Audition as a Trigger of Head Movements

Benjamin Cohen-Lhyver, Sylvain Argentieri, Bruno Gas

Audition as a Trigger of Head Movements

B. Cohen-Lhyver, S. Argentieri, B. Gas

Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France

Summary. In multimodal realistic environments, audition and vision are the prominent two sensory modalities that work together to provide us with a best possible understanding of the perceptual contents of the world. Yet, when designing artificial binaural systems, this collaboration often not honored. Instead, substantial effort is made to construct best performing purely-auditory-analysis systems, sometimes with goals and ambitions that reach beyond human capabilities. It is often not considered that, what enables us to perform so well in complex environments, is the ability of, (i), using more than one source of information, for instance, visual in addition to auditory one and, (ii), making assumptions about the objects to be perceived on the basis of a-priory knowledge. In fact, the human capability of inferring information from one modality to another one helps substantially to efficiently analyze the complex environments that humans face everyday. Along this line of thinking, this chapter addresses the effects of *attention reorientation* triggered by audition. Accordingly, it discusses mechanism that lead to appropriate motor reactions, such as head movements for putting our visual sensors toward an audiovisual object of interest. After presenting some of the neuronal foundations of multimodal integration and motor reactions linked to auditory-visual perception, some ideas and issues from the field of a robotics are tackled. This is accomplished by referring to computational modeling. Thereby some biological bases are discussed as underlie active multimodal perception, and it is demonstrated how these can be taken into account when designing artificial agents endowed with human-like perception.

1 Introduction

Assume the following situation: A listener in a lecture hall attends a talk of a fellow researcher. The conference room is almost full, and everyone has reached a seat. Yet people keep sparingly moving in all along the talks, trying to make as little noise as possible while they thread their way through the rows to find an available chair. While the talk is still going on, a sharp, vivid, but muffled due to the distance, sound of a small glass breaking on the floor of the lecture hall reaches the listener's right ear. A first observable reaction,

Fig. 1. *Attention reorientation* caused by the occurrence of an unpredictable stimulus leading to head movement towards the audiovisual source. This motor reaction enables the visual sensors to acquire supplemental data about the object of interest – after Corbetta et al. (2008).

will very likely to be the *turn-to reflex*, namely, listeners quickly turn their heads towards the object that has caused the sound.

Why?

Such head movements are an attempt to guide the optical sensors (eyes) to spatial areas of interest, namely, to enable an analysis complementary to the one that has already been performed beforehand by the auditory modality. This primary analysis is indeed responsible for the alerting mechanism. Reactions triggered in such a way are generally termed *attention reorienting*. In the case discussed here, the reaction was initiated by auditory cues – see Fig. 1. Turning our head in a case like this is a manifestation of the need to *focus* on a particular object of interest that occurs in an environment.

Attention reorienting is an observable consequence of the integration of multiple complex mechanisms giving humans the ability to react quickly to complex environments. In particular, head movements are triggered by various different signals and situations, for instance, danger signals, but also unexpected perceptual objects such as stimuli requiring our attention or carrying an interest with respect to a task to accomplish.

In the situation described above – besides the notion of danger signal – a main characteristics of the "falling glass" object is its obvious rareness in the context given and, consequently, its low predictability. In other words, neither any perceptual clue nor any prior information has supported anticipation of this object. But whatever the origins of the head movements are, they resulted in putting the optical sensors towards an area that required deeper analyses of the perceptual information.

What does actually "cause" head-turn reactions?

As stated above, audition is a modality capable of triggering movements of the head towards an unpredictable object. However, will that reaction also occur in a situation were glasses are constantly falling and breaking on the ground? In other words, how important is the context in which an object is occurring, and how does this context with consequent motor reaction? A same object can thus either trigger motor reactions in a specific environment, or remain completely unnoticed in a different environment. A dog barking in a kennel would certainly provoke a different reaction in you than if barking in your bedroom, given that it is not your own dog. The difference between these two environments is the *predictability* of the object to occur. Thus, it has to be concluded that the occurrence of an auditory, visual, or audiovisual object is not an inherent attribute of the corresponding signals. It is the context that determines consequent motor reactions. Predictability is thus a key in understanding the mechanism of attention reorienting.

In particularly, all these considerations are of importance when it comes to the design of artificial agents endowed with human-like multimodal perception capabilities. Such agents aim at understanding complex environments in a similar way as humans do. Thus, they have to be able to process the different kinds of signals as perceived by their dedicated sensors, artificial ears and eyes for instance, but also to know how to combine them appropriately to form a multimodal perceptual world. The technologies of both sound processing and image processing, that is, a *multimodal approach* are needed to provide these robots with an adequate comprehension of the world. However, at least in the robot community, audition and vision are often considered as two separate senses with distinct information channels, each used to form perceptual worlds in their own particular way.

