelihoods as defined in Portello *et al.* (2013) from the channel-time-frequency decomposition of the binaural signals.

Following similar ideas, an active audio-motor speaker-localization scheme was defined in Marković *et al.* (2013) on the basis of a particle filter entailing a curvefitting of source-azimuth likelihoods by a mixture of unnormalized circular von-Mises distributions and wrapped-Cauchy distributions. However, due to the underlying particle-filtering engine, some estimator inconsistencies were noticed.

Besides stochastic filtering, sound-source localization during the robot movement on the basis of a microphone array was performed in Sasaki *et al.* (2010) through triangulation and fused with the movement through the *random-sample-consensus* (RANSAC) method.

5.3.2 Audio SLAM

Although the benefits rendered by the auditory modality when vision fails had been long acknowledged, audio SLAM solutions as proposed in the literature are often incomplete. Some of these approaches are reviewed hereafter.

To start with, some works focus on audio-map building. For instance, Martinson and Schultz (2009) and Sasaki *et al.* (2010) proposed a framework for the exploration of auditory

scenes while estimating the robot position in a known geometric map. A slightly different approach was developed in Kallakuri *et al.* (2013), where the geometric map was computed from a classical SLAM framework. Then, the inverse problem, which aims at localizing the robot itself, can be considered. In Manzanares *et al.* (2011), the robot location was computed by combining the sensing of a permanent sound source – an industrial machine – with prior knowledge about the acoustic environnement.

Some other works deal with sound localization for robots navigation. Among them, Huang *et al.* (1999) designed a real-time sound-localization system on the basis of a bio-inspired model, and used SONAR for obstacle detection. The robot was able to approach the sound source while avoiding obstacles. In Wang *et al.* (2004) 24 microphones were placed on the walls to localize a mobile, talking robot and control its navigation.

Further, Dellaert *et al.* (2003) utilized sounds to gather range measurements between robots, and processed these to solve a range-only SLAM problem. Therein, the robots themselves were used as landmarks. Finally, Hu *et al.* (2011) set a framework for simultaneous localization of a mobile robot and multiple sound sources. One of methods used, based on estimation of the time delays between microphones, was applied to compute a farfield-source direction – which was used as an observation in a bearing-only simultaneous localization and mapping procedure. The problem was solved through the FASTSLAM algorithm (Thrun *et al.*, 2005). Besides, Steckel and Peremans (2013) created a geometry map by combining a biomimetic navigation model and a biomimetic SONAR.

5.3.3 Towards situation-based motion for active-information-based localization

Planned situation-based motions This paragraph describes two represenative approaches where head movements have been planned and executed to improve localization, in other words, reflective motions.

- *Active paradigm of Kumon* et al. *(2010) to improve speech recognition* This motion planning approach enabled a mobile monaural auditory robot to maximize the so-called *confidence measurement* of a speech-recognition system. In other words, based on the assumption that the better the confidence measure the better the speech recognition rate, it was aimed at planning the robot movement in such a way as to improve the recognition accuracy. Interestingly, some ideal listening spots – *"sweet spots"* – were identified in a noisy environment, in the vicinity of which recognition is improved

- *Active paradigm of Martinson* et al. *(2011) to improve sound-source localization accuracy* Instead of planning microphone movements inside the environment, this contribution proposed to dynamically modify their positions. It was shown that

significant improvements could be obtained by optimizing the relative microphone positions through an attraction/repulsion model

Situation-based reflex motions Up to our knowledge, no methodology has yet been developed for active/information-based situation-based reflexive motions. Ongoing work is being developed by LAAS-CNRS – not funded by TWO!EARS. The results will be open to TWO!EARS when available.

5.3.4 Sensor-based reflex motions and actively reconfigurable sensors

In Sec. 5.2, it was stated that sensor-based motion control is a valuable and widespread alternative to situation-based approaches, at least in vision. However, in audition, approaches that recover the world geometry look more pervasive. Nevertheless, two relevant contributions can be listed.

- A sensor-based solution, termed *audio servo* was proposed in Kumon *et al.* (2003) on the basis of raw audio information preprocessed by a digital signal processor
- Reconfigurable sensors were also addressed in Kumon and Noda (2011). In this project, the design of an active soft pinna, in analogy to those of cats, was proposed. The system is able to move or deform the pinna shape dynamically. Further, some fundamental characteristics such as the influence of materials on the directivity of the pinna were studied

5.3.5 Other sensorimotor feedback in robot audition

In robotics, perception has for long been considered a pure bottom-up process, so that actions would be the result of sensory analysis only. This historical viewpoint on perception is being questioned, all the more because the exploratory abilities of robotics platforms can be exploited to improve the analysis and understanding of the environment. For example, an active strategy for auditory-space learning with application to sound-source localization has been proposed by Bernard *et al.* (2012), Bernard (2014).

Such strategies rely on a general theoretical approach to perception, widely known as *sensorimotor-contingencies theory* (O'Regan and Noe, 2001, Philipona *et al.*, 2003). This strategy builds on the notion that action is to be envisaged at the same level as perception. In other words, action and perception interact in forming an internal representation of the auditory space. As a first step, an active hearing process is performed while the learning an auditory-motor map. Next, this map is used for *a-priori* passive sound localization.

Let us depict all kinds of environment, motor states and sensations that an agent can consider as the respective manifolds, $\mathcal{E}$, $\mathcal{M}$, and $\mathcal{S}$. A sensory state, $s \in \mathcal{S}$, is given as a function of the current motor and environment states, $m \in \mathcal{M}$, and $e \in \mathcal{E}$, through a sensorimotor law, Φ, so that $s = \Phi(m, e)$. e models the spatial and spectral properties so the acoustic scenen and the sound sources in it. m models the agent's body configuration whereas Φ represents the body-environment interactions and neural processing giving rise to the sensation, s. Moreover the sensory space, $\mathcal{S}$, lies on a low-dimensional manifold the topology of which is similar to the embodying space and, consequently, the learning of spacial perception becomes the learning of such a manifold. Such a process has been applied to auditory-space learning using nonlinear dimensionality-reduction techniques (Aytekin *et al.*, 2008)(Deleforge and Horaud, 2011).

Classical localization methods express a source location in terms of angle or range in an Euclidean physical space. In contrast, the sensorimotor approach directly links perception and action in an internal representation of space. There, a spatial position is directly expressed as a motor state and, as such, it does not implies any notion of space. Given a motor space, $\mathcal{M}$, and an environment state, $e \in \mathcal{E}$, the source localization problem can thus be defined as the estimation of the motor state, $\tilde{m}$, as follows.

$$\tilde{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} |\Phi(m, e) - \Phi(m_0, e_0)|, \tag{5.1}$$

where $|.|$ denotes a distance metric and $\Phi(m_0, e_0)$ represents a reference sensory state that has to be approximated.

In the case of sound source localization, $\Phi(m_0, e_0)$ correspond to a source localized in front of the listener with the head in the rest position, that is, the most obvious case of azimuthal localization.

A simple behavior enabling head-to-source orientation can be implemented from ILD cues as follows. Once a sound is perceived, the agent orients its head toward the loudest side, while the ILD is non-zero. Once this behavior is completed, that is, when the ILD reaches 0, the head of the agent is oriented towards the sound source. This active hearing process allows for *a-posteriori* localization, with $\tilde{m}$ being given after motion as the difference between the initial and final motor states. Such evoked behavior, linking the initial sensory state in $\mathcal{S}$ to the final motor state in $\mathcal{M}$, provides the sensorimotor association required for an *a-priori* passive localization.

Figure 5.3 shows an auditory space representation after learning of high dimensional ILD cues from 1000 auditory stimuli. Each point, corresponding to a different sensory state, is associated with its localization estimate, $\tilde{m}$, computed from the orientation behavior.

After learning of such an association, it becomes possible to localize new percepts based

on neighborhood relationships. Suppose a new stimulus corresponding to a sensory state, $s \in \mathcal{S}$, perceived by the agent. s is firstly projected in the sensory-space representation and, if this projection has close neighbors – s_1 in Fig. 5.3 – its corresponding motor state, $\tilde{m}$, is interpolated from the neighborhood, giving a passive localization estimation. If the projection is outlying in an area with no neighbors – s_2 in Fig. 5.3 – this sensory state is not yet represented and $\tilde{m}$ can not be estimated passively. In this case the orienting behavior is executed, giving an active estimate of $\tilde{m}$.

This method requires almost no *a priori* knowledge on either the agent or the environment. It mainly depends on the knowledge of the auditory-space-representation dimension, which is typically 2D or 3D, and on a dimension-reduction technique that is robust enough to estimate the non-linear embedding of complex environments in an efficient "hard-wired" evoked behavior.

The sensorimotor theory (Poincaré, 1945, O'Regan and Noe, 2001) claims that the brain is initially a naive agent that interacts with the world via an unknown set of afferent and efferent connexions, with no a priori knowledge about its own motor capacities or the space it is immersed in. The agent therefore extracts this knowledge by analyzing the consequences of its own movements on its sensory perceptions, building a sensorimotor representation of its embodying space. Consequnetly, the perceptive capabilities of robots might be based on sensorimotor flow analysis, that is, on analysis of the sensory consequences of its own actions. The aim is then to extract sensorimotor contingencies that depict the interaction capabilities of the robots with respect to its environment. Importantly, such contingencies

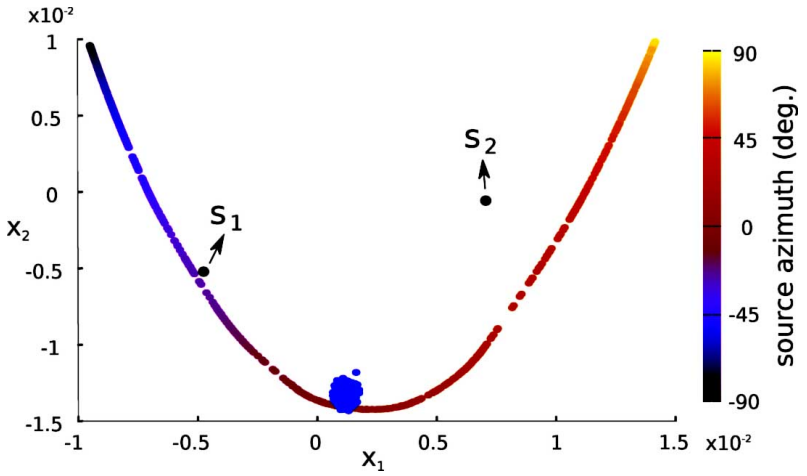


Figure 5.3: Two-dimensional manifold of auditory space, learned from a set of 1000 sensory states obtained *before* orienting behavior (parabolic curve). Each state corresponds to a sound source of random azimuth in $[-90°, 90°]$. See also projections on the manifold of sensory states obtained *after* the orienting behavior approximating the reference state $\Phi(m_0, e_0)$ (cluster of points close to $x_1 = 0$). Further, new percepts, such as s_1 and s_2, can be localized on the manifold

need not necessarily be given *a priori* to the robots by the engineers, but can as well be discovered/learned autonomously by the robots themselves.

The design of robotic systems is largely dictated by purely human intuition about how we perceive the world. In order to develop truly autonomous robots, it is necessary to step away from this intuition and let robotic agents develop their own way of perceiving. Robots should start from scratch and gradually develop perceptual notions, under no prior assumptions but exclusively by looking into its sensorimotor experience and identifying repetitive patterns and invariants.

As one of the most fundamental perceptual constructs, *space*, cannot be an exception to this requirement. Recently, it has been showen (Laflaquiere, 2013, Laflaquiere *et al.*, 2015) that the notion of space as environment-independent cannot be deduced solely from exteroceptive information, as it is highly variable and mainly determined by the contents of the environment.

Yet, the environment-independent definition of space can be approached by looking into the functions that link the motor commands to changes in exteroceptive inputs. A redundant robotic arm – Fig. 5.4 – has been simulated with a retina installed at its end-point, and showing how such an agent can learn the configuration space of its retina. The resulting manifold has the topology of the Cartesian product of a plane and a circle, and corresponds to the planar position and orientation of the retina.

The results of this work highlight the fact that the approach entirely relies on the properties of the raw sensorimotor flow. On the contrary, most body-schema-acquisition studies hypothesize some spatial knowledge provided *a-priori* to the robot via sensory-input pre-processing or through additional knowledge of the agent's structure.

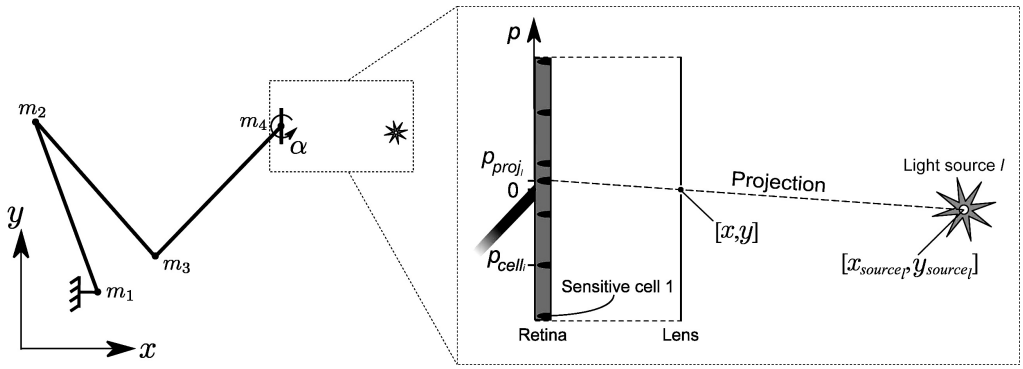


Figure 5.4: Agent being a three-segment arm equipped with four motors and a light-sensitive retina (left). The retina is regularly covered with six light-sensitive cells. Light from the light sources is projected onto the retina through a pinhole lens. Each cell's excitation is a Gaussian function of its distance to the light projection

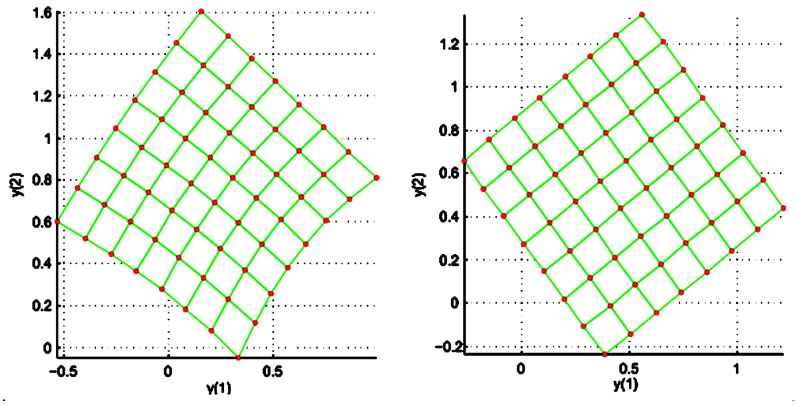


Figure 5.5: Examples of the internal map of the external configuration of the terminal point of the arm obtained by the sensorimotor algorithm in (Laflaquiere, 2013)

Although the notion of space is intuitively associated with visual inputs, and more generally with exteroception, it has been shown that considering them exclusively cannot lead to an environment-independent notion. The authors thus propose to look for spatial properties not in the exteroceptive flow of the agents but in the structure of their sensorimotor interactions with the world. From the agents' point of view, the sensorimotor laws are just functions that map the motor outputs to sensory inputs, and both the properties of the external world and of its own properties are simply constraints on the shape of these functions.

In particular, the structure of the external space as such manifests itself through the constraint that certain properties of these functions should not depend on the objects present in the environment. It is suggested that the discovery of such constraints can lead the agents to discover the structure of external space. The open issue is then to show how external objects in the world can be expressed as sensorimotor laws, just as it is done for space description.

5.4 Conclusions with respect to Two!Ears

From the literature on sensorimotor feedback, the following suggestions can be inferred with respect to Two!Ears, among others.

- Sensorymotor feedback at the reflexive level, such as the *"turn-to-reflex"* of the head, can help to disambiguate already at the sensoric level, such as solving front-back confusion and/or elevation estimation of sound sources

- By solving ambiguities in the recognition process at a low level and in a "hardwired" manner, sensorimotor feedback has the potency of speeding-up auditory scene analysis substantially.
- Sensorimotor feedback at the reflexive level, that is, with participation of cognitive stages of the model system, can control the sensor positions and, thus, guide these and/or even the complete robots along paths and into positions where they can perform their respective taks in an optimal way

6 Conclusion

In the Two!Ears project, the inclusion of various feedback loops in a comprehensive computational model of the binaural auditory system is a prominent feature. Although there is physiological evidence that, in the human auditory system, feedback between all stages of the system is possible – at least in in principle – we have restricted ourselfes within Two!Ears to feedback loops that have a likely functional relevance for the model system at large.

See Chap. 1 for a complete list of all considered feedback loops as had been set up for the proposal. An updated list, which reflects the current discussions within the consortium at the end of the first project year, is provided in the attached document, D4.1 Part C.

The current document reports on a literature survey regarding the foundations of four feedback mechanisms, namely, *head-and-body-movement, attention-driven, olivocochlear* and *sensorimotor* feedback.

Some hints regarding possible opportunities of exploiting the knowledge gained from this survey for the Two!Ears model system are listed at the end of each chapter.

Further details, particularily, with regard to feedback mechanisms that have already been tackled in the current implementation process of the Two!Ears system, are provided in the attached document "Supporting information on feedback in Two!Ears"– D4.1 Part D.