In order provide more evidence of the relevance of a thorough multimodal understanding of the world, the question will be addressed of *"How can audition be utilized as a trigger for head movements towards objects of interest?"* that is, how can one modality, for example, audition, be used for requisition of another modality, for instance, vision, to the end of gaining a better understanding of a multimodal environment? To address this question, four key neuronal phenomena are discussed in the second section of this chapter, which form a solid basis of the comprehension of multimodal integration, motor reactions, and prediction abilities of the human sensory cortices. First will be introduced the Superior Colliculus, a brain structure that is responsible for the cross-modal integration of audio and visual information, and for a consequent motor reaction depending on this incoming data. Secondly will be described the Reverse Hierarchy Theory (RHT), formalized in the first decade of the 2000's, and that proposes a powerful hypothesis about how audio and visual streams of data are processed both along a complementary bottom-up and top-down manner. Thirdly, the Mismatch Negativity phenomenon will be described, a phenomenon that illustrates how sensitive to unpredictable inputs

the sensory cortices are. Finally, the concept of Saliency will be discussed, for its strong involvement in sensory data analysis, especially for the detection, and subsequent reaction, to discontinuities in the sensory flow. Understanding these mechanisms provide helpful hints for designing artificial intelligences aimed at being integrated in robots that are to be furnished with human-like perception. As an example, a computational model of the head-turn reflex driven by auditory information will be described in the third section, namely the Head Turning Modulation (HTM) model. A conclusive section will then sum up the chapter.

2 Neuronal roots

There is extensive literature available concerning the relevant phenomena mentioned above. It includes binaural audition and sound processing by dedicated cortical areas, binocular vision, and image processing by other dedicated cortical areas, multimodal integration, attention computing, and motor reactions – both in reflexive as well as in reflective behavior – compare Blauert and Brown (2018), this volume.

Consequently, the following descriptions are restricted to biological foundations of *attention reorientation caused by audition*. Four neuronal mechanisms are dealt with in this context. These are mechanisms that represent primary biological components to be understood and considered when designing artificial agents with attentional capabilities driven by multimodal perception. Importantly, even though these phenomena are all involved in attention reorientation in their own particular way, they will here be described separately, for biological evidence of direct links between them in the processing of audiovisual signals has not been extensively studied, as for now.

2.1 Superior Colliculus

The *Superior Colliculus* (SC), is a good example for illustrating how important multimodal integration is in the analysis of sensory information. It is now widely accepted that multimodal integration is crucial even for unimodal perceptual flow analysis (Atilgan et al., 2018), in particular when it comes to designing artificial systems that use auditory, auditory, visual, or any other sensory modality. Taking for example the *cocktail party effect* (Cherry (1953); Cherry and Taylor (1954)) and the related analysis of auditory and visual information, the following two approaches to cross-modal interaction are conceivable.

- One may consider auditory and vision as two distinct modalities being separately processed through different and well characterized channels; and only the results of these analyses in each perceptual modality being used for further analyses and integration

- One may, alternatively, consider that cross-modal integration is already performed at low levels of the participating modal pathways, thus benefiting as early as possible from each available source of information.

The first approach is actually guided by a common misconception, namely, the assumption that sensory cortices process information solely from the sensors they they are directly connected to: the auditory cortex processes only auditory input sounds from the ears, while the visual cortices only process visual inputs from the eyes. However, there is ample evidence nowadays that a strict separation of different modalities and the accompanying neural areas does not exist. For instance, various studies have shown the ability of the visual cortex to also process sounds (Shams et al., 2005; Iurilli et al., 2012; Vetter et al., 2014). Others have found in return that the auditory cortex can also process visual input (Sharma et al., 2000; Belin et al., 2000; Finney et al., 2001). Of course, the auditory cortex has the major role in sound processing, and the visual cortex is far from contributing as much as the former one in sound processing. But in the context discussed here, the question is not how important the cross-modal contribution is, but rather the fact that it *does* exist at all.

The SC is a suitable candidate for the location where cross-modal integration actually happens in the central nervous system. Perhaps nowhere is the convergence of modalities more evident than there, as asserted by Meredith and Stein (1986) on the basis of a extensive review of research works on multimodal integration in mammal brains. Located in the brainstem, the SC is organized in seven layers, split into two functional units. One of these receives sensory inputs (mainly from vision, audition, and proprioception), the other one generates motor commands on the basis of this sensory input. These motor commands can, for instance, be eye saccades (Moschovakis, 1996), body movements (Stein et al., 2004), in particular head movements(May, 2006).