Bibliography

Ahissar, M. and Hochstein, S. (**2004**), "The reverse hierarchy theory of visual perceptual learning." *Trends in Cognitive Sciences* **8**(10), pp. 457–64, URL `http://www.ncbi.nlm.nih.gov/pubmed/15450510`. (Cited on pages 12 and 13)

Aldebaran (**2014**), "The Nao robot," URL `http://www.aldebaran.com/en/humanoid-robot/nao-robot`. (Cited on page 55)

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (**1988**), "Active vision," *Intl. J. Computer Vision* **1**(4), pp. 333–356. (Cited on page 69)

Arleo, A. and Gerstner, W. (**2000**), "Modeling rodent head-direction cells and place cells for spatial learning in bio-mimetic robotics," in *Sixth Intl. Conf. Simulation of Adaptive Behaviour*. (Cited on page 25)

Asoh, H., Asano, F., Yoshimura, T., Yamamoto, K., Motomura, Y., Ichimura, N., Hara, I., and Ogata, J. (**2004**), "An application of a particle filter to Bayesian multiple sound source tracking with audio and video information fusion," in *Intl. Conf. Information Fusion (FUSION'2004)*. (Cited on page 75)

Assmann, P. F. and Summerfield, Q. (**2004**), "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. Ainthworth, A. Popper, and R. R. Fay, Springer, New York NY, pp. 231–308. (Cited on page 49)

Aytekin, M., Moss, C., and Simon, J. (**2008**), "A sensorimotor approach to sound localization," *Neural Computation* **20**. (Cited on page 79)

Baranes, A. and Oudeyer, P. (**2009**), "R-IAC: Robust intrinsically motivated exploration and active learning," *IEEE Trans. Autonomous Mental Development* **1**(3), pp. 155–169. (Cited on pages 30 and 32)

Baranes, A. and Oudeyer, P. (**2010**), "Intrinsically motivated goal exploration for active motor learning in robots: A case study," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2010)*, pp. 1766–1773. (Cited on pages 30, 32, 33, 34, 35, and 40)

Barto, A. G., Singh, S., and Chentanez, N. (**2004**), "Intrinsically motivated learning

of hierarchical collections of skills," in *Intl. Conf. Developmental Learning (ICDL)*. (Cited on page 32)

Bear, M. F., Connors, B. W., and Paradiso, M. A. (**2007**), *Neuroscience: Exploring the brain (3rd ed.)*., Lippincott Williams\&Wilki. (Cited on page 53)

Beck, D. and Kastner, S. (**2009**), "Top-down and bottom-up mechanisms in biasing competition in the human brain," *Vision Res.* **49**, pp. 1154–1165. (Cited on page 44)

Begum, M. and Karray, F. (**2011**), "Visual attention for robotic cognition: A survey," in *IEEE Trans. Autonomous Mental Development*, vol. 3, pp. 92–105. (Cited on page 51)

Berlyne, D. E. (**1950**), "Novelty and curiosity as determinants of exploratory behavior," *British J. Psychology* **41**(1-2), pp. 68–80. (Cited on pages 30, 33, 36, and 40)

Berlyne, D. E. (**1965**), *Structure and direction in thinking*, New York: John Wiley & Sons. (Cited on pages 30 and 40)

Bernard, M. (**2014**), "Audition active et intégration sensorimotrice pour un robot autonome bioinspiré," Ph.D. thesis, UPMC, Paris, France. (Cited on page 78)

Bernard, M., Pirim, P., de Cheveigné, A., and Gas, B. (**2012**), "Sensorimotor learning of sound localization from an auditory evoked behavior," in *IEEE Intl. Conf. Robotics and Automation (ICRA'2012)*, pp. 91–96. (Cited on page 78)

Blair, H. T. and Sharp, P. E. (**1995**), "Anticipatory head direction signals in anterior Thalamus: Evidence for a thalamocortical circuit that integrates angular head motion to compute head direction," *J. Neuroscience* **15**(9), pp. 6260–6270. (Cited on page 9)

Blauert, J. (**1974, 2nd ed.1997**), *Spatial Hearing - the psychophysics of human sound localization*, MIT Press, Cambridge MA. (Cited on pages 14 and 46)

Blauert, J. and Braasch, J. (**2005**), "Acoustical communication: the Precedence Effect," in *Proc. Forum Acusticum, Budapest*, OPAKFI, H–Budapest, pp. 15–19. (Cited on page 46)

Blauert, J. and Jekosch, U. (**2012**), "A layer model of sound quality," *J. Audio Engr. Soc.* **60**, pp. 4–12. (Cited on page 48)

Bourgault, F., Makarenko, A., Williams, S., Grocholsky, B., and Durrant-Whyte, H. (**2002**), "Information based adaptive robotic exploration," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2002)*, EPFL Lausanne, Swiherland. (Cited on page 73)

Braitenberg, V. (**1986**), *Vehicles: Experiments in synthetic psychology*, A Bradford Book.

(Cited on page 68)

Bregman, A. S. (**1991**), "Using quick glimpses to decompose mixtures," in *Music, language, speech and brain*, edited by J. Sundberg, L. Nord, and R. Carlson, MacMillan, UK–London, pp. 284–289. (Cited on page 50)

Bregman, A. S. (**1993**), "Auditory scene analysis: Hearing in complex environments," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigang, Clarendon Press, UK-Oxford, chap. 2, pp. 10–36. (Cited on page 50)

Broadbent, D. E. (**1958**), *Perception and communication*, Pergamon Press, GB–Oxford. (Cited on page 45)

Brockhaus Konversationslexikon (**1869**), "entry: Aufmerksamkeit (attention)," URL `http://www.retrobibliothek.de/retrobib/seite.html?id=121408`. (Cited on page 43)

Brooks, R. (**1991**), "Intelligence without representation," *Artificial Intelligence* **47**, pp. 139–159. (Cited on page 68)

Brown, G. J., Ferry, R. T., and Meddis, R. (**2010**), "A computer model of auditory efferent suppression: Implications for the recognition of speech in noise," *J. Acoust. Soc. Am.* **127**(2), pp. 943–954. (Cited on pages 63, 64, and 65)

Brown, M., Venecia, R., and Guinan, J., J.J. (**2003**), "Responses of medial olivocochlear neurons," *Exp.Brain Res.* **153**(4), pp. 491–498, URL `http://dx.doi.org/10.1007/s00221-003-1679-y`. (Cited on page 58)

Brown, M. A. and Sharp, P. E. (**1995**), "Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and the nucleus accumbens," *Hippocampus* **5**(3), pp. 171–188. (Cited on page 7)

Bundesen, C. and Habekost, T. (**2008**), *Principles of visual attention*, Oxford Univ. Press, GB–Oxford. (Cited on page 44)

Canudas de Wit, C., Khennouf, H., Samson, C., and Sordalen, O. (**1993**), "Nonlinear control design for mobile robots," in *Recent Trends in Mobile Robots*, edited by Y. F. Zheng, World Scientific, World Scientific Series in Robotics and Automated Systems, Vol. 11. (Cited on page 71)

Capdepuy, P., Polani, D., and Nehaniv, C. L. (**2007**), "Maximization of potential information flow as a universal utility for collective behaviour," in *IEEE Symp. Artificial Life*, pp. 207–213, URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4218888`. (Cited on page 31)

Carrasco, M. (**2011**), "Visual attention: The past 25 years," *Vision Res.* **51**, pp. 1484–1525.

(Cited on page 44)

Chaumette, F. and Hutchinson, S. (**2006**), "Visual servo control, Part I: Basic approaches," *IEEE Robotics and Automation Mag.* **13**(4), pp. 82–90. (Cited on page 72)

Chaumette, F. and Hutchinson, S. (**2007**), "Visual servo control, Part II: Advanced approaches," *IEEE Robotics and Automation Mag.* **14**(1), pp. 109–118. (Cited on page 72)

Cherry, E. C. (**1953**), "Some experiments upon the recognition of speech with one and two ears," *J. Acoust. Soc. Am.* **25**, pp. 975–979. (Cited on page 49)

Cherry, E. C. (**1954**), "Some further experiments upon the recognition of speech with one and two ears," *J. Acoust, Soc. Am.* **26**, pp. 554–559. (Cited on page 49)

Cherubini, A. and Chaumette, F. (**2013**), "Visual navigation of a mobile robot with laser-based collision avoidance," *Intl. J. Robotics Research* **32**(2), pp. 189–205. (Cited on pages 69 and 73)

Chevalier, G., Vacher, S., Deniau, J., and Desban, M. (**1985**), "Disinhibition as a basic process in the expression of striatal functions. I. The striato-nigral influence on tecto-spinal/tecto-diencephalic neurons," *Brain Research* **334**, pp. 215–226. (Cited on page 10)

Chiu, H., Williams, S., Dellaert, F., Samarasekera, S., and Kumar, R. (**2013**), "Robust vision-AAided navigation using sliding-window factor graphs," in *IEEE Intl. Conf. Robotics and Automation (ICRA'2013)*. (Cited on page 71)

Clark, N. R., Brown, G. J., Jürgens, T., and Meddis, R. (**2012**), "A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise," *J. Acoust. Soc. Am.* **132**(3), pp. 1535–1541. (Cited on page 64)

Cooke, M. (**1993**), *Modelling auditory processing and organization*, Cambridge Univ. Press. (Cited on page 51)

Cooke, M. (**2003**), "Glimpsing speech," *J. Phonetics* **31**, pp. 579–584. (Cited on page 50)

Cooke, M., Lu, Y., Lu, Y., and Horaud, R. (**2007**), "Active hearing, active speaking," in *Intl. Symp. Auditory and Audiological Research (ISAAR'2007)*, Marienlyst, Helsigør, Denmark. (Cited on page 69)

Cools, A. R. (**1980**), "Role of the neostriatal dopaminergic activity in sSequencing and selecting behavioural strategies: Facilitation of processes involved in selecting the best strategy in a stressful situation," *Behavioural Brain Research* **1**, pp. 361–378. (Cited on page 10)

Cooper, N. P. and Guinan, J. J. J. (**2003**), "Separate mechanical processes underlie fast and slow effects of medial olivocochlear efferent activity." *J. Physiology* **548**(Pt 1), pp. 307–312. (Cited on page 59)

Cover, T. and Thomas, J. (**1991**), *Elements of Information Theory*, Wiley. (Cited on page 73)

Cuperlier, N., Quoy, M., and Gaussier, P. (**2007**), "Neurobiologically inspired mobile robot navigation and planning," *Frontiers in Neurorobotics* . (Cited on pages 7, 22, 23, 24, and 25)

Cuperlier, N., Quoy, M., Giovannangeli, C., Gaussier, P., and Laroque, P. (**2006**), "Transition cells for navigation and planning in an unknown environment," *From Animals to Animats* **4095**, pp. 286–297. (Cited on pages 8 and 22)

Dai, P., Scharf, B., and Buus, S. (**1991**), "Effective attenuation of signals in noise under focused attention," *J. Acoust. Soc. Am.* **89**, pp. 2837–2842. (Cited on page 47)

Darrow, K. N., Maison, S. F., and Liberman, M. C. (**2006**), "Cochlear efferent feedback balances interaural sensitivity," *Nature: Neuroscience* **9**(12), pp. 1464–1476. (Cited on page 61)

Darwin, C. J. (**2007/2010**), "Listening to speech in the presence of other sounds," in *The Perception of Speech: From Sound to Meaning*, edited by B. Moore, L. Tyler, and W. Marslen-Wilson, Oxford univ. press, UK–Oxford, chap. 7, pp. 151–170, 1st published online in 2007. (Cited on page 49)

de Boer, J. and Thornton, A. R. D. (**2007**), "Effect of subject task on contralateral suppression of click evoked otoacoustic emissions." *Hearing Res.* **233**(1-2), pp. 117–123. (Cited on page 62)

de Boer, J., Thornton, A. R. D., and Krumbholz, K. (**2012**), "What is the role of the medial olivocochlear system in speech-in-noise processing?" *J. Neurophysiology* **107**(5), pp. 1301–1312. (Cited on page 60)

Deleforge, A. and Horaud, R. (**2011**), "Learning the direction of a sound source using head motions and spectral features," Tech. Rep. 7529, INRIA. (Cited on page 79)

Dellaert, F., Alegre, F., and Martinson, E. (**2003**), "Intrinsic localization and mapping with 2 applications: Diffusion mapping and Marco Polo localization," in *IEEE Intl. Conf. Robotics and Automation (ICRA'2003)*, vol. 2, pp. 2344–2349. (Cited on page 77)

Deniau, J. and Chevalier, G. (**1985**), "Disinhibition as a basic process in the expression of striatal functions. II. The striato-nigral influence on thalamocortical cells of the ventromedial thalamic nucleus," *Brain Research* **334**, pp. 227–233. (Cited on page 10)

Denzler, J. and Brown, C. (**2002**), "Information-theoretic sensor data selection for active object recognition and state estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence* , pp. 145–157. (Cited on page 74)

Deutsch, J. A. and Deutsch, D. (**1963**), "Attention: Some theoretical considerations," *Psychol. Rev.* **40**, pp. 80–90. (Cited on page 45)

Diehl, R. L., Lotto, A. J., and Holt, L. L. (**2004**), "Speech perception," *Ann. Rev. Psychology* **55**, pp. 149–179. (Cited on page 49)

Dolan, D. and Nuttall, A. (**1988**), "Masked cochlear whole-nerve response intensity functions altered by electrical-stimulation of the crossed olivocochlear bundle," *J. Acoust. Soc. Am.* **83**(3), pp. 1081–1086. (Cited on pages 59 and 64)

Durand Petiteville, A., Courdesses, M., Cadenat, V., and Baillion, P. (**2010**), "On-line estimation of the reference visual features. application to a vision based long range navigation task," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2010)*, Taipei, Taiwan. (Cited on page 73)

Durrant-Whyte, H. and Bailey, T. (**2006**), "Simultaneous localization and mapping: Part I," *IEEE Robotics and Automation Magazine* **13**(2), pp. 99–110. (Cited on page 70)

EARS (**2014**), "Embodied audition for robots," URL `http://robot-ears.eu/`. (Cited on page 55)

Elfes, A. (**1989**), "Using occupancy grid for mobile robot perception and navigation," *Computer* **22**(6), pp. 46–57. (Cited on page 37)

Engelke, U., H., K., Zepernick, H.-J., and Ndjiki-Nya, P. (**2011**), "Visual attention," *IEEE Sign. Proc. Mag.* , pp. 50–59. (Cited on pages 44 and 45)

Espiau, B., Chaumette, F., and Rives, P. (**1992**), "A new approach to visual servoing in robotics," *IEEE Trans. Robotics Automation* **8**(3), pp. 313–326. (Cited on pages 69 and 72)

Evers, C., Moore, A. H., and Naylor, P. A. (**2014**), "EARS – Embodied audition for robots," URL `http://robot-ears.eu/wp-content/uploads/Poster_UKSpeech_IMPERIAL_09_06_2014.pdf`. (Cited on page 55)

Fabre-Thorpe, M. (**2003**), "Visual categorization: accessing abstraction in non-human primates." *Philosophical Trans. Royal Society London. Series B : Biological Sciences* **358**, pp. 1215–1223. (Cited on page 51)

Feder, H., Leonard, J., and Smith, C. M. (**1999**), "Adaptive mobile robot navigation and mapping," *Intl. J. Robotics Research* **18**, pp. 650–668. (Cited on page 73)

Ferry, R. T. (**2008**), "Auditory processing and the medial olivocochlear efferent system," Ph.D. thesis, University of Essex, Colchester, UK. (Cited on page 65)

Ferry, R. T. and Meddis, R. (**2007**), "A computer model of medial efferent suppression in the mammalian auditory system," *J. Acoust. Soc. Am.* **122**(6), pp. 3519–3526. (Cited on pages 63 and 64)

Fletcher, H. (**1940**), "Auditory pattern," *Rev. Modern Physics* **12**, pp. 47–65. (Cited on page 47)

Folio, D. (**2007**), "Stratégies de commande référencées multi-capteurs et gestion de la perte du signal visuel pour la navigation d'un robot mobile," Ph. d. thesis, Université Paul Sabatier, LAAS, LAAS , Toulouse, France. (Cited on page 71)

Frintrop, S. (**2006**), *VOCUS: A visual attention system for object detection and goal-directed search*, vol. 3899, Springer. (Cited on page 51)

Frintrop, S., Rome, E., and Christensen, H. I. (**2010**), "Computational visual attention systems and their cognitive foundations," . (Cited on pages 51 and 52)

Furukawa, K., Okutani, K., Nagira, K., Otsuka, T., Itoyama, K., Nakadai, K., and Okuno, H. (**2013**), "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2013)*, pp. 3943–3948. (Cited on page 74)

Gaussier, P., Revel, a., Banquet, J. P., and Babeau, V. (**2002**), "From view cells and place cells to cognitive map learning: processing stages of the hippocampal system." *Biological Cybernetics* **86**(1), pp. 15–28, URL `http://www.ncbi.nlm.nih.gov/pubmed/11918209`. (Cited on page 8)

Ghitza, O., Messing, D., Delhorne, L., Braida, L., Bruckert, E., and Sondhi, M. (**2006**), "Towards predicting consonant confusions of degraded speech," Cloppenburg, Germany. (Cited on pages 63 and 64)

Goldstein, J. L. (**1990**), "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," *Hearing Research* **49**, pp. 39–60. (Cited on page 63)

Green, D. and Swets, J. A. (**1966**), *Signal detection theory and psychophysics*, John Wiley and Sons, New York NY. (Cited on page 47)

Greene (**1960**), "Psychoacoustics and detection theory," *J. Acoust. Soc. Am.* **32**, pp. 1189–1203. (Cited on page 47)

Gregg, M. K. and Samuel, A. G. (**2008**), "Attention and the organizational proper-

ties of sound," *J. Exp. Pychol.: Human Perception and Performance* **5**, pp. 168–175. (Cited on page 48)

Grey, W. (**1950**), "An imitation of life," *Scientific American* , pp. 42–45. (Cited on page 67)

Grey, W. (**1951**), "A machine that learns," *Scientific American* , pp. 60–63. (Cited on page 67)

Grocholsky, B. (**2006**), "Information theoretic control of multiple sensor platforms," Ph.D. thesis, University of Sydney. (Cited on page 74)

Grocholsky, B., Makarenko, A., and Durrant-Whyte, H. (**2003**), "Information-theoretic coordinated control of multiple sensor platforms," in *IEEE Intl. Conf. Robotics and Automation (ICRA'2003)*, Taipei, Taiwan. (Cited on page 73)

Guinan, J. J. (**1996**), "Physiology of olivocochlear efferents," in *The Cochlea*, edited by P. Dallos, A. N. Popper, and R. R. Fay, Springer-Verlag, pp. 435–502. (Cited on page 57)

Guinan, J. J. (**2010**), "Cochlear efferent innervation and function," *Current Opinion Otolaryngol Head Neck Surgery* **18**(5), pp. 447–453. (Cited on pages 57 and 61)

Guinan, J. J. (**2014**), "Olivocochlear efferent function: issues regarding methods and the interpretation of results," *Frontiers in System Neuroscience* **8**, pp. 1–5. (Cited on page 57)

Guinan, J. J. and Stankovic, K. M. (**1996**), "Medial efferent inhibition produces the largest equivalent attenuations at moderate to high sound levels in cat auditory-nerve fibers," *J. Acoust. Soc. Am.* **100**, pp. 1680–1690. (Cited on page 64)

Gurney, K., Prescott, T. J., and Redgrave, P. (**2001**a), "A computational model of action selection in the basal ganglia. I. A new functional anatomy." *Biological Cybernetics* **84**(6), pp. 401–10, URL `http://www.ncbi.nlm.nih.gov/pubmed/11417052`. (Cited on pages 10 and 27)

Gurney, K., Prescott, T. J., and Redgrave, P. (**2001**b), "A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour." *Biological Cybernetics* **84**(6), pp. 411–23, URL `http://www.ncbi.nlm.nih.gov/pubmed/11417053`. (Cited on page 27)

Hafter, E. R., Sarampalis, A., and Loui, P. (**2008**), "Auditory attention and filters," in *Auditory perception of sound sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay, Springer, Berlin–Heidelberg–New York, pp. 115–142. (Cited on pages 44 and 47)

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (**2005**), "Microstructure of a spatial map in the entorhinal cortex," *Nature* **436**, pp. 801–806. (Cited on page 10)

Hampson, R. E., Heyser, C. J., and Deadwyler, S. A. (**1993**), "Hippocampal cell firing correlates of delayed match-to-sample performance in the rat." *Behavioral Neuroscience* **107**(5), pp. 715–739. (Cited on page 8)

Harkrider, A. W. and Bowers, C. D. (**2009**), "Evidence for a cortically mediated release from inhibition in the human cochlea." *J. Am. Acad. Audiology* **20**(3), pp. 208–215. (Cited on page 62)

He, J. and Yu, Y. (**2009**), "Role of the descending control in the auditory pathway," in *Oxford Handb. of Auditory Science,*, edited by A. Rees and A. R. Palmer, Oxford Univ. Press„ New York, NY, vol. 2: The auditory brain, pp. 247–268. (Cited on page 1)

Hirst, W., Spelke, E., Reaves, C. C., Caharack, G., and Neisser, U. (**1980**), "Dividing attention with alternation or automaticity," *J. Exper. Psychol.:General* **109**, pp. 98–117. (Cited on page 45)

Hok, V., Save, E., and Poucet, B. (**2005**), "Coding for spatial goals in the prelimbic-infralimbic," *PNAS* **102**(12), pp. 4602–4607. (Cited on page 7)

Hörnstein, J., Lopes, M., Santos-Victor, J., and Lacerda, F. (**2006**), "Sound localization for humanoid robots - Building audio-motor maps based on the HRTF," in *IEEE Intl. Conf. Intell. Robots and Systems*, pp. 1170–1176. (Cited on page 52)

Hu, J., Chan, C., Wang, C., Lee, M., and Kuo, C. (**2011**), "Simultaneous localization of a mobile robot and multiple sound sources using a microphone array," *Advanced Robotics* **25**, pp. 135–152. (Cited on page 77)

Huang, J., Supaongprapa, T., Terakura, I., Wang, F., Ohnishi, N., and Sugie, N. (**1999**), "A model-based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems* , pp. 199–209. (Cited on page 77)

Huang, X. and Weng, J. (**2002**), "Novelty and reinforcement learning in the value system of developmental robots." in *Lund University Cognitive Studies.* (Cited on page 31)

Hutchinson, S., Hager, G., and Corke, P. (**1996**), "A tutorial on visual servo control," *IEEE Trans. Robotics and Automation* **12**(5), pp. 651–670. (Cited on page 72)

Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., and Imura, J. (**2009**), "Ego noise suppression of a robot using template subtraction," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems, (IROS'2009)*, pp. 199–204. (Cited on page 74)

Ince, G., Nakadai, K., Rodemann, T., Imura, J., Nakamura, K., and Nakajima, H. (**2011**), "Incremental learning for ego noise estimation of a robot," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2011)*, pp. 131–136. (Cited on page 74)

Itti, L. and Baldi, P. (**2009**), "Bayesian surprise attracts human attention," *Vision Research* **49**, pp. 1295–1306. (Cited on pages 46 and 52)

Jarvis, R. A. and C., B. (**1986**), "Robot navigation: Touching, seeing and knowing," in *Australian Conf. on Artificial Intelligence*, Melbourne. (Cited on page 37)

Jeffery, K. J. and Hayman, R. (**2004**), "Plasticity of the hippocampal place cell representation," *Rev. Neuroscience* **15**, pp. 309–331. (Cited on page 8)

Jekosch, U. (**2005**), "Assigning meaning to sounds: Semiotics in the context of product-sound design," in *Communication Acoustics*, edited by J. Blauert, Springer, D–Berlin, Heidelberg, chap. 8, pp. 193–221. (Cited on pages 48 and 50)

Jennings, S. G., Heinz, M. G., and Strickland, E. A. (**2011**), "Evaluating adaptation and olivocochlear efferent feedback as potential explanations of psychophysical overshoot," *J. Assoc. Res. Otolaryngology* **12**(3), pp. 345—360. (Cited on pages 63 and 65)

Jennings, S. G., Strickland, E. A., and Heinz, M. G. (**2009**), "Precursor effects on behavioral estimates of frequency selectivity and gain in forward masking," *J. Acoust. Soc. Am.* **125**(4), pp. 2172—2181. (Cited on page 64)

Johnston, J. C. and McCann, R. (**2006**), "On the locus of dual-task interference: Is there a bottleneck at the stimulus classifications stage?" *Quarterly J. Exper. Psychology* **59**, pp. 694–719. (Cited on page 45)

Jones, M. R. and Yee, W. (**1993**), "Attending to auditory events: The role of temporal organization," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigand, Clarendon Press, UK–Oxford, chap. 4, pp. 69–112. (Cited on pages 44 and 48)

Julian, B., Karaman, S., and Rus, D. (**2013**), "On mutual information-based control of range sensing robots for mapping applications," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2013)*, Tokyo, Japan. (Cited on page 73)

Kallakuri, N., Even, J., Morales, Y., Ishi, C., and Hagita, N. (**2013**), "Probabilistic approach for building auditory maps with a mobile microphone array," in *IEEE Intl. Conf. Robotics and Automation (ICRA'2013)*, Karlsruhe, Germany. (Cited on page 77)

Kawase, T., Delgutte, B., and Liberman, M. C. (**1993**), "Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones." *J. Neurophysiology* **70**(6), pp. 2533–2549. (Cited on page 59)

Kim, C., Mason, R., Member, A. E. S., and Brookes, T. (**2013**), "Head movements made by listeners in experimental and real-life listening activities," *J. Audio Engineering Soc.* **61**, pp. 425–438. (Cited on pages 18 and 19)

Kitano, H., Okuno, H., Nakadai, K., Sabisch, T., and Matsui, T. (**2000**), "Design and architecture of SIG the humanoid: an experimental platform for integrated perception in RoboCup humanoid challenge," *IEEE/RSJ Intl. Conf. Intell. Robots and Systems* **1**, pp. 181–190 vol.1. (Cited on page 53)

Kropotov, J. D. and Etlinger, S. C. (**1999**), "Selection of actions in the basal ganglia-thalamocortical circuits: review and model." *Intl. J. Psychophysiology* **31**(3), pp. 197–217, URL `http://www.ncbi.nlm.nih.gov/pubmed/10076774`. (Cited on page 10)

Kubie, J. L. and Ranck, J. (**1983**), "Sensory-behavioral correlates in Iidividual hippocampus neurons in three situations: Space and context." in *Neurobiology of the Hippocampus*, Seifert W, ed., pp. 433–447. (Cited on page 8)

Kuehn, B., Schauerte, B., Kroschel, K., and Stiefelhagen, R. (**2012**), "Multimodal saliency-based attention: A lazy robot's approach," in *IEEE Intl. Conf. Intell. Robots and Systems*, pp. 807–814. (Cited on page 52)

Kumar, U. A. and Vanaja, C. S. (**2004**), "Functioning of olivocochlear bundle and speech perception in noise." *Ear Hear* **25**(2), pp. 142–146. (Cited on page 60)

Kumon, M., Fukushima, K., Kunimatsu, S., and Ishitobi, M. (**2010**), "Motion planning based on simultaneous perturbation stochastic approximation for mobile auditory robots," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2010)*, pp. 431–436. (Cited on page 77)

Kumon, M. and Noda, Y. (**2011**), "Active soft pinnae for robots," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2011)*, pp. 112–117. (Cited on page 78)

Kumon, M., Sugawara, T., Miike, K., Mizumoto, I., and Iwai, Z. (**2003**), "Adaptive audio servo for multirate robot systems," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2003)*. (Cited on page 78)

Laflaquiere, A. (**2013**), "Approche sensori-motrice de la perception de l'espace pour la robotique autonome (sensorimotor approach of space perception for autonomous robotics)," Ph.D. thesis, Pierre and Marie University, Paris. (Cited on pages 81 and 82)

Laflaquiere, A., O'Regan, J., Argentieri, S., Gas, B., and Terekhov, A. (**2015**), "Learning agent's spatial configuration from sensorimotor invariants," *Robotics and Autonomous Systems* . (Cited on page 81)

Langhans, A. and Kohlrausch, A. (**1992**), "Differences in auditory performance between monaural and diotic conditions. I: Masked thresholds in frozen noise," *J. Acoust. Soc. Am.* **91**, pp. 3456–3470. (Cited on page 62)

Larsen, E. and Liberman, M. C. (**2010**), "Contralateral cochlear effects of ipsi-

lateral damage: No evidence for interaural coupling," *Hearing Research* **260**(1–2), pp. 70 – 80, URL http://www.sciencedirect.com/science/article/pii/S0378595509002871. (Cited on page 61)

Laurent, I. (**2014**), "The neuromorphic vision toolkit," URL http://ilab.usc.edu/toolkit/home.shtml. (Cited on page 51)

LaValle, S. (**2011**), "Motion planning," *IEEE ASSP Robotics Automation Magazin* **18**(1), pp. 79–89. (Cited on page 71)

Lavie, N. (**2005**), "Distracted and confuses? Selected attention under load," *Trends in Cognitive Sciences* **9**, pp. 75–82. (Cited on page 45)

Lever, C., Wills, T., Cacucci, F., Burgess, N., and O'Keefe, J. (**2002**), "Long-term plasticity in hippocampal place-cell representation of environmental geometry," *Letters to Nature* **416**, pp. 90–94. (Cited on page 7)

Lilaonitkul, W. and Guinan, J. J. (**2009**), "Human medial olivocochlear reflex: Effects as functions of contralateral, iIpsilateral, and bilateral elicitor bandwidths," *J. Assoc. Res. Otolaryngology.* **10**, pp. 459–470. (Cited on page 62)

Locke, L. (**1689**), *An essay concerning human understanding*, re-edited by P. Nidditch 1979, Oxford Univ. Press, GB–Oxford. (Cited on page 43)

Lotto, A. J. and Sullivan, S. C. (**2008**), "Speech as a sound source," in *Auditory perception of sound sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay, Springer, Berlin–Heidelberg–New York, vol. 29, chap. 10, pp. 281–305. (Cited on page 49)

Lu, Y. and Cooke, M. (**2010**), "Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners," *Speech Communication* . (Cited on page 75)

Macedo, L. (**2004**), "Modeling forms of surprise in artificial agents: Empirical and theoretical study of surprise functions," in *26th Annual Conf. Cognitive Science Soc.* (Cited on pages 34 and 36)

Macedo, L. and Cardoso, A. (**2004**), "Using CBR in the exploration of unknown environments with an autonomous agent," in *Advances in Case-Based Reasoning*, edited by P. Funk and P. A. González Calero, Springer Berlin Heidelberg, vol. 3155 of *Lecture Notes in Computer Science*, pp. 272–286, URL http://dx.doi.org/10.1007/978-3-540-28631-8_21. (Cited on page 34)

Macedo, L. and Cardoso, A. (**2005**), "The role of surprise, curiosity and hunger on exploration of unknown environments populated with entities," in *2005 Portuguese Conf. Artificial Intelligence*, Ieee, pp. 47–53, URL http://ieeexplore.ieee.org/lpdocs/

epic03/wrapper.htm?arnumber=4145922. (Cited on pages 34, 35, 36, and 40)

Machmer, T. and Moragues, J. (**2009**), "Robust impulsive sound source localization by means of an energy detector for temporal alignment and pre-classification," in *Proc. Europ. Sig. Proc.*, pp. 1409–1412. (Cited on page 52)

Maison, S., Micheyl, C., and Collet, L. (**2001**), "Influence of focused auditory attention on cochlear activity in humans." *Psychophysiology* **38**(1), pp. 35–40. (Cited on page 62)

Makarenko, A. A., Williams, S. B., Bourgault, F., and Durrant-whyte, H. F. (**2002**), "An experiment in integrated exploration," in *IEEE Intl. Conf. Robots and Systems*. (Cited on pages 20 and 21)

Malis, E. and Chaumette, F. (**2000**), "2D1/2 visual servoing with respect to unknown objects through a new estimation scheme of camera displacement," *Intl. J. Computer Vision* **37**(1), pp. 79–97. (Cited on page 72)

Mansard, N. and Chaumette, F. (**2007**), "Task sequencing for high-level sensor-based control," *IEEE Trans. Robotics* **23**(1), pp. 60–72. (Cited on page 73)

Manzanares, M., Bolea, Y., and Grau, A. (**2011**), "Robust audio localization for mobile robots iniIndustrial environments," in *Advances in Sound Localization*, edited by P. Strumillo, InTech. (Cited on page 77)

Marković, I. and Petrović, I. (**2010**), "Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering," *Robotics and Autonomous Systems* **58**(11), pp. 1185–1196. (Cited on page 75)

Marković, I., Portello, A., Danès, P., Petrović, I., and Argentieri, S. (**2013**), "Active speaker localization with circular likelihoods and bootstrap filtering," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2013)*, Tokyo, Japan. (Cited on page 76)

Markus, E. J., Barnes, C. A., McNaughton, B. L., Gladden, V. L., and Skaggs, W. E. (**1994**), "Spatial information content and reliability of hippocampal CA1 neurons: Effects of visual input," *Hippocampus* **4**(4), pp. 410–421. (Cited on page 8)

Martinson, E., Apker, T., and Bugajska, M. (**2011**), "Optimizing a reconfigurable robotic microphone array," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2011)*, pp. 125–130. (Cited on page 77)

Martinson, E. and Schultz, A. (**2009**), "Discovery of sound sources by an autonomous mobile robot," *Autonomous Robots* **27**, pp. 221–237. (Cited on page 76)

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., and Moser, M.-B. (**2006**), "Path integration and the neural basis of the 'cognitive map'." *Nature Reviews: Neu-*

roscience **7**(8), pp. 663–78, URL http://www.ncbi.nlm.nih.gov/pubmed/16858394. (Cited on page 8)

Merriam-Webster Dictionary (**accessed 2014.05.21**), "entry: attention," URL http://www.merriam-webster.com/dictionary/attention. (Cited on page 43)

Messing, D. P., Delhorne, L., Bruckert, E., Braida, L. D., and Ghitza, O. (**2009**), "A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise," *Speech Communication* **51**(8), pp. 668–683. (Cited on page 64)

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (**2008**), "The iCub humanoid robot : an open platform for research in embodied cognition." in *Proc. 8th Worksh. Performance Metrics for Intelligent Systems*, pp. 50–56. (Cited on page 52)

Meyer, J.-a. (**1996**), "From natural to artifcial life: Biomimetic mechanisms in animat designs," *Robotics and Autonomous Systems* , pp. 1–26. (Cited on page 24)

Meyer, J.-A., Guillot, A., Girard, B., Khamassi, M., Pirim, P., and Berthoz, A. (**2005**), "The Psikharpax project: towards building an artificial rat," *Robotics and Autonomous Systems* **50**(4), pp. 211–223, URL http://linkinghub.elsevier.com/retrieve/pii/S0921889004001757. (Cited on pages 26, 27, 28, and 29)

Milford, M. J., Wyeth, Gordon, F., and Prasser, D. (**2004**), "RatSLAM : A Hippocampal model for simultaneous localization and mapping," in *IEEE Intl. Conf. Robotics and Automation*. (Cited on pages 25 and 26)

Miller, G. A. and Licklider, J. C. R. (**1950**), "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**, pp. 167–173. (Cited on page 50)

Mink, J. W. and Thach, W. T. (**1993**), "Basal ganglia intrinsic circuits and their role in behavior." *Current Opinion in Neurobiology* **3**(6), pp. 950–7, URL http://www.ncbi.nlm.nih.gov/pubmed/8124079. (Cited on page 10)

Mishra, S. K. and Lutman, M. E. (**2014**), "Top-down influences of the medial olivocochlear efferent system in speech perception in noise," *PLoS ONE* **9**(1), pp. e85756. (Cited on page 60)