By binding quick motor reactions to sensory inputs, the SC is thought to play an important role in attentional reactions, in particular *exogeneous* ones[1]. Two major phenomenon have been observed in attentional reactions in which the SC is involved, namely,

- If two cross-modal stimuli are sufficiently overlapping in space and time, a synergistic effect will be observed in the multimodal neurons of the SC,
- This synergy will be more pronounced when the modality of the stimuli is the less influential one in the neurons of the SC – a phenomenon called *multimodal enhancement.*

Moreover, multimodal integration is dependent on the *congruence* of the perceived stimuli: when two or more stimuli arise from the same perceptual entity, like an audiovisual object for instance, or when they share perceptual

[1] That is, reactions caused by the stimuli themselves, in opposition to *endogeneous* ones as are caused in a goal-driven way.

attributes, like an audiovisual click[2]. Interestingly, when there is a conflict, that is an *incongruence* between auditory and visual information supposedly belonging to the same perceptual object, vision usually takes over the other modalities: a phenomenon (Hay et al., 1965) named *visual capture*. For instance, Pick et al. (1969) showed that the visual spatial position of an object is not alterable by incongruent auditory stimuli. According to the review on *visual capture* by (Posner et al., 1976), the reason why vision takes the lead on other modalities might be explained by the "relatively weak capacity of visual inputs to alert the organism to their occurrence." Thus, attention is preferably put on visual analysis to counterbalance the relative inherent lack of saliency of visual stimuli. However, it has to be kept in mind that the relative importance of visual dominance has been reconsidered by Spence and Driver (1994, 1996, 1997a,b) and later by Turatto et al. (2002). These findings are crucial for the understanding of how multimodal information is gathered and integrated. In fact, visual and auditory information are not considered equal in multimodal object formation and, consequently, with regard to potential reactions to their appearance in an environment.

As compared to visual scenes, auditory scenes are inherently more prone to salient objects. Nevertheless, some particular cases of auditory capture over vision have been observed and reported (see Gebhard and Mowbray (1959) for instance). A later hypothesis by Welch and Warren (1980) provides a plausible explanation of the underlying mechanisms of visual or auditory capture. Obviously, vision is particularly adapted to *spatial analysis* whereas auditory fits particularly *temporal analysis*. This hypothesis, called *modality appropriateness*, is based on the specifics of the sensors themselves.

More recently, Fendrich and Corballis (2001) used an experimental paradigm after Welch and Warren (1980) that led to the observation of a more pronounced effect of auditory capture versus visual capture. Interestingly, the authors have introduced the notion of *Intersensory Temporal Locking* (ITL), thus providing a more comprehensive explanation of the different observed phenomena of modal capture. The ITL, supported by a prior study of Scheier et al. (1999), is defined as a mechanism allowing the sensory cortices to solve potential temporal ambiguities in the perception of multimodal stimuli and offers a good basis for the understanding of when either auditory or vision lead perception, and what kind of stimuli triggers such modal capture.

In addition, the experiments of Shams et al. (2001) and Shams et al. (2002), both leading to the observation of auditory capture over visual capture, combined to the opposite results obtained a decade before by Saldana and Rosenblum (1993). Shams issued a statement as to which

> "The discontinuous stimulus in one modality alters the percept of the continuous stimulus in the other modality, yet not as strongly vice versa."

[2] An *audiovisual click* is a quick and simple sound, such as a pure tone section, presented together with a visual object, such as a dot or a cross of equal duration.

In summing up, all the studies mentioned above lead to the conclusion that multimodal perception consists of more than sole concatenation of auditory and visual data for forming the representation of multimodal objects in higher cerebral areas. The phenomena of auditory capture, visual capture, modality appropriateness, or discontinuity vs. continuity of perceived signals, indicate that auditory and vision are definitely working together closely, whereby each modality of the two mutually benefits from this advantage.

2.2 The Reverse Hierarchy theory

Consider the cases of the voice of somebody talking in a completely silent room in contrast to talking in a very crowded and noisy place (and again, we are close to the cocktail party situation). This raises the following question:

Are identical stimuli in different surroundings processed in the same way?