Moddemeijer, R. (**1988**), "An information theoretical delay estimator," in *Ninth Symposium on Information Theory in the Benelux*, pp. 121–128. (Cited on page 53)

Moore, B., Tyler, L., and Marslen-Wilson, W. (Eds.) (**2010**), *The perception of speech: from sound to meaning*, Oxford UK.Press, UK–Oxford. (Cited on page 49)

Moore, B. C. J. (**1997**), *An introduction to the psychology of hearing*, Academic Press, GB–London, 4th ed. (Cited on page 47)

Moray, N. (**1959**), "Attention in dichotic listening: Effective cues and the influence of the instructions," *Quarterly J. Exper .Psychol.* **11**, pp. 56–60. (Cited on page 49)

Moser, E. I., Kropff, E., and Moser, M.-B. (**2008**), "Place cells, grid cells, and the brain's spatial representation system." *Annual Rev. Neuroscience* **31**, pp. 69–89, URL `http://www.ncbi.nlm.nih.gov/pubmed/18284371`. (Cited on pages 9 and 10)

Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., and Sayd, P. (**2006**), "Real time localization and 3D reconstruction," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR'2006)*. (Cited on page 71)

Muller, R. U. and Kubie, J. L. (**1987**), "The effects of changes in the environment hippocampal cells on the spatial firing of," *J. Neuroscience* **7**(7), pp. 1951–1968. (Cited on page 8)

Murata, T. (**1989**), "Petri nets: Properties, analysis and applications," *Proc. IEEE* **77**(4), pp. 541 – 580. (Cited on page 54)

Murphy, G. L. (**2004**), *The big book of concepts*, MIT Press. (Cited on page 53)

Murray, N. (**1970**), *Attention: Selective processes in vision and hearing*, Academic Press, New York NY. (Cited on page 45)

Nahum, M., Nelken, I., and Ahissar, M. (**2008**), "Low-level information and high-level perception: the case of speech in noise." *PLoS Biology* **6**(5), pp. e126, URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2386842&tool=pmcentrez&rendertype=abstract`. (Cited on page 12)

Nakadai, K., Lourens, T., Okuno, H., and Kitano, H. (**2000**a), "Active audition for humanoids," in *Nat. Conf. Artificial Intelligence, AAAI-2000*, Austin, TX, pp. 832–839. (Cited on pages 53 and 69)

Nakadai, K., Matsui, T., Okuno, H., and Kitano, H. (**2000**b), "Active audition system and humanoid exterior design," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2000)*, Takamatsu, Japon, pp. 1453–1461. (Cited on page 69)

Nelken, I. and Ahissar, M. (**2006**), "High-level and low-level processing in the auditory system : The role of primary auditory cortex," *Dynamic of Speech Production and Perception* , pp. 5–12. (Cited on pages 12 and 13)

Nicola, S. M., Yun, I. a., Wakabayashi, K. T., and Fields, H. L. (**2004**), "Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative stimulus task." *J. Neurophysiology* **91**(4), pp. 1840–65, URL `http://www.ncbi.nlm.nih.gov/pubmed/14645377`. (Cited on page 7)

Nobre, K. and Kastner, S. (**2014**), *The Oxford handb. of attention*, Oxford Univ. Press,

UK–Oxford. (Cited on page 44)

O'Keefe, J. and Nadel, L. (**1978**), *The hippocampus as a cognitive map*, Oxford University Press. (Cited on pages 8 and 10)

Okuno, H., Nakadai, K., Hidai, K., Mizoguchi, H., and Kitano, H. (**2001**), "Human-robot interaction through real-time auditory and visual multiple-talker tracking," *IEEE/RSJ Intl. Conf. Intell. Robots and Systems.* **3**. (Cited on pages 52 and 53)

O'Regan, J. K. and Noe, A. (**2001**), "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences* **24**, pp. 939–973, URL `http://journals.cambridge.org/article_S0140525X01000115`. (Cited on pages 78 and 80)

Oudeyer, P.-y. and Kaplan, F. (**2008**), "How can we define intrinsic motivation?" in *Intl. Conf. Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems.* (Cited on pages 30, 32, and 40)

Oxford Dictionary (**accessed 2014.05.21**), "entry: attention," URL `http://www.oxforddictionaries.com/definition/english/attention`. (Cited on page 43)

Perrett, S. and Noble, W. (**1997**a), "The contribution of head motion cues to localization of low-pass noise." *Perception & Psychophysics* **59**(7), pp. 1018–1026, URL `http://www.ncbi.nlm.nih.gov/pubmed/9360475`. (Cited on page 17)

Perrett, S. and Noble, W. (**1997**b), "The effect of head rotations on vertical plane sound localization." *J. Acoust. Soc. Am.* **102**(4), pp. 2325–2332, URL `http://www.ncbi.nlm.nih.gov/pubmed/9348691`. (Cited on page 17)

Philipona, D., O'Regan, J., and Nadal, J. (**2003**), "Is there something out there? Inferring space from sensorimotor dependencies," *Neural Computation* **15**(9), pp. 2029–2049. (Cited on page 78)

Poincaré, H. (**1945**), "L'espace et la géométrie (Space and geometry)," *Revue de métaphysique et de Morale* . (Cited on page 80)

Portello, A., Bustamante, G., Danès, P., Piat, J., and Manhès, J. (**2014**), "Active Localization of an Intermittent Sound Source from a Moving Binaural Sensor," in *Forum Acustium (FA'2014)*, Krakow, Poland. (Cited on page 76)

Portello, A., Danès, P., and Argentieri, S. (**2011**), "Acoustic models and Kalman filtering strategies for active binaural sound localization," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2011)*, San Francisco, CA. (Cited on page 76)

Portello, A., Danès, P., and Argentieri, S. (**2012**), "Active binaural localization of intermittent moving sources in the presence of false measurements," in *IEEE/RSJ Intl. Conf.*

on Intell. Robots and Systems (IROS'2012), Vilamoura, Portugal. (Cited on page 76)

Portello, A., Danès, P., Argentieri, S., and Pledel, S. (**2013**), "HRTF-based source azimuth estimation and activity detection from a binaural sensor," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2013)*, Tokyo, Japan. (Cited on page 76)

Posner, M. and Cohen, Y. (**1984**), "Components of visual orienting," in *Attention and performance*, edited by H. Bouma and D. G. Bouhuis, MIT press, Cambridge MA, vol. X, pp. 531–555. (Cited on page 46)

Prescott, T. J., Redgrave, P., and Gurney, K. (**1999**), "Layered control architectures in robots and vertebrates," *Adaptive Behavior* **7**(1), pp. 99–127, URL `http://adb.sagepub.com/cgi/doi/10.1177/105971239900700105`. (Cited on page 10)

Quirk, G. J., Muller, R. U., and Kubie, J. L. (**2008**), "The firing of hippocampal rat ' s recent experience place cells in the dark depends on the," *J. Neuroscience* **10**(6), pp. 2008–2017. (Cited on page 8)

Raake, A. and Blauert, J. (**2013**), "Comprehensive modeling of the formation process of sound-quality," in *Proc. QoMEX*, AU-Klagenfurt. (Cited on page 48)

Raffo, G., Farines, J., Becker, L., and Moreno, U. (**2011**), "Tutorial 1: Mobile robotics," in *Brazilian Symp. Computing System Engineering (SBESC)*, pp. 206–207. (Cited on page 70)

Ragozzino, M. E., Detrick, S., and Kesner, R. P. (**1999**), "Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning." *J. Neuroscience* **19**(11), pp. 4585–94, URL `http://www.ncbi.nlm.nih.gov/pubmed/10341256`. (Cited on page 7)

Ranó, I. (**2007**), "On taxis for control and its qualitative solution on mobile robots," in *Proc. Intell. Autonomous Vehicles*. (Cited on page 68)

Ravulakollu, K. K., Erwin, H., and Burn, K. (**2011**), "Improving robot-human communication by integrating visual attention and auditory localization using a biologically inspired model of superior colliculus," *Advanced Materials Res.* **403 - 408**, pp. 4711–4717. (Cited on page 54)

Redgrave, P., Prescott, T. J., and Gurney, K. (**1999**), "The basal ganglia: A a vertebrate solution to the selection problem?" *J. Neuroscience* **89**(4), pp. 1009–1023. (Cited on page 10)

Redish, A. D. (**2001**), "The hippocampal debate: are we asking the right questions?" *Behavioural Brain Research* **127**(1-2), pp. 81–98, URL `http://www.ncbi.nlm.nih.gov/pubmed/11718886`. (Cited on page 7)

Reiter, E. R. and Liberman, M. C. (**1995**), "Efferent-mediated protection from acoustic overexposure: relation to slow effects of olivocochlear stimulation." *J. Neurophysiology* **73**(2), pp. 506–514. (Cited on page 61)

Reynolds, J. and Desimone, R. (**2000**), "Competitive mechanisms subserve selective visual attention," in *Image, Language, Brain*, edited by Y. Marantz, Y. Miyashita, and W. O'Neil, The MIT press, Cambridge MA, pp. 233–247. (Cited on page 44)

Rodemann, T., Joublin, F., and Goerick, C. (**2009**), "Audio proto objects for improved sound localization," *2009 IEEE/RSJ Intl. Conf. Intelligent Robots and Systems* , pp. 187–192URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5354698`. (Cited on pages 14 and 15)

Rosenthat, D. F. and Okuno, H. G. (Eds.) (**1998**), *Computational auditory scene analysis*, Lawrence Erlbaum, Mahwak NY. (Cited on page 51)

Roy, N., Mccallum, A., and Com, M. W. (**2001**), "Toward optimal active learning through Monte-Carlo estimation of error reduction," in *Intl. Conf. Machine Learning*. (Cited on page 31)

Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., and Pfeifer, R. (**2008**), "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," in *Proceedings - IEEE Intl. Conf. Robotics and Automation*, pp. 962–967. (Cited on page 52)

Russell, I. J. and Murugasu, E. (**1997**), "Medial efferent inhibition suppresses basilar membrane responses to near characteristic frequency tones of moderate to high intensities," *J. Acoust. Soc. Am.* **102**, pp. 1734–1738. (Cited on page 63)

Ryan, R. and Deci, E. (**2000**), "Intrinsic and extrinsic motivations: Classic definitions and new directions." *Contemporary Educational Psychology* **25**(1), pp. 54–67, URL `http://www.ncbi.nlm.nih.gov/pubmed/10620381`. (Cited on page 30)

Saab, B. J., Georgiou, J., Nath, A., Lee, F. J. S., Wang, M., Michalon, A., Liu, F., Mansuy, I. M., and Roder, J. C. (**2009**), "NCS-1 in the dentate gyrus promotes exploration, synaptic plasticity, and rapid acquisition of spatial memory." *Neuron* **63**(5), pp. 643–56, URL `http://www.ncbi.nlm.nih.gov/pubmed/19755107`. (Cited on page 7)

Sanchiz, J. and Fisher, R. (**2000**), "Viewpoint estimation in three-dimensional images taken with perspective range sensors," *IEEE Trans. Pattern Analysis and Machine Intelligence* **22**(11), pp. 1324–1329. (Cited on page 74)

Santangelo, V. and Spence, C. (**2007**), "Is the exogenous orienting of spatial attention truly automatic? A multisensory perspective," *Conciousness and Cognition* **17**, pp. 989–1015. (Cited on page 46)

Sasaki, Y., Thompson, S., Kaneyoshi, M., and Kagami, S. (**2010**), "Map-generation and Identification of Multiple Sound Sources from Robot in Motion," in *IEEE/RSJ Intl. Conf. Intell. Robots and Systems (IROS'2010)*, Taipei, Taiwan, pp. 437–443. (Cited on page 76)

Scardovi, L. (**2005**), "Information based control for state and parameter estimation," Ph.D. thesis, University of Genoa. (Cited on page 74)

Scharf, B. (**1970**), "Critical bands," in *Foundations of Modern Auditory Theory*, edited by J. Tobias, Academic Press, New York NY, vol. 1, pp. 155–202. (Cited on page 47)

Scharf, B., Quigly, S., Aoki, C., Peachy, N., and Revves, A. (**1987**), "Focused attention and frequency selectivity," *J. Acoust. Soc. Am.* **42**, pp. 215–223. (Cited on page 47)

Schauerte, B., Kühn, B., Kroschel, K., and Stiefelhagen, R. (**2011**), "Multimodal saliency-based attention for object-based scene analysis," in *IEEE Intl. Conf. on Intell. Robots and Systems*, pp. 1173–1179. (Cited on page 52)

Schauerte, B. and Stiefelhagen, R. (**2013**), "Wow! Bayesian surprise for salient acoustic event detection," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP'2013)*, pp. 6402–6406. (Cited on page 51)

Schlauch, R. S. and Hafter, E. R. . (**1991**), "Listening bandwidth and frequency uncertainty in pure-ton signal detection," *J. Acoust. Soc. Am.* **90**, pp. 1332–1339. (Cited on page 47)

Schmidhuber, J. (**1991**), "Curious model-building control systems," in *Intl. Joint Conf. on Neural Networks*, Singapore, pp. 1458–1463. (Cited on page 33)

Schofield, B. R. (**2009**), "Structural organization of the descending auditory pathway," in *Oxford Handb. of Auditory Science, Vol. 2: The Auditory Brain*, edited by R. A. and A. R. Palmer, Oxford Univ. Press, New York, NY, pp. 43–64. (Cited on pages 1 and 2)

Shamma, S. (**2008**), "On the emergence and awareness of auditory objects." *PLoS biology* **6**(6), pp. e155, URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2435155&tool=pmcentrez&rendertype=abstract`. (Cited on pages 12, 13, and 14)

Shamma, S. (**2013**), "Cortical processes for navigating complex acoustical environments (invited talk)," in *Third BINHAAR Symp.*, CNRS/LAAS- Univ. de Toulouse Paul Sabatier. (Cited on page 1)

Shannon, C. E. (**1948**), "A mathematical theory of communication," *Bell System Technical J.* **27**, pp. 379–423, 623–656. (Cited on page 36)

Shinn-Cunningham, B. G. (**2008**), "Object-based auditory and visual attention," *Trends*

in Cogn. Scs. **12**, pp. 182–186. (Cited on page 48)

Shirai, Y. and Inoue, H. (**1973**), "Guiding a robot by visual feedback in assembling tasks," *Pattern Recognition* **5**, pp. 99–108. (Cited on page 71)

Smith, D. W., Aouad, R. K., and Keil, A. (**2012**), "Cognitive task demands modulate the sensitivity of the human cochlea." *Frontiers of Psychology* **3**, pp. 30. (Cited on page 62)

Sommerlade, E. and Reid, I. (**2008**), "Information-theoretic active scene exploration," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'2008)*. (Cited on page 74)

Souères, P., Cadenat, V., and Djeddou, M. (**2003**), "Dynamical sequence of multi-sensor based tasks for mobile robots navigation," in *IFAC Symp. on Robot Control (SYROCO'2003)*. (Cited on page 73)

Spence, C. and Santangelo, V. (**2010**), "Auditory attention," in *Oxford Handb. of auditory science*, edited by C. J. Plack, Oxford Univ. Press, vol. 3: Hearing, pp. 249–270. (Cited on page 44)

Spence, C. J. and Driver, J. (**1994**), "Covert spatial attention in audition: Exogenous and endegenous mechnisms," *J. Exp. Psychology* **20**, pp. 555–574. (Cited on page 46)

Sridhar, T. S., Liberman, M. C., Brown, M. C., and Sewell, W. F. (**1995**), "A novel cholinergic "slow effect" of efferent stimulation on cochlear potentials in the guinea pig." *J. Neuroscience* **15**(5 Pt 1), pp. 3667–3678. (Cited on page 59)

Stachniss, C., Dirk, H., and Burgard, W. (**2004**), "Exploration with active loop-closing for FastSLAM," in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems (IROS'2013)*. (Cited on page 21)

Stanford Encyclopedia of Philosopy (**accessed 2014.05.21**), "entry: attention," URL `http://plato.stanford.edu/entries/attention/`, last revision 2013. (Cited on page 43)

Steckel, J. and Peremans, H. (**2013**), "BatSLAM: simultaneous localization and mapping using biomimetic sonar," *PloS one* **8**. (Cited on page 77)

Summerfield, C. and Egner, T. (**2003**), "Attention and decision making," in *The Oxford handb. of attention*, edited by A. C. Nobre and S. Kastner, Oxford Univ. press, GB–Oxford, pp. 837–864. (Cited on page 45)

Taha, S. a., Nicola, S. M., and Fields, H. L. (**2007**), "Cue-evoked encoding of movement planning and execution in the rat nucleus accumbens." *J. Physiology* **584**(Pt 3), pp. 801–18, URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?`

artid=2276984&tool=pmcentrez&rendertype=abstract. (Cited on page 7)

Taube, J. S. (**2007**), "The head direction signal: origins and sensory-motor integration." *Ann.l Rev. Neuroscience* **30**, pp. 181–207, URL http://www.ncbi.nlm.nih.gov/pubmed/17341158. (Cited on page 11)

Taube, J. S. and Muller, R. U. (**1998**), "Comparisons of head direction cell activity in the postsubiculum and anterior thalamus of freely moving rats." *Hippocampus* **8**(2), pp. 87–108, URL http://www.ncbi.nlm.nih.gov/pubmed/9572715. (Cited on page 9)

Thrun, S., Burgard, W., and Fox, D. (**1998**), "A probabilistic approach to concurrent mapping and localization for mobile robots," *Autonomous Robots* **5**(3-4), pp. 253–271. (Cited on page 69)

Thrun, S., Burgard, W., and Fox, D. (**2005**), *Probabilistic robotics*, The MIT Press. (Cited on pages 71 and 77)

Thurlow, W. R., Mangels, J. W., and Runge, P. S. (**1967**), "Head movements during sound localization," *J. Acoust. Soc. Am.* **42**(2), pp. 489–493, URL http://scitation.aip.org/content/asa/journal/jasa/42/2/10.1121/1.1910605. (Cited on pages 16 and 17)

Thurlow, W. R. and Runge, P. S. (**1967**), "Effect of induced head movements on localization of direction of sounds," *J. Acoust. Soc. Am.* **42**(2), pp. 480–488, URL http://scitation.aip.org/content/asa/journal/jasa/42/2/10.1121/1.1910604. (Cited on page 16)

Treisman, A. (**2003**), "Conciousness and perceptual binding," in *The unity of conciousness: Binding, integration and dissociation*, edited by A. Cleeremans, Oxford Univ. Press, GB–Oxford, pp. 95–113. (Cited on page 45)

Treisman, A. M. and Gelade, G. (**1980**), "A feature-integration theory of attention." *Cognitive Psychology* **12**, pp. 97–136. (Cited on pages 45 and 51)

Trifa, V. M., Koene, A., Morén, J., and Cheng, G. (**2007**), "Real-time acoustic source localization in noisy environments for human-robot multimodal interaction," in - *IEEE Intl. Worksh. Robot and Human Interactive Communication*, pp. 393–398. (Cited on pages 52 and 53)

Tsilfidis, A., Westerman, A., Buchholz, J., Georganti, E., and Mourjopoulos, J. (**2013**), "Binaural dereverberation," in *The technology of binaural listening*, edited by J. Blauert, Springer, Berlin–Heidelberg–New York, chap. 14, pp. 359–396. (Cited on page 46)

Tsotsos, J., Culhane, S., W.Y.K., W., Lai, Y., Davis, N., and Nuflo, F. (**1995**), "Modeling visual attention via selective tuning," *Artificial Intelligence* **78**(1–2), pp. 507–545, special Volume on Computer Vision. (Cited on page 69)

Valin, J., Michaud, F., and Rouat, J. (**2006**), "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP'2006)*. (Cited on page 75)

van Noorden (**1977**), "Minimum differnces of level and frequency for perceptual fission if tone sequences ABAB," *J. Acoust. Soc. Am.* **61**, pp. 1041–1045. (Cited on page 48)

van Noorden, L. (**1975**), "Temporal coherence in the perception of tone sequences," PhD. thesis, Techn. Univ. NL-Eindhoven, cited after Jones 1993. (Cited on page 48)

von Ehrenfels, C. (**re-edited 1990**), *Christian von Ehrenfels: Philosophische Schriften: 4. Metaphysik – (Philosophical writings. 4. metaphysics)*, Philosophia Verlag, D-München. (Cited on page 50)

Wallach, H. (**1938**), "On sound localization," *J. Acoust. Soc. Am.* **10**(1), pp. 83, URL `http://scitation.aip.org/content/asa/journal/jasa/10/1/10.1121/1.1902063`. (Cited on page 15)

Walther, T. and Cohen-Lhyver, B. (**2014**), "Multimodal feedback in auditory-based active scene exploration," in *Proc. Forum Acusticum, PL–Krakòw*. (Cited on page 54)

Wang, Q. H., Ivanov, T., and Aarabi, P. (**2004**), "Acoustic robot navigation using distributed microphone arrays," *Information Fusion* **5**, pp. 131–140. (Cited on page 77)

Ward, D., Lehmann, E., and Williamson, R. (**2003**), "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Processing* **11**(6), pp. 826–836. (Cited on page 75)

Warren, R. (**1982**), *Auditory perception: a new synthesis*, Pergamon Press, New York NY. (Cited on page 50)

Warren, R. M., Obusek, C. J., and Ackroff, J. M. (**1972**), "Auditory induction: Perceptual synthesis of absent sounds," *Science* **176**, pp. 371–383. (Cited on page 49)

Weiss, L., Sanderson, A., and Neuman, C. (**1987**), "Dynamic sensor-based control of robots with visual feedback," *IEEE Trans. Robotics and Automation* **3**(5), pp. 404–417. (Cited on pages 69 and 72)

Wenhardt, S., Deutsch, B., Hornegger, J., Niemann, H., and Denzler, J. (**2006**), "An information-theoretic approach for next-best-view planning in 3D reconstruction," in *Intl. Conf. Pattern Recognition (ICPR'2006)*. (Cited on page 74)

Westheimer, G. (**1999**), "Gestalt theory reconfigured: Max Wertheimer's anticipation of recent developments in visual neuroscience," *Perception* **28**(1), pp. 5–15, URL `http://www.perceptionweb.com/abstract.cgi?id=p2883`. (Cited on page 14)

Wikipedia (**accessed 2013.11.02**), "entry: attention," . (Cited on page 43)

WillowGarage2014 (**2014**), "The PR2 robot," URL `http://wiki.ros.org/pr2_description`. (Cited on page 54)

Wilson, W., Williams Hulls, C., and Bell, G. (**1996**), "Relative end-effector control using cartesian position based visual servoing," *IEEE Trans. Robotics and Automation* **12**(5), pp. 684–696. (Cited on page 72)

Wirth, S. and Pellenz, J. (**2007**), "Exploration transform: A stable exploring algorithm for robots in rescue environments," *IEEE Intl. Worksh. Safety, Security and Rescue Robotics* , pp. 1–5URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4381274`. (Cited on pages 21, 37, 38, and 39)

Wolfe, J. M. (**1999**), "Inattentional amnesia," in *Fleeting Memories*, edited by V. Coltheart, MIT Press, Cambridge, MA, pp. 71–94. (Cited on page 48)

Wolfe, J. M. (**2007**), "Guided search 4.0 current progress with a model of visual search," *J. Vision* **1**, pp. 99–119. (Cited on page 51)

Wolfe, J. M. and Horowitz, T. S. (**2004**), "What attributes guide the development of of visual attention, and how do they do it?" *Nature Rev. Neuroscience* **5**, pp. 1–7. (Cited on page 46)

Wood, E. R., Dudchenko, P. A., and Eichenbaum, H. (**1999**), "The global record of memory in hippocampal nNeuronal activity," *Nature* **397**, pp. 613–616. (Cited on page 8)

Wright, B. A. and Dai, H. (**1994**), "Detection of unexpected tones with short and long durations," *J. Acoust. Soc. Am.* **95**, pp. 931–938. (Cited on page 47)

Würtz, R. P. (**2008**), "Introduction: Organic computing," *Understanding Complex Systems* **2008**, pp. 1–6. (Cited on page 53)

Yamauchi, B. (**1997**), "A Frontier-based approach for autonomous exploration," in *IEEE Intl. Symposium Computational Intelligence in Robotics and Automation*, IEEE, pp. 146–151. (Cited on page 37)

Yan, R., Tee, K. P., Chua, Y., Huang, Z., and Li, H. (**2013**), "An attention-directed robot for social telepresence," in *1st Intl. Conf. Human-Agent Interaction*, pp. —-. (Cited on pages 54 and 55)

Young, B. G., Fox, G. D., and Eichenbaum, H. (**1994**), "Correlates of hippocampal complex-spike cell activity in rats performing a nonspatial radial maze task." *J. Neuroscience* **14**(11), pp. 6553–6563. (Cited on page 8)

Young, P. T. (**1931**), "The role of head movements in auditory localization," *J. Exp. Psychology* **14**(2), pp. 95–124, URL `http://content.apa.org/journals/xge/14/2/95`. (Cited on page 14)

Yu, Y., Mann, G. K. I., and Gosine, R. G. (**2010**), "An object-based visual attention model for robotic applications." *IEEE Trans. Systems, Man and Cybernetics. Part B, Cybernetics* **40**, pp. 1398–1412. (Cited on pages 53 and 56)

Zaheer Aziz, M., Mertsching, B., Shafik, M. S. E. N., and Stemmer, R. (**2006**), "Evaluation of visual attention models for robots," in *Fourth IEEE Intl. Conf. Computer Vision Systems, ICVS'06*, vol. 2006, p. 20. (Cited on page 52)

Zelinsky, A. (**1988**), "Robot navigation with learning," *Australian Computer J.* **20**(2), pp. 85–93. (Cited on page 37)

Zelinsky, A. (**1991**), "Environment exploration and path planning algorithms for a mobile robot using sonar," Ph.D. thesis, Univ. Wollongong, New South Wales, Australia. (Cited on page 37)

Zelinsky, A. (**1994**), "Using path transforms to guide the search for findpath in 2D," *Intl. J. Robotic Research* **13**(4), pp. 315–325. (Cited on page 39)

Zhao, W. and Dhar, S. (**2011**), "Fast and Slow Effects of Medial Olivocochlear Efferent Activity in Humans," *PLoS One* **6**(4), pp. e18725. (Cited on pages 59 and 61)

Zilany, M. S. and Bruce, I. C. (**2006**), "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *J. Acoust. Soc. Am.* **120**(3), pp. 1446–66. (Cited on pages 63 and 65)

Zwicker, E. (**1961**), "Subdivision of the audible frequency range into critical bands [abstract]," *J. Acoust. Soc. Am.* **33**, pp. 248. (Cited on page 47)

Zwicker, E. (**1965**), "Temporal effects in simultaneous masking and loudness," *J. Acoust. Soc. Am.* **38**, pp. 132–141. (Cited on page 60)

FP7-ICT-2013-C TWO!EARS Project 618075

Deliverable D4.1, part C

Listing of feedback loops

Thomas Walther *

November 26, 2014

* The TWO!EARS project (http://www.twoears.eu) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 618075.

Project acronym:	TWO!EARS
Project full title:	Reading the world with TWO!EARS

Work package:	4
Document number:	D4.1, part C
Document title:	Listing of feedback loops
Version:	1

Delivery date:	30th November 2014
Dissemination level:	Restricted
Nature:	Report

Editor(s):	Thomas Walther, RUB
Author(s):	Thomas Walther
Reviewer(s):	Klaus Obermayer

2 Listing of feedback loops

The enumeration in Table 1 represents an update of "Table 1: Examples of entry ports for feedback and possible actions induced" of the project proposal (part B, page 15). We have now modified the original list entries to reflect ongoing development in the signal-generation/processing stages, the blackboard architecture, and the virtual-reality mechanisms.

The modifications result in more precise feedback-port definitions as compared to the original formulation. Further, the ports are now assigned to concrete parts/interfaces of the evolving Two!Ears system framework. Note that the term "in progress" in Table 1 indicates that assignment of the corresponding entry ports is pending and subject to further investigation. In addition to the sharpened feedback-port definitions in Table 1, the information sources that actually "feed" the given ports have been revisited.

Significant progress as been made in all fields of Two!Ears allows to also update "Table 2: Points of origin of feedback to be delivered to the feedback entry ports # 1–5 [...], and expected functional improvements" (see proposal, part B, page 16). While the expected functional improvements largely remain unaltered in Table 2, the information sources have been re-labeled in order to reflect recent advances in the system framework.

Table 1: Examples of entry ports for feedback and possible actions induced

Port#	**Potential entry port for feedback**	**Possible control actions**
1	• KEMAR head rotating on KEMAR torso • KEMAR head on PR2 platform • Simulated robots in MORSE/SSR	• Rotational and translatory sensor movements (currently up to 3 DOF)
2	• WP2 signal manager, data objects, and signal processors	• Adjustment of filter bandwidths and shapes, focusing on specific spectral regions • Adjustment of operation points and dynamic ranges of operation

Table 1: Examples of entry ports for feedback and possible actions induced (2)

3	• Monaural and binaural processing stages, SOC/MSO/LSO - level modules, in progress	• Adjustment of time-windows, time constants and spectral regions • Task-specific employment of additional processing steps, e.g., lateral and contra-lateral inhibition, precedence preprocessing, dereverberation
4	• Binaural-activity-mapping stage, IC - level modules, in progress	• Setting time constants for contra-lateral inhibition • Providing masks for dedicated analyses of binaural-activity maps • Focusing on specific spectral regions • Adjustment of operation points and dynamic ranges • Provision of non-auditory sensory data, e.g. from vision, proprioception, sensorimotor cues
5	• WP3 blackboard architecture (graphical model, knowledge sources, scheduler)	• Provision of external knowledge, e.g., salient features, object-building schemata, rule-systems • Knowledge of the situational history, communicative intention of sound sources • Task-specific expert knowledge, internal references • Provision of non-auditory knowledge, e.g., from visual scene analyses

Table 2: Points of origin of feedback to be delivered to the feedback entry ports # 1-5 (see Table 1), and expected functional improvements

Port#	**Source of feedback signals and/or symbolic feedback information**	**Expected functional improvements**
1	• Binaural-processing stage • Visual cues from the (real or simulated) robot's cameras • WP3 blackboard architecture (graphical model, knowledge sources, scheduler)	• Turning the acoustic sensors into optimal position (turn-to reflex) • Advanced movements of the head-&-torso platform (active exploration, e.g., dynamic weighting)
2	• Modules operating on the SOC/MSO/LSO level • WP3 pre-segmentation stage • WP3 blackboard architecture (graphical model, knowledge sources, scheduler)	• Increasing the signal-to-noise ratio, increasing spectral and temporal selectivity • Paying attention to specific signal features to deliver specific additional information as required by the cognitive stage
3	• Binaural-activity-mapping stage • WP3 pre-segmentation stage • WP3 blackboard architecture (graphical model, knowledge sources, scheduler)	• Activation of specific (computationally more expensive) signal processing procedures, such as echo cancelling, de-reverberation, precedence-effect preprocessing • Re-evaluation (reconsideration) to solve ambiguities
4	• Visual cues from the (real or simulated) robot's cameras • Sensorimotor cues from the (real or simulated) head-&-torso platform • WP3 pre-segmentation stage • WP3 blackboard architecture (graphical model, knowledge sources, scheduler)	• Optimal positioning of the head-&-torso platform (task-specific) • Improvement of object recognition, auditory grouping, aural stream segregation, aural scene analysis, attention focusing

Table 2: Points of origin of feedback to be delivered to the feedback entry ports # 1-5 (see Table 1), and expected functional improvements (2)

5	• External knowledge sources • Visual cues from the (real or simulated) robot's cameras • Acoustic cues from the (real or simulated) KEMAR head's microphones	• Improvement of scene understanding, assignment of meaning, quality judgements, attention focusing

FP7-ICT-2013-C TWO!EARS Project 618075

Deliverable D4.1, part D

Supporting information on feedback in TWO!EARS

Benjamin Cohen-L'hyver, Patrick Danès, Chungeun Kim,
Armin Kohlrausch, Thomas Walther*

November 26, 2014

* The TWO!EARS project (`http://www.twoears.eu`) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 618075.

Project acronym:	TWO!EARS
Project full title:	Reading the world with TWO!EARS

Work package:	4
Document number:	D4.1, part D
Document title:	Supporting information on feedback in TWO!EARS
Version:	1

Delivery date:	30th November 2014
Dissemination level:	Restricted
Nature:	Report

Editor(s):	Thomas Walther, RUB
Author(s):	Benjamin Cohen-L'hyver, Patrick Danès, Chungeun Kim, Armin Kohlrausch, Thomas Walther
Reviewer(s):	Klaus Obermayer

Contents

1 Supporting information on feedback in Two!Ears

This document addresses the integration of feedback-related concepts and mechanisms into the current architecture of the Two!Ears framework. Although the supporting information filed below is not a mandatory part of D4.1, we decided to add this report in order to provide additional insight into important aspects of our key achievements in the actual project period. We highlight progress in *olivocochlear feedback* (section 1.1), discuss advances in *active exploration* and *multi-modal feedback methods* (section 1.2) and shortly step into the description of concepts employed for sensorimotor feedback within Two!Ears (section 1.3). Each section first subsumes the gain that the Two!Ears project might expect from the proposed feedback loops. Focus is then geared towards theoretical considerations on the integration of these feedback mechanisms into the actual system architecture, taking into account results from D3.2 and D5.1. Eventually, basic proof-of-concept applications for active exploration and multi-modal feedback in Two!Ears are presented in section 1.2.

1.1 Olivocochlear feedback in Two!Ears

The olivocochlear feedback loop has an ascending path, connecting the inner hair cells in the inner ear via afferent connections with neurons in the superior olivary complex in the brainstem. These neurons are, at the same time, the origin of efferent fibers, which connect back to the inner ear, terminating on the outer hair cells. Because these hair cells are intrinsically interwoven with the active behavior of the basilar membrane, activation of the feedback activity will affect the nonlinear properties of the basilar membrane, changing the nonlinear input-output characteristics of this peripheral stage of auditory transformation. In the context of the Two!Ears modeling framework, this feedback loop is located rather peripheral, and it will thus have an influence on all sounds being analyzed in the modeling framework.

The olivocochlear feedback path connects auditory neurons in the olivary complex via descending connections with the inner ear. In Two!Ears, we will only consider feedback loops originating in the Medial Olivary Complex (MOC) and ignore those that originate

from neurons in the Lateral Olivary Complex (LOC) (see D4.1, part B, chapter 4 for an extensive description). The efferent projections from the MOC project both to the outer hair cells on the ipsilateral and on the contralateral side. In consequence, acoustic activation of one ear will create a MOC reflex in both ears. In humans, the strengths of the ipsi- and the contralateral MOC feedback effect seem to be approximately equal [14].

The MOC feedback has both a *reflexive* and a *reflective* component. The physiological properties of the former have been addressed in a number of studies (see D4.1, part B, chapter 4), and also have been the object of computational modeling. Contrary, scientific insights about the role and the properties of attentional control of the MOC effect (the reflective component), are much less conclusive and need further evaluation. Note that the realization of olivocochlear feedback in the Two!Ears framework will be sufficiently flexible to experiment with attentional control of the MOC feedback connections.