A recent model of perceptual information analysis, the *Reverse Hierarchy Theory* (RHT), puts the following insight to the fore. The informational context in which stimuli are perceived has an impact on the deepness and thoroughness of their analysis. RHT has been introduced and put into a formalized algorithm by (Hochstein and Ahissar, 2002; Ahissar and Hochstein, 2004), (Nelken and Ahissar, 2006) and recently (Nahum et al., 2008). The core of this theory is to bridge between high-level representations of perceived signals (such as auditory objects) to correlated low-level cues (such as frequency spectrum, ITD, or ILD). As to the latter ones, it is of interest whether these cues are necessary or not for taking high-level decisions, such as to initiate adequate motor actions. On the one hand, a rule is that the more difficult a discrimination task is, the more low-level attributes gain in relevance for refining auditory stream analysis, for example, for solving ambiguities. On the other hand, if the informational context is simple, the high-level representation of the perceptual streams (i.e. objects) will be usable directly, thus making deeper and more thorough analyses of the streams dispensable.

The RHT is thus also linked to internal representations of the world and, specifically, to the ways in which perceptual streams are combined to achieve a unified and robust perception of multimodal entities, perceptual objects. Indeed, as (Shamma, 2008) sums up,

> If the high "objects" and their "low-levels cues" are congruent, the feed-forward process is rapid, and the use of all available salient cues is effective and comprehensive.

Thus, in addition to the capabilities of perceptual streams analysis due to powerful features extraction, the ability to rapidly provide access to high-level representations of the perceptual world is quite astonishing as well. This is due to the fact that high-level representations include temporal integration and prior assumptions about incoming sensory information – see Sec. 2.3.

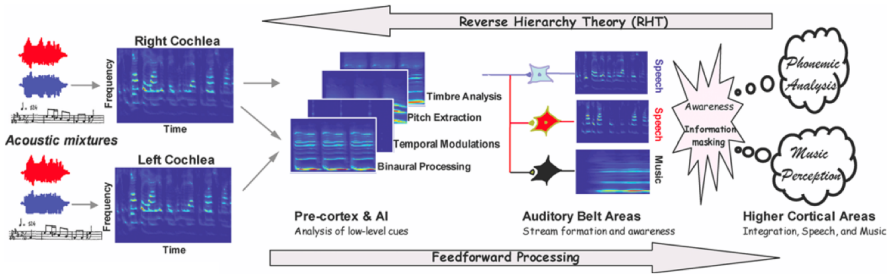

Fig. 2. Schematic of the Reverse Hierarchy Theory – after Shamma (2008)

Further, the RHT helps to understand attentional processes. In cases of incongruent perceptual streams such as, for instance, two males speaking in approximately the same spatial region, the theory postulates that the mechanisms that require low-level cues to disambiguate the two streams will be easily disrupted by competing cues. The way perceptual information streams are processed by the sensory areas of the brain, in particular those dealing with vision and audition, has for long been interpreted as almost exclusively being dominated by bottom-up processing. With the RHT however, there is now an innovative attempt for explaining the links between the traditional sensor-to-cortex pathway and the cortex-to-sensor one, namely, that they are activated depending on the complexity of the information to be processed. Consequently, RHT is of help when constructing artificial agents equipped with human-like perception. It suggest to process the data that such agents acquire with respect to the context in which they have been collected. In particular, it shows that making assumptions about *what is coming up next* in a scene can be useful for accelerating and simplifying the processing of sensory information. To be sure, the existence of such processes in humans implies that their brain has prediction abilities. And with these prediction abilities comes also the potential validation or invalidation of them by real perceived information. As for the other phenomena involved with attention, whenever there is a difference between 'what is expected' and 'what is perceived', a cerebral reaction, possibly followed by a motor one, is triggered. To illustrate this, the next section introduces the *Mismatch Negativity* phenomenon, a physiological reaction to deviant incoming sensory information with respect to the prediction made by these cortices.

2.3 Mismatch negativity

Can the apparition of stimuli be anticipated?

Anticipation, or prediction, is the ability to have a strong belief about what is coming up next. This ability has the potency of considerably accelerating

processing of perceptual data. Further, it enables the sensory cortices to detect inconsistent, salient and/or incongruent, objects. "Inconsistent, salient or incongruent" objects are such that somehow do not fit prior predictions. Consequently, they may require special reactions, such as a motor commands to redirect the sensors in order to get additional data that would help understand the origin of the observed unpredictability. As an example, imagine a strong male with an angry face uttering with high pitch and very calm voice: "Yester*phinge*, I was in the elephant". The following lists three cases in which an anticipation is initiated. Yet, it may turn out be wrong in the end.