1.1.1 Biological/psychoacoustic insights and existing technical approaches/ models

In order to model effects of the olivocochlear feedback loop in an auditory model, it is required to first include a *nonlinear* basilar membrane module to represent the filtering processes in the inner ear. In our system, this module is currently realized using a Dual-Resonance Non-Linear (DRNL) implementation. This nonlinear inner ear model, which stays closer to current physiological knowledge, parallels the gammatone filterbank implementation. Peripheral nonlinearities provide additional challenges for central pattern recognition and matching stages. Thus, when evaluating the role of reflexive and reflective MOC feedback in the Two!Ears model, one also has to evaluate the overall consequences of switching between a linear, and a nonlinear inner-ear filter.

The technical approach chosen here is based on a DRNL implementation with parameter settings that have extensively been evaluated in a psychoacoustic context [13]. The structure of the employed model allows higher system stages to change the model's nonlinear characteristics.

The main cause for the invocation of efferent MOC feedback might be seen in the reduction of nonlinear amplification in the corresponding inner ear section at low and medium signal levels. This, in consequence, makes the corresponding input-output characteristic less compressive and thus increases the contrast of stimulus onsets. Much of what is known of this feedback loop describes it as a reflexive system, where the MOC effect strength in the inner ear is directly coupled to the neural activity in ascending parts of the hearing system. Competing with such observations that indicate a purely reflexive behavior, there exist a number of recent studies that link the strength of the MOC feedback effect to selective

auditory attention, thus providing a reflective component. The results of these studies are, however, currently not conclusive whether attending to a sound increases or decreases the strength of the MOC effect on basilar-membrane properties. The Two!Ears framework will allow to evaluate the influence of cognitive control of peripheral nonlinearities on auditory performance in various scenarios.

1.1.2 Integration into the system architecture

The finding that MOC activity affects basilar membrane responses is reflected in the ongoing evolution of recent computer models (D4.1, part B, chapter 4). In particular, the current version of the DRNL filterbank model introduced here incorporates MOC efferent suppression in the form of a gain factor applied to the peripheral signal path of the nonlinear part of the model. Implementation of the MOC feedback mechanism in the Two!Ears framework uses the same approach – the basilar membrane nonlinearity is modeled as a DRNL filterbank, connected with the MOC efferent attenuation stage through the proposed gain factor.

As described in D2.1 and D3.1, the software framework for the periphery models incorporates an object-oriented architecture. The DRNL model has been developed in this manner, in the form of an auditory front-end processor object named drnlProc. The current structure of drnlProc follows the DRNL filterbank implementation of Jepsen *et al.* [13]. Its internal processing blocks are originally based on those used in a previous DRNL model as suggested by Meddis *et al.* [19], and are adjusted to better represent the findings from human physiological data as suggested by Lopez-Poveda and Meddis [15]. The DRNL processor can be used to replace the gammatone filterbank processor. This allows to compare the outputs of both processors on the basilar membrane level. The detailed structure of the drnlProc object specific to the Two!Ears framework is described in D2.2.

Integration of the MOC feedback follows the approach used by Ferry and Meddis [10], and Clark *et al.* [8]. Firstly, as suggested in Ferry and Meddis [10] and summarized in D4.1, part B, chapter 4, the MOC feedback was implemented within the DRNL filterbank model as the attenuation to be applied to the nonlinear path of the DRNL structure. This is an open-loop structure, where the attenuation caused by the MOC feedback at each of the filterbank's channels is controlled externally (using additional input arguments in the drnlProc object). Further development towards a closed-loop structure will follow, similar to work of Clark *et al.* [8]. A separate processor object will be developed which determines the amount of the MOC feedback. A new feature of this processor compared to the previous versions is that it will incorporate "reflexive" and "top-down" components, in order to enable feedback-related simulation and repeat experiments that investigate the role of the MOC mechanism (as discussed in D4.1, part B, chapter 4). Figure 1.1 describes the conceived MOC processor architecture within the auditory front-end framework.

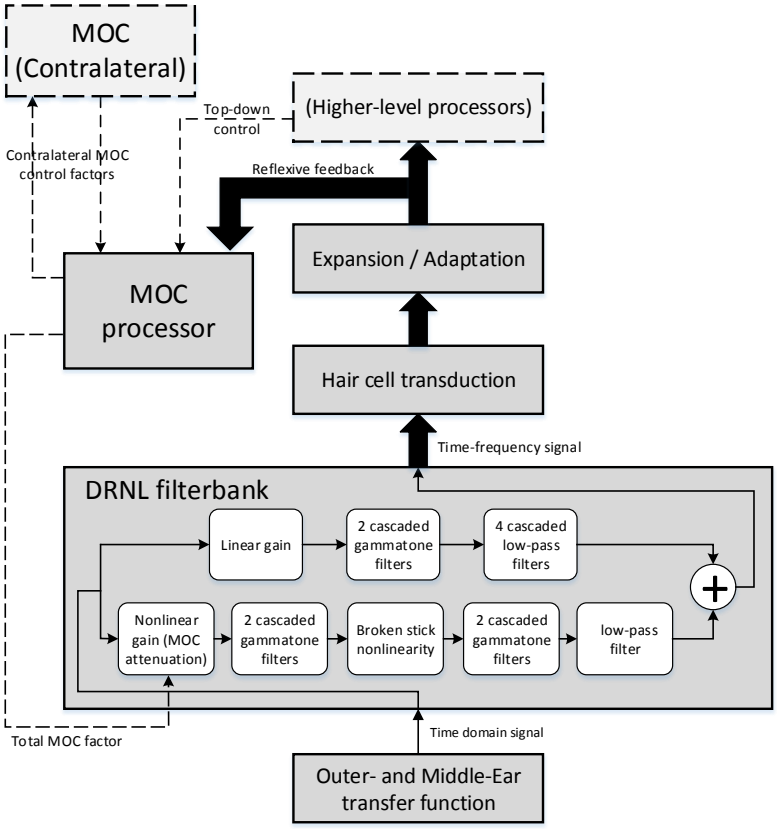


Figure 1.1: Conceived MOC processor design within the auditory front-end framework. The processor monitors the output at the auditory nerve stage, accepts control from higher-level cognitive stages, and sends and receives MOC factors to/from its contralateral instance. The final output is applied to the nonlinear path of the DRNL filterbank.

The reflexive closed-loop feedback will be implemented as to enable the MOC processor to run along with the auditory front-end modules of WP2, and to monitor the output from the auditory nerve stage. It will further allow the processor to adjust the strength of MOC feedback using internal parameters that are to be determined through comparison of experimental output to recent physiological findings (such as the cochlear rate-level functions [11]). The processor will then be extended to accept external control from higher-level cognitive stages. Currently, the employed control signal is a gain factor that can be applied to the DRNL nonlinear path. This allows the relevant cognitive stages to determine how much, if any, MOC efferent activity will be initiated, depending on the tasks and decisions made.

1.1.3 Use of olivocochlear feedback in existing and planned scenarios

The implemented MOC feedback will be tested at the auditory front-end level, and within the scenarios defined for Two!Ears. At the auditory front-end level, the performance of the peripheral models using basilar membrane nonlinearity and MOC feedback will be compared to models employing a linear basilar membrane model. A number of monaural and binaural internal representations and features that psychophysically relate to the functions of the olivocochlear system will be extracted, cf. D4.1, part B, chapter 4. Changes in features such as ratemap and ILD might reveal the impact of the basilar membrane nonlinearity and of the MOC feedback on speech perception and/or localization performance in given Two!Ears scenarios.

1.1.4 Outlook

It should be noted that recent physiological findings on MOC unmasking effects are sometimes controversial. Further, the reflective modulation mechanism of MOC activity (based, e.g., on cognitive features and attention) has not yet been revealed clearly. Going on, we will investigate performance gains induced by the application of MOC feedback in the scenarios proposed in Two!Ears, in order to see in which conditions the use of this feedback modality is advisable. Controlling factors in our experiments might be: a) assigned tasks, b) the cognitive models relating attention to the induced amount of feedback, c) target/masker signal properties. We assume that the proposed MOC feedback framework will serve as a platform to conduct further MOC-related studies. Using an efficient experimental design, our system will allow to examine task-related MOC activities and will make the simulation results easily comparable to currently available data.

1.2 Active exploration and multi-modal feedback in TWO!EARS

Given the search and rescue (S&R) situation in Fig. 1.2, the use of system inherent top-down/bottom-up loops becomes evident: the depicted robot cannot observe the procumbent victim directly, due to a wall obstacle blocking the direct line of sight. Nevertheless, the machine's acoustic sensors, namely, the ears of the artificial KEMAR head, acquire the victim's distress calls. While this acoustic sensation is likely not sufficient to perform an accurate localization of the procumbent person, the machine can use the incoming audio cues to set up an initial "working hypothesis", expecting a human victim behind the blocking obstacle. To verify or falsify this hypothesis, the robot will employ basic active exploration techniques using the panning capabilities of its sensor head: to that end, it "scans" the acoustic environment very much like humans do in order to "sharpen" its internal acoustic world model and get a more precise notion of the victim's true position.

Based on this enhanced position hypothesis, more advanced active exploration schemes kick in: the robot engages its path planning and motor units to approach the inferred position while continuously evaluating information from the head cameras to avoid collisions with given obstacles. As soon as the envisaged target comes into camera focus, the robot's cognitive units are enabled in order to determine the validity of the working hypothesis and accordingly either call a human rescuer to evacuate the victim, or restart the localization process.

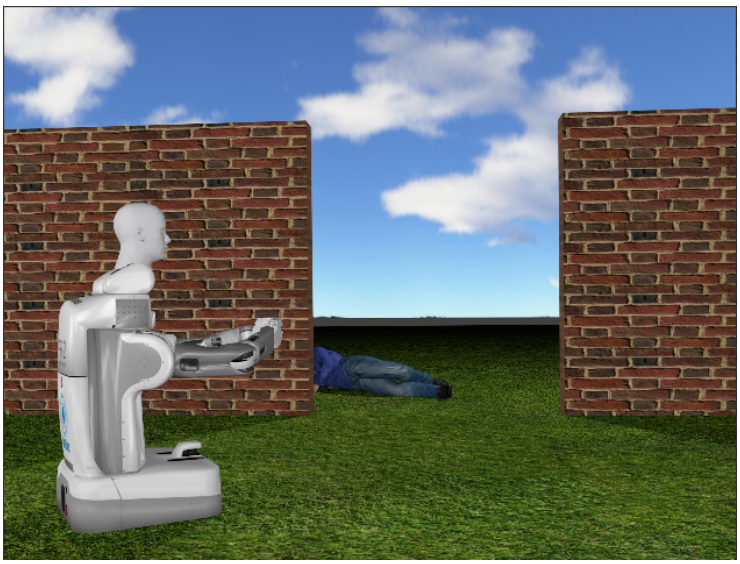


Figure 1.2: Possible search and rescue scenario

Note that the use of visual information significantly enhances the given rescue task, and essentially fits the hosting Two!Ears framework with multi-modal feedback capabilities. However, though vision is generally a very powerful cue, its reliability might be extenuated by the adverse conditions often found in S&R scenarios: lack of illumination, smoke, and environmental clutter (debris) easily overstrain the capabilities of contemporary image analysis software. Worse, the same holds for purely auditory cues: concurring sound sources (e.g., fire, yelling persons, sirens), reverberation, and obstacles are likely to render non-augmented CASA mechanisms powerless in certain situations.

Nevertheless, cues from single modalities can be combined using appropriate weighting schemes: in smoke or darkness, for instance, auditory hints receive high weight and are thus trusted more, while visual cues might become dominant in strongly reverberant acoustic environments with multiple sources and normal illumination. This strategy of cue fusion is very much in the spirit of the multi-modal feedback paradigm advocated in the Two!Ears framework: if one modality is not sufficient to perform well in a given task (here: S&R), other modalities are exploited whenever available.

Note that in the above S&R tasks, the mobile front end of the Two!Ears system has to fulfill its objectives in highly adverse and dangerous environments. Under such conditions, a thorough real-world testing of appropriate active exploration strategies would be tedious and likely endanger the physical robot, if not the human operators. To counter that issue, a well-tailored virtual test environment (VTE) will allow to quickly and safely assess complex feedback functionalities in silico, using arbitrarily complex input scenarios.

A-priori knowledge might be integrated into the VTE in order to decide on the importance of detected stimuli, and thus perform attention focusing. The algorithmic structure of the VTE can well benefit from routines provided by other project partners; in turn, insights gained from the simulation can be disseminated to other project participants. Such knowledge cross-feed is likely to yield advantages for the whole Two!Ears project. Further, the VTE allows to test multi-modal feedback early in the project's lifecycle, enabling, for instance, visual detection and categorization of observable entities.

Stepping from S&R tasks to quality of experience (QoE) scenarios, the virtual mobile front end of the VTE might scan the emulated sound field of audio-presentation systems like Ambisonics: by translating in the x/y direction and rotating its head, the robot might, for instance, find and explore the sweet spot according to the expectation of high-level quality experts. As in the above search and rescue paradigm, the virtual testbed shows potential to significantly speed up testing in QoE applications.

It should be kept in mind that the Two!Ears project eventually aims at setting up experiments in S&R and QoE in real-world environments. To that end, communication interfaces used by the VTE and the physical robot have to be unified as much as possible.

In ideal case, the Two!Ears system architecture should be able to toggle between the virtual and the real robot in a completely transparent manner.

As the elements of the Two!Ears system that are concerned with feedback will be developed over a time span of almost two years, many of these elements are not immediately available for constructing and testing of feedback procedures. This is particularly true for methods that deal with more abstract and/or more complex functions, like active exploration and multi-modal feedback methods. Both of these feedback paths will require the Two!Ears system to be endowed with sophisticated visual processing methods, e.g., for visual object detection/recognition, audio-visual speaker identification, or vision-based collision avoidance. To enable early testing of such techniques for visual processing, the VTE mimics the robot's cameras and allows to capture visual data from the simulated environments.

Our project partners can test their own feedback-related ideas in the virtual environment, take advantage of the visual data provided by the VTE and set up cooperation with WP4's feedback routines early. By that strategy, potential issues might be detected and eliminated long before final system assembly. While experimenting with the feedback loops in the virtual environment, environmental variates and labels will be identified that are definitely needed for reliable feedback, thus allowing for algorithmic streamlining.

A first VTE realized in the Two!Ears context is the Bochum Experimental Feedback Testbed (BEFT) [31], on which we report in the following.

1.2.1 The Bochum Experimental Feedback Testbed

BEFT integrates a virtual 3D visualization environment based on the OGRE 3D rendering engine [24], and hosts a mobile front end - currently a CAD model of the PR2 robot, respectively a personation of a KEMAR dummy head mounted on a rotational axis. The testbed allows to read in further scene components directly from XML files. This way, simulated entities like persons (e.g., victims), walls, terrains, and so on, can easily be added to a scenario.

The entities convey physical parameters, like distance and azimuth (w.r.t. the robot), or percentage of occlusion. Based on these parameters, BEFT simulates category labeling for each entity: "hand-crafted" degradation functions [31] weaken given a-priori knowledge of the entities' true categories in order to emulate, as closely as possible, uncertainty in category estimation caused by sensor noise or algorithmic issues.

According to the estimated parameters and the inferred category labels, the virtual mobile front end can be actuated (e.g., via active exploration mechanisms) in order to

update and enhance the parameter/label estimates and sharpen the robot's internal world model.

Note that BEFT was intended to operate on the cognitive rather than on the signal level, allowing for very early feedback testing and multi-modal analysis, by skipping Two!Ears's signal processing and pre-segmentation stages that were "under construction" at that time. However, BEFT's 3D display capabilities and its abilities to handle robot control based on the "Robot Operating System" [29] (ROS) middleware are clearly limited. The first issue might hamper simulation of visually challenging scenarios; the latter problem, however, would cause major re-work of algorithms tested in BEFT in order to port these methods to physical robot devices operating in real-world scenarios.

Figure 1.3: MORSE simulation of a KEMAR head and torso

1.2.2 The MORSE robotics simulator

As this is clearly unacceptable, the MORSE robotics simulator [20] will inherit from BEFT and become the standard VTE for visual simulation in Two!Ears. Note that MORSE is based on the BLENDER 3D modeling/simulation software [6], using Bullet [7] to enable physically plausible behavior of the simulated environment. As MORSE has the ability to operate different robotic middlewares, e.g., ROS, porting issues can be minimized by enabling the Two!Ears framework to control/read out the virtual robotic front-end (motors, cameras, microphones, etc.) using exactly the same methods that will be employed

to control/read out the physical device in later project stages. Further, MORSE comes with multiple pre-defined industrial components (sensors, actuators, controllers, and robotic platforms) which can be assembled into complete robots using straightforward Python™ scripting.

1.2.3 Integrating MORSE with Two!Ears

Several modifications of the original system architecture become mandatory (s. Fig. 1.4) to integrate MORSE with Two!Ears. In order to ensure the transparent communication announced above, MORSE uses the same middleware (ROS) that will be used on the physical robot. Communication in ROS is handled via so-called "topics" that transfer information between MORSE-internal ROS "nodes" (s. D5.1 for more details) and corresponding "external" ROS nodes that can be generated using GenoM3 [18]. These external nodes are enabled to handle custom c/c++ code – for instance, to perform complex data preprocessing – and are connected to Two!Ears's *robot interface* by a 'Matlab® bridge' interlayer designed by one of our project partners (s. D5.1).

The robot interface controls the signal flow between the virtual/real robot and the knowledge sources (KSs): motion commands (e.g., head turn, platform motion) from the KSs are routed to the ROS nodes and, in parallel, to the auralization component in order to update the robotic front-end's pose. Visual sensor information – e.g., from the cameras – is in turn routed back to the KSs from the installed ROS nodes, together with acoustic information from the auralization component or the robot's microphones. Note that currently the control aspects and the audio component of the robot interface are realized, however, video information will become available in the further course of the project. To actually fuse acoustic and visual cues, a final decision on the optimal strategy for data synchronization will be made.

With the given system architecture, it becomes clear that the robot interface is extremely important for feedback-related knowledge sources: the active exploration KS will make intense use of the interface's motion control capabilities; further, the multi-modal feedback KS relies on the visual and acoustic information that the robot interface is going to provide. The following section analyzes the structure of the aforementioned knowledge sources in greater detail. Focus will be on the information that each KS receives from the robotic front-end and the features that are sent by the KSs to the blackboard layer of Two!Ears.

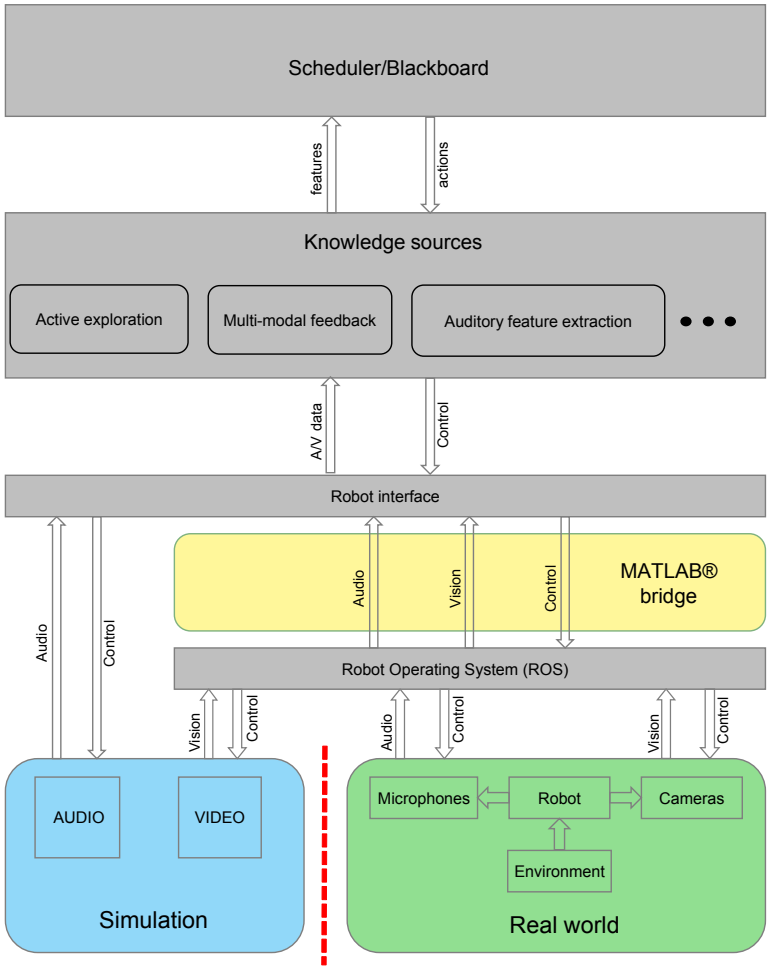


Figure 1.4: Integrating audio and video simulation (MORSE) with Two!Ears

1.2.4 Knowledge source for active exploration

Relying on the above insights, the basic properties of a purposeful knowledge source for active exploration can be subsumed as follows: first, the KS needs access to robot control mechanisms via the robot interface. A proof-of-concept application has already been realized that makes use of head panning control and allows to identify and remove false positives in acoustic source localization. For scenarios to be set up in later project stages,

active exploration capabilities have to be extended to platform heading control and to the control of platform acceleration/velocity. To that end, intelligent behavioral strategies will be encoded in the active exploration KS; while these action planning strategies remain quite simplistic in the case of the already existing scenarios, human-like behavior and problem solving strategies might serve as a paradigm for action planning in more complex scenes. A first realization of such sophisticated active exploration methods is the *dynamic weighting* scheme introduced below.

1.2.5 Dynamic Weighting

Assume there is a glass falling off a table. Consider the observation of the dropping glass as a first *event*. Once the glass touches the ground, a specific sound should be emitted (second event). These two events are *congruent*: the second one can be predicted as an expected consequence of the first event. If this natural chain of events breaks because the glass drops without emitting an audible stimulus, an *incongruency* results that will immediately catch an observer's attention. As the observer focuses the glass on the floor she/he acquires additional information: maybe a carpet on the ground prevented the glass from breaking, or maybe the "glass" was made of plastic.

The *Dynamic Weighting module* (DWmod) aims at mimicking a human observer's capabilities in situations like the one described above by taking a *weighting* approach (note that the excerpts below have largely been taken from [31]): DWmod simulates a feedback loop that would biologically extend between the cochlea and the primary auditory cortex. The module mimics two biological phenomena that occur in auditory processing, namely: (i) Turn-to-reflex movements, and (ii) attentional filtering of sound sources. DWmod is based on the *congruency* of incoming auditory events. As in geometry or algebra, the congruency of two or more objects is a measure of the similarities of some features of these objects. In perception, congruency relates to perceptual and semantic features that link two objects (for example, seeing a dog and hearing it bark). If the reliability of an auditory stimulus is low, ambiguities due to lack of information or unexpected behavior of an object become likely.

The dynamic weighting module

In the Two!Ears framework, DWmod will then trigger a turn-to-reflex that possibly changes the robot's current task depending on the potential importance weight or threat of an incoming stimulus. For the present work, DW is based on the labels of all relevant environmental objects. The dynamic weighting module is able to evaluate incoming stimuli according to their importance and possibly triggers a quick turn-to reaction or changes the robot's current task. At each discrete time step $t \leq T$, with T being the length of

the simulation (here, $T = 70$), the weight of each perceived object is computed. DWmod updates the weights in a discrete manner. The time interval between two time steps can be changed. For any newly detected object, the corresponding weight will be set to 0.5. If an object that is congruent emits a novel sound stimulus, the DW module assigns less importance to the onset of this stimulus, potentially suppressing the turn-to-reflex, while quickly decreasing the corresponding object's weight.

Weights are limited to lie between 0 and 1, where 0 represents a highly congruent sound object and 1 indicates a highly incongruent and/or dangerous sound object. To that end, bifurcating logarithmic functions are used to model the dynamics of the machine's intended reaction:

$$w_+(t_c) = \begin{cases} \frac{log(t_c)}{2*log(n)} + c & \text{if } 1 \leq t_c < T_{max} \\ \frac{log(T_{max})}{2*log(n)} + c & \text{if } t_c \geq T_{max} \end{cases} \tag{1.1}$$

$$w_-(t_c) = \begin{cases} 2c - w_1(t_c) & \text{if } 1 \leq t_c < T_{max} \\ 2c - w_2(t_c) & \text{if } t_c \geq T_{max} \end{cases} \tag{1.2}$$

Here, w_+ is the ascending part of the bifurcation, and w_- sketches the descending part. Let $c = (0.25, 0.5, 0.75)$ be the state of congruency, depending on the activity and the label of a given object. The higher c, the less congruent the new auditory object is. t_c represents an "internal" step counter that is started at the time of stimulus onset and is reset when either $t_c > T_{max}$ (to avoid deadlock situations), or when the object's activity changes. Here, we choose $T_{max} = 10$. If any congruent object is suddenly detected as incongruent (for example, a person that was walking and is now yelling), the object's weight will abruptly increase to 0.75 and then further increase to 1. If the person stops yelling, that does not necessarily mean that the person is not in danger anymore. Thus, the corresponding object's weight slowly decreases to a residual value of 0.5.

Once weights have been computed, DWmod outputs the object that should be focused on by the agent. This output consists of two parts:

- *turnHeadToObject*: controls head turning. This part of the output is used to enable the robotic agent to continue its current task while focusing attention on a different object/event. For instance, assume that the agent goes towards a person yelling and suddenly hears crackling of a nearby fire. The agent should then not stop its way towards the yelling person, but should turn its head towards the fire in order to monitor it and acquire more information about this potential threat.

- *moveToObject*: set to redefine the task of the robotic agent by providing it with a new goal. For instance, assume that the agent heads towards some landmark in a non-S&R scenario, and suddenly hears a person yelling. The moveToObject part of

the output can then be used to change the robot's goal/task in order to rescue the endangered person.

Preliminary results

In an early version of DWmod, experiments have been conducted in a virtual S&R scenario using Matlab® (see [31]). The simulated scene integrates the robotic agent, a wall obstacle and different sound sources (human victims, car, fire, ...) that become active at different time steps. The first task of the virtual agent has been to approach the "car" sound source, given distractions from the other sound sources: one person yells in an interleaved, but regular pattern, another person generates continuous "walking" sounds. Further distraction comes from a starting fire. At $t = 0$, the fire is not allowed to harm any person in the scenario. This restriction is later alleviated, causing the walking person to become endangered by the approaching fire. For completeness, note that the robotic agent is supplied with a-priori knowledge of the position and the category labels of each sound source.

The simulation was designed to test the agent's ability to determine whether an overheard acoustic stimulus is important (high weight) or not (low weight). At the moment, notion of importance directly relates to the threat caused by each object: if an object's auditory "fingerprint" signals danger (e.g., siren wailing, fire crackling, etc.), the agent immediately has to focus this object. In the contrary case that the auditory fingerprint signals an object being endangered (e.g., person yelling, person falling), the robot not only has to focus the object, but should also start a rescue attempt.

Figure 1.5 shows the results of our simulation. The temporal course of the weights of different objects (person1, person2, fire, and car) is sketched. At timestep $t = 36$, the fire starts to endanger the walking person, causing the robot to re-focus its attention to person1, and to shift its task from "approach car" to "rescue person1".

Our simulation show promising results in attention focusing and early decision making. Nevertheless, we enhanced the basic version of DWmod as indicated in the next paragraph.

Dynamic weighting module – enhancements

Rule-based approaches to early decision making (including DWmod) are interesting, as they allow straightforward encoding of purposeful behavioral strategies. However, these approaches turn out to become very complex as the number of scenarios/objects/action patterns increases. There exist two approaches to counter that issue. The first would be

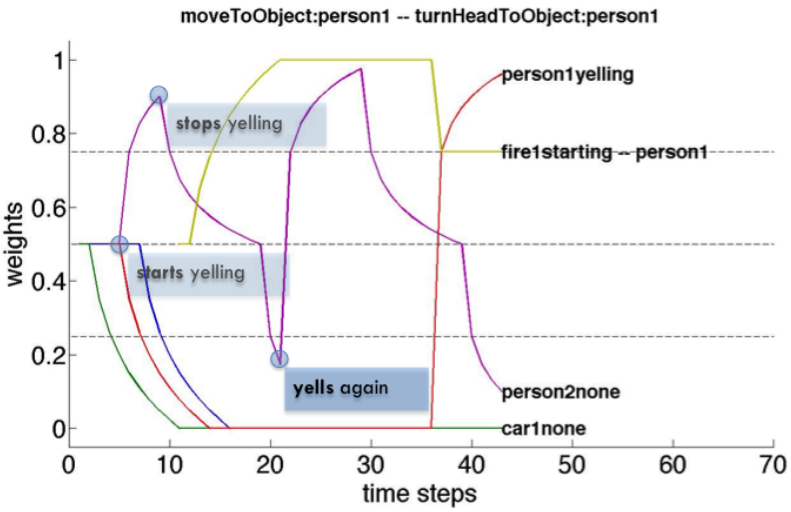


Figure 1.5: Dynamic weighting applied to a baseline S&R simulation. See text for details.

to implement a continuous learning algorithm that re-trains the knowledge base/action planner of the DW module on each occurrence of a new object/stimulus. However, this method requires integration of WP3 classifiers into DWmod, thus significantly increasing computation time. The second approach requires to enhance DWmod's capabilities in generalization in order to allow the system to get adapted to new kinds of objects. Within Two!Ears, we decided to pursue the latter strategy.

Thus, the main ambitions for version 2 of DWmod are:

- Work on acoustic data acquired in real time. To that end, we will connect DWmod to the WP2 signal processors and the WP3 knowledge sources (see Fig. 1.6). We plan to acquire real-time acoustic data by either using the microphones of the Kemar head or by utilizing an external sound card (such as the "RME Babyface");
- Characterize objects in a more general way, for instance by employing a notion based on entropy [30]: the onset of an unexpected stimulus increases the environmental entropy and catches the robot's attention.

- Relate the psychophysical notion of *motivation*[1] to Two!Ears active exploration feedback mechanisms.
- Integrate vision, since objects used in Two!Ears are mostly audio-visual objects.

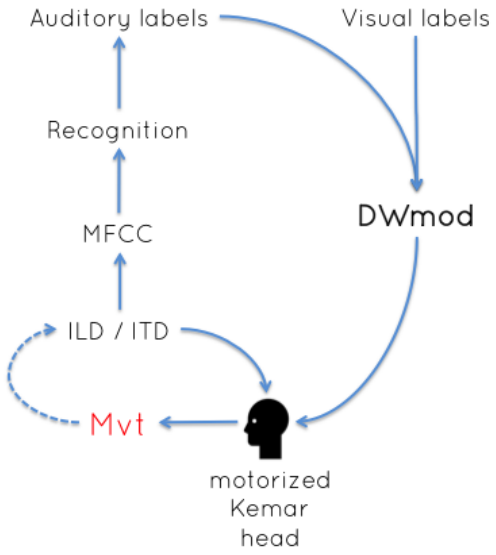


Figure 1.6: Scheme of the DWmod version under development. See text for details.

Dynamic Weighting Virtual Testbed Environment

In order to make further tests on the new version of DWmod, a basic VTE (DWvte) is developed under Matlab®. Figure 1.7 shows the graphical user interface of DWvte. For completeness, note that DWvte heavily relies on the sound processors provided by WP2, and connects the sound capturing modules (e.g, sound card, KEMAR microphones) to Matlab®. This connection is established via a virtual machine that runs ROS and GenoM3 and has been provided by WP5.

1 See [3, 4] on the psychophysics notion of motivation, [26, 1, 2] on the robotic implementation of intrinsic motivation, [16, 17] on the notion of curiosity and surprise, and [12, 23] for contributions on the understanding of how visual and auditory information streams are processed. See also *Chapter 2* of the literature survey for a dedicated chapter on these notions of motivation.

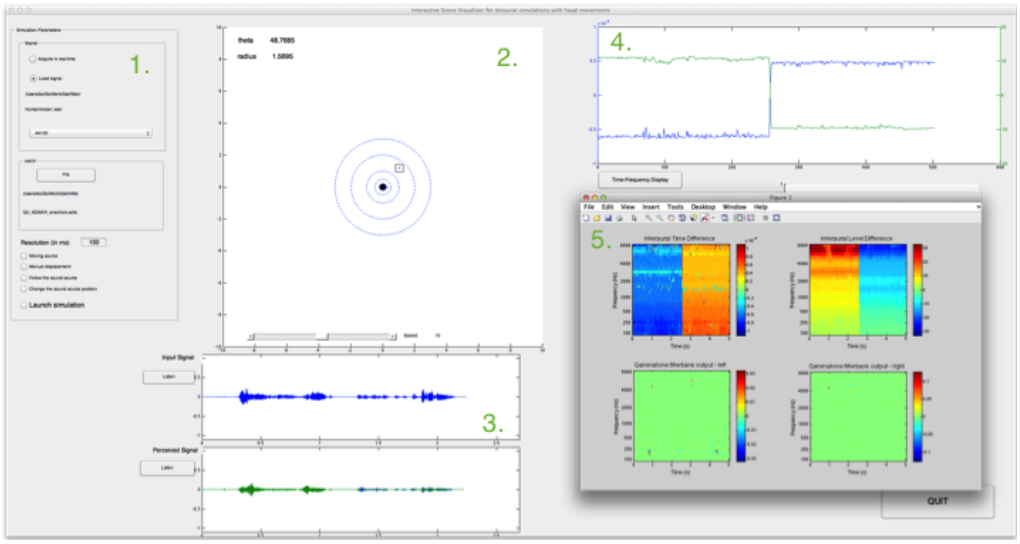


Figure 1.7: Dynamic Weighting Virtual Testbed Environment

Conclusion

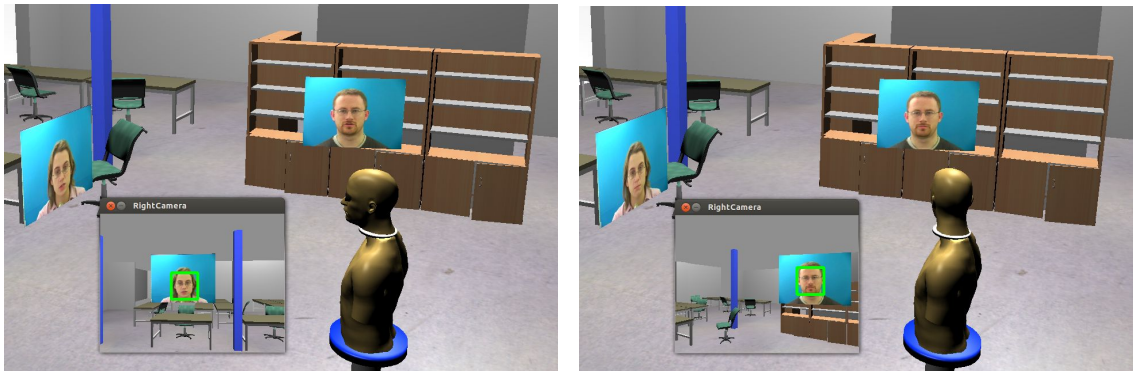
Note that motor commands triggered by DWmod can also be issued by higher level system stages in order to realize feedback methods in active exploration and attention focusing tasks. Within TWO!EARS, DWmod can be seen as to reside in between low-level *reflexive* stages, and high-level cognitive processing stages.

With the DWvte, we aim at providing an easy-to-use, graphical user interface that can be employed to test DWmod in several dynamic scenarios. The virtual environment is helpful to visualize and understand complex modules and algorithms that are going to be developed in the next project phase. With MORSE integration into TWO!EARS becoming more and more plenary, functionalities of DWmod will be transferred into TWO!EARS's knowledge sources; thus DWvte will be superseded in later system versions.

1.2.6 Knowledge source for multi-modal scene analysis

While TWO!EARS primarily relies on acoustic cues for both S&R and QoE, visual information is assumed to significantly enhance estimation results provided by the auditory subsystems. Following this idea, a multi-modal KS becomes mandatory in TWO!EARS: this knowledge source will integrate audio and video data, and is intended to allow for sophisticated cue weighting together with hypotheses testing based on multi-modal information. To that end, the corresponding KS realization requires access to the (virtual) robot's cameras and microphones as provided by the robot interface. As stated above, there are

some synchronization issues that have to be resolved in order to effectively combine data from the visual and acoustic sensor entities; while this synchronization is essential for tasks like audio-visual speech recognition, basic detection tasks can potentially be accomplished by the multi-modal KS without syncing the information flow.



(a) **(b)**

Figure 1.8: Face detection in the MORSE simulator

Note that first experiments in visual scene analysis have already been conducted in MORSE, paving the way for the construction of multi-modal knowledge sources: we set up a first proof-of-concept application to demonstrate visual processing capabilities in the MORSE environment: an emulated KEMAR-like head/torso combination is embedded in a standard MORSE scenario, see Figure 1.8. The artificial head is allowed to rotate freely and has two virtual cameras attached. Videos of human speakers are projected into the virtual environment using basic video texturing methods.

Note that the videos are currently chosen from the audio-visual corpus of Cooke *et al.* [9]. One of the robot's cameras is connected to an external ROS node that was created using GenoM3 [18]. This external node is enabled to perform fast image processing based on the OpenCV library [32]: images from the virtual camera are streamed into the node and are then analyzed with OpenCV's face detection engine.

Figure 1.8 shows the results of face detection in MORSE: found faces are marked by green rectangles; the corresponding face regions could be propagated to higher system layers, e.g., to perform audio-visual speaker identification. Note that there are still some issues to be solved with respect to synchronization between the audio and the video data stream, s. above. Also, the above toy application is not yet fully integrated into Two!Ears. To eventually become consistent with the framework's architecture, the face detection mechanism will have to be encapsulated in a knowledge source. Further, communication with the visual processing ROS node has to be established via Two!Ears's robot interface.

1.3 Sensorimotor feedback

Sensorimotor feedback loops are intended to model hardwired behavior, and to seamlessly interweave sensory stimulation and motion. They are located at the reflex level, and act on short time scales; the *turn-to-reflex* may be seen as a typical example of sensorimotor feedback. Tight integration of motion and sensory stimulation complies with recent developments in embodied cognition [25], postulating that our sensory experience arises from mastering sensorimotor contingencies by learning the variation of stimuli as a function of bodily movement. In robotics, the synthesis of so-called "active" binaural auditory functions, which incorporate the motor commands of the sensor, has long been acknowledged [21]. These active functions aim at overcoming the limitations of their passive counterparts – *e.g.*, issues with front-back ambiguities, distance non-observability. More, active binaural functions may be used to perceive a sound source in the "auditory fovea" [22] while keeping the engineering design simple.

1.3.1 A three-stage framework for active source localization

In the vein of [28], three fundamental stages have been identified, the first two being related to the analysis of the sensorimotor flow, the third being feedback in itself (s. D4.1, part B, chapter 5). These three stages are defined as follows (cf. Figure 1.9)

(A) Short-term detection: Estimation of the spatial arrangement/number of active sources from the analysis of the binaural stream over small time snippets.

(B) Audio-motor binaural localization: Assimilation of the data over time, and combination with the motor commands of the sensor, so as to arrive at a first level of active localization.

(C) Information-based feedback control of the binaural sensor: Feedback control of the sensor motion so as to improve the fusion performed in audio-motor binaural localization.

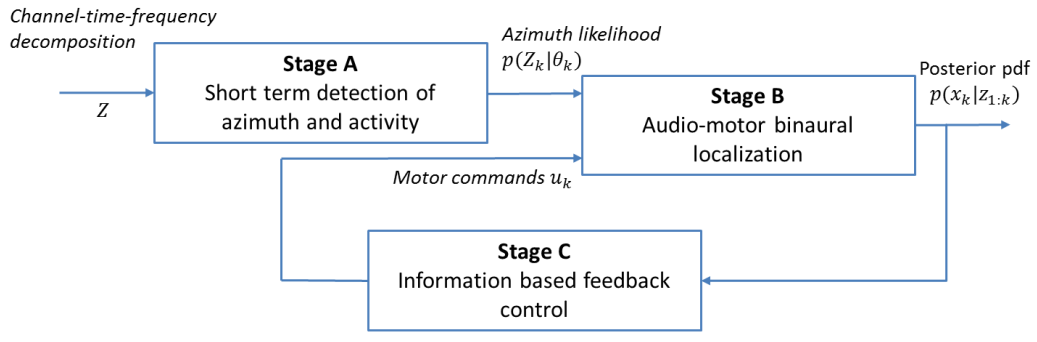


Figure 1.9: Three-stage active binaural localization.

The extraction of spatial source characteristics in stage ***(A)*** can be performed through maximum likelihood estimation on the basis of the channel-time-frequency distribution of the binaural signals over a group of time frames. Herein, it is assumed that the signals sensed in each frame are jointly Gaussian and wide-sense stationary, and that the relative motion between the binaural sensor and the sources is negligible. A closed-form separable solution can be obtained for the single source case – leading to the likelihood $p(Z_k|\theta_k)$ of the azimuth θ_k at time step k if the problem is planar. Multiple sources can be handled via the "Expectation-Maximization" algorithm if they are W-disjoint orthogonal, that is, if at most one source has significant energy in each "bin" of the channel-frame-frequency decomposition of the binaural stream [27]. Source activity detection can be addressed through information criteria. An advantage of [27] is that it handles scattering – by capitalizing on prior knowledge of the left and right HRTFs – separately from the statistics of the noise. However, since the sensed noise statistics are affected by sensor motion, the noise estimator requires a nontrivial online update procedure. The approach has not yet been extended to reverberant environments.

The assimilation over time of the output history from stage ***(A)*** with the motor commands of the sensor can naturally be performed in a stochastic filtering scheme, which is the cornerstone of stage ***(B)***. Let $\mathbf{x}_k$ be the estimated state vector of the underlying stochastic state space model. Assume that $\mathbf{x}_k$ describes the sensor-to-source relation(s) at time step k. The control input to the model is constituted by the motor commands of the binaural sensor. The stochastic state equation then depicts the effect of sensor motion on localization (assuming, e.g., rigid body kinematics). If the sources move as well, their absolute positions must be inserted in the unknown state vector, and their dynamics are described by an autonomous system with unknown initial conditions. Here, the challenges are threefold: first, due to model nonlinearities, the consistency of the filter must be carefully examined. Even if the filter relies on perfectly known noise statistics, its approximation of the state posterior density function can be overly "optimistic" or inconsistent due to overestimation of range, etc. Second, the filter must be endowed with self-initialization abilities, as well as with routines that cope with false measurements and source intermittence. Third, data association problems predictably occur in the case of multiple sources.

Also, sensor motion obviously affects the quality of localization. As mentioned, the architecture of stage ***(C)*** relates to the design of the feedback controller. Here, the idea is to define a criterion that judges the quality of exploration based on the parameters of the posterior probability distribution function of the state. If the synthesis of the control law is guided by this factor, other competing objectives have to be included, for instance, the energy of the control signal, the error to be stabilized on average, etc. The challenge is to bridge the gap between the mathematical statement of the problem and a tractable implementation.

1.3.2 Implementation in the TWO!EARS framework and usefulness for the case studies of WP6

The proposed three-stage framework complies with the blackboard architecture that constitutes the core of the cognitive part of TWO!EARS. The input to stage ***(A)*** is an input to the graphical model. Stage ***(B)*** is carried out by an expert in the architecture. The posterior probability distribution function that this expert computes is stored in a node of the graphical model. Stage ***(C)*** is carried out by an expert, too. From the robotics viewpoint, all three stages must be implemented as functional modules, as they are subject to severe time and communication constraints.

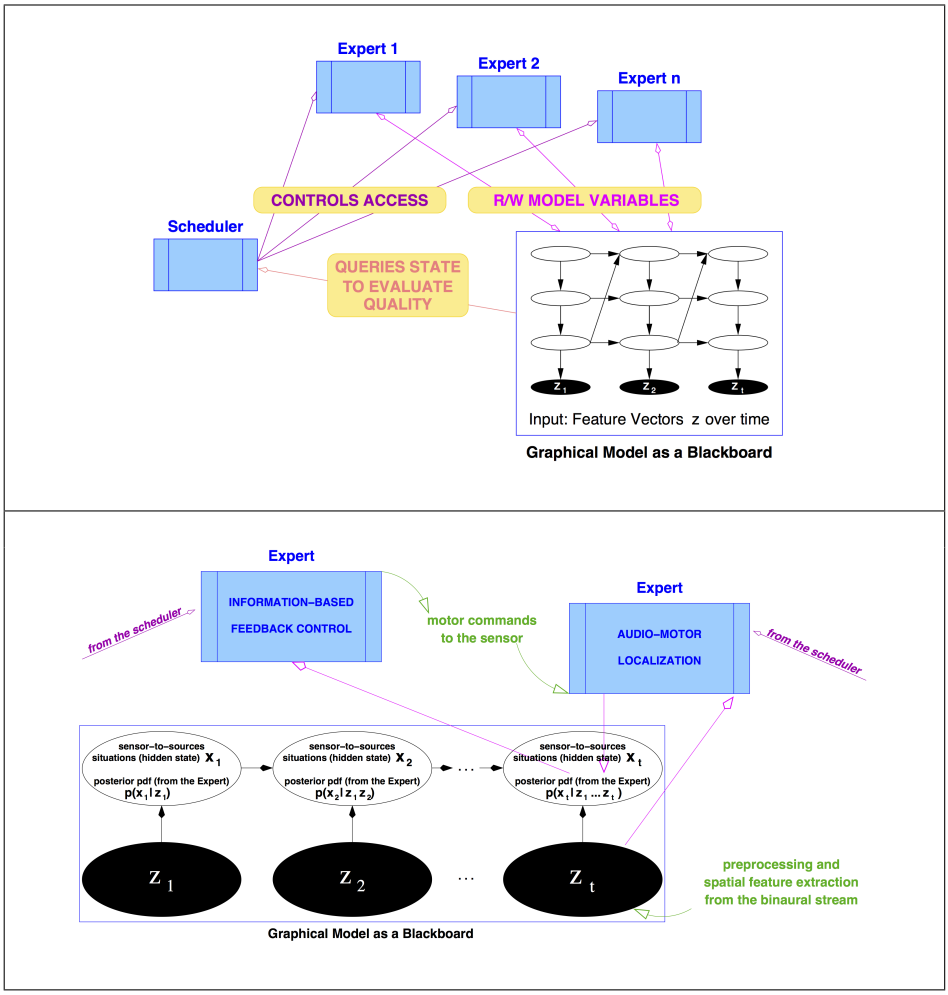


Figure 1.10: Top: Schematic plot of the blackboard architecture of TWO!EARS. Bottom: Schematic plot of the sensorimotor feedback within this architecture, cf. [5]

Bibliography

[1] Baranes, A. and Oudeyer, P. Y. (**2010**), "Intrinsically Motivated Goal Exploration for Active Motor Learning in Robots: A Case Study," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 1766–1773. (Cited on page 16)

[2] Baranès, A. and Oudeyer, P.-y. (**2013**), "R-IAC : Robust Intrinsically Motivated Exploration and Active Learning," in *IEEE Transactions Autonomous, Mental Development*, vol. 1, pp. 155–169. (Cited on page 16)

[3] Berlyne, D. E. (**1954**), "A Theory of Human Curiosity," *British Journal of Psychology* **45**(3), pp. 180–191. (Cited on page 16)

[4] Berlyne, D. E. (**1965**), *Structure and direction in thinking*, Wiley. (Cited on page 16)

[5] Blauert, J., Kolossa, D., and Danès, P. (**2014**), "Feedback loops in engineering models of binaural listening," in *JASA-EL*, under review. (Cited on page 21)

[6] Blender Foundation (**2014**), "Blender - 3D open source animation suite," URL `http://www.blender.org/`. (Cited on page 9)

[7] Bullet (**2014**), "The Bullet Physics Engine," URL `http://bulletphysics.org/wordpress/`. (Cited on page 9)

[8] Clark, N. R., Brown, G. J., Jürgens, T., and Meddis, R. (**2012**), "A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise." *The Journal of the Acoustical Society of America* **132**(3), pp. 1535–41. (Cited on page 3)

[9] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (**2006**), "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America* **120**, pp. 2421–2424. (Cited on page 18)

[10] Ferry, R. T. and Meddis, R. (**2007**), "A computer model of medial efferent suppression in the mammalian auditory system," *The Journal of the Acoustical Society of America* **122**(6), pp. 3519–3526. (Cited on page 3)

[11] Guinan, J. J. (**2006**), "Olivocochlear efferents: anatomy, physiology, function, and

the measurement of efferent effects in humans." *Ear and hearing* **27**(6), pp. 589–607. (Cited on page 4)

[12] Hochstein, S. and Ahissar, M. (**2002**), "View from the Top : Hierarchies and Reverse Hierarchies Review," *Neuron* **36**(3), pp. 791–804. (Cited on page 16)

[13] Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**), "A computational model of human auditory signal processing and perception." *The Journal of the Acoustical Society of America* **124**(1), pp. 422–38. (Cited on pages 2 and 3)

[14] Lilaonitkul, W. and Guinan, J. J. (**2009**), "Human Medial Olivocochlear Reflex: Effects as Functions of Contralateral, Ipsilateral, and Bilateral Elicitor Bandwidths," *J Assoc Res Otolaryngol.* **10**, pp. 459–470. (Cited on page 2)

[15] Lopez-Poveda, E. A. and Meddis, R. (**2001**), "A human nonlinear cochlear filterbank," *The Journal of the Acoustical Society of America* **110**(6), pp. 3107–3118. (Cited on page 3)

[16] Macedo, L. and Cardoso, A. (**2005**), "The role of Surprise, Curiosity and Hunger on Exploration of Unknown Environments Populated with Entities," in *2005 Purtuguese Conference on Artificial Intelligence*, Ieee, pp. 47–53, URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4145922`. (Cited on page 16)

[17] Macedo, L. and Cardoso, A. (**2012**), "The Exploration of Unknown Environments Populated with Entities by a Surprise-Curiosity-Based Agent," *Cognitive Systems Research* **19**(20), pp. 62–87. (Cited on page 16)

[18] Mallet, A. and Herrb, M. (**2011**), "Recent developments of the GenoM robotic component generator," in *6th National Conference on Control Architectures of Robots*, INRIA Grenoble Rhône-Alpes, Grenoble, France. (Cited on pages 10 and 18)

[19] Meddis, R., O'Mard, L. P., and Lopez-Poveda, E. A. (**2001**), "A computational algorithm for computing nonlinear auditory frequency selectivity," *The Journal of the Acoustical Society of America* **109**(6), pp. 2852–2861. (Cited on page 3)

[20] MORSE (**2014**), "MORSE, the Modular OpenRobots Simulation Engine," URL `https://www.openrobots.org/wiki/morse/`. (Cited on page 9)

[21] Nakadai, K., Lourens, T., Okuno, H., and Kitano, H. (**2000**), "Active Audition for Humanoid," in *Nat. Conf. Artificial Intelligence, AAAI-2000*, Austin, TX. (Cited on page 19)

[22] Nakadai, K., Okuno, H., and Kitano, H. (**2002**), "Exploiting Auditory Fovea in Humanoid-Human Interaction," in *Nat. Conf. Artificial Intelligence, AAAI-2002*, Edmonton, Canada. (Cited on page 19)

[23] Nelken, I. and Ahissar, M. (**2006**), "High-level and Low-level Processing in the Auditory System : The Role of Primary Auditory Cortex," *Dynamic of Speech Production and Perception* , pp. 5–12. (Cited on page 16)

[24] OGRE (**2014**), "OGRE - Open Source 3D Graphics Engine," URL `http://www.ogre3d.org/`. (Cited on page 8)

[25] O'Regan, J. and Noe, A. (**2001**), "A Sensorimotor Account of Vision and Visual Consciousness," *Behavioral and Brain Sciences* **24**(5), pp. 939–1031. (Cited on page 19)

[26] Oudeyer, P.-y. and Kaplan, F. (**2008**), "How Can We Define Intrinsic Motivation?" in *Epigenetics Robotics: Modeling Cognitive Development in Robotic Systems*. (Cited on page 16)

[27] Portello, A., Bustamante, G., Danès, P., and Mifsud, A. (**2014**), "Localization of Multiple Sources from a Binaural Head in a Known Noisy Environment," in *IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS'2014)*, Chicago, IL. (Cited on page 20)

[28] Portello, A., Bustamante, G., Danès, P., Piat, J., and Manhès, J. (**2014**), "Active Localization of an Intermittent Sound Source from a Moving Binaural Sensor," in *Forum Acustium (FA'2014)*, Krakow, Poland. (Cited on page 19)

[29] ROS (**2014**), "The Robot Operating System," URL `http://www.ros.org/`. (Cited on page 9)

[30] Shannon, C. E. (**1948**), "A Mathematical Theory of Communication," *The Bell System Technical Journal* **27**(3), pp. 379–423. (Cited on page 15)

[31] Walther, T. and Cohen-L'hyver, B. (**2014**), "Multimodal feedback in auditory-based active scene exploration," in *Proc. Forum Acusticum*, Kraków, Poland. (Cited on pages 8, 12, and 14)

[32] WillowGarage (**2014**), "Open Source Computer Vision Library," URL `http://sourceforge.net/projects/opencvlibrary/`. (Cited on page 18)