1. The characteristics of the voice (an angry face would anticipated a loud, low-pitched voice)
2. The semantic content of the speech, that is, certain words have a higher probability to occur in the given context ("... in the elephant")
3. The words themselves, given the context and the initial syllables ("Yester*phinge*" instead of "Yesterday").

For all three cases the following holds. If what is perceived does not match prior expectation, a quick reaction is triggered. One of the first reaction to these unexpected objects occurring in a predictable stream of information can be observed in the sensory areas (such as the auditory or visual cortices) in terms of a particular neuronal response, the *Mismatch Negativity* (MMN).

This effect, when elicited, signals a quick attentional response to objects that do not match the expectations of the sensory areas. Discovered by Näätänen et al. (1978), the MMN can thus be described as a quick specific reaction to the *incongruence* of an auditory or visual object with regard to the short-term context in which it appears. MMN is particularly present in the auditory areas (Molholm et al., 2005), specifically in the temporal superior cortex and the frontal cortex (Alho, 1995). It occurs at around 100–200 ms after the deviant stimulus. For instance, when in a repeated sequence of sounds of a center-frequency of 1000 Hz, unexpectedly a sound at 1032 Hz is presented, it will be recognized as deviant from the predictable sequence perceived so far. The neuronal reaction to this deviant sound will show up as the MMN – see Figure 3. MMN has also been observed when there are amplitude or timbre variation ((Näätänen and Alho, 1995)). It is thus an effect that is linked either the apparition of new percepts or to variations in the perceptual attributes of ongoing ones.

Mismatch negativity is certainly an indication of a reaction to an unpredictable stimulus. Yet, its role in the formation of a perceptual world model has also to be seen under the following aspects. By being able on the basis of only three or four occurrences of a stimulus, to infer a rule that enables the prediction of the next stimulus to appear, the sensory cortices can speed up the processing of the incoming stream of stimuli by just checking if the actual perceived stimulus matches the prediction. If it does match, there is no need to fully process the stimulus, and computation time is saved (behavior to be linked to the RHT, see above). However, if it does, a warning signal (the

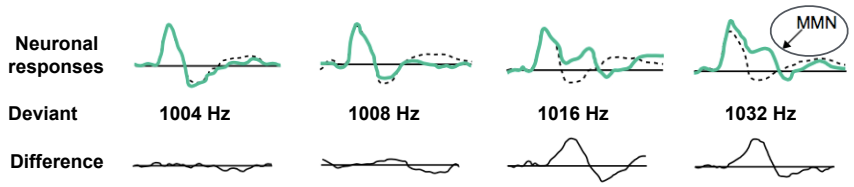

Fig. 3. *Mismatch negativity.* **Upper curve:** Neural responses recorded in 80 % of the occurrences of randomly presented sounds of 1000 Hz center frequency **(black dotted lines)**, of deviant sounds at different center frequencies in 20 % of the occurrences **(green lines)**. **Lower curve:** Differences of the responses to deviant sounds as compared to the 1000 Hz reference – after Näätänen et al. (2007).

MMN) is generated to potentially initiate a motor reaction such as, for instance, a head movement. This reaction is a way to motivate a deeper analysis of the unpredictable stimulus, for instance, by bringing other available and relevant sensors into play that gather additional information for the analysis.

Friston (2005) has highlighted the fact that the brain's internal representations of the world can be utilized to predict what most probably happens next in the environment. Along this line of thinking, Lochmann and Deneve (2011) introduced the notion of *predictive coding* for the causing of inference with regard to sensory objects that are not directly recognizable from sensory cues. Arnal and Giraud (2012) in their review of cortical oscillations and sensory predictions, listed several mechanisms that allow the auditory cortex to predict the point in time when a stimulus is most likely to happen in the given context. In fact, MMN accompanies all prediction processes in the brain. Yet, for two reasons these processes are more than just simple anticipation: (i), it makes the analysis of the perceptual scene faster and, (ii), it represents a powerful way of revealing unpredictable changes in the perceptual stream of information, especially in the auditory one.

Mismatch Negativity teaches us that when auditory (and visual) stimuli are processed, the sensory cortices are very soon able to form a *predictable sequence*, thus enabling instant detection of perceptual irregularities. But even more so, the MMN reveals already at this stage of sensory information processing, that there is no stimulus standing out per se: ~~As Fig. ?? illustrates schematically,~~ a stimulus can be detected as deviant, or *incongruent*, in a certain sequence of perceptual objects, but it would be rated as "normal" in different sequence. Consequently, the relative importance of a stimulus with regard to whether is will in fact trigger an attentional reaction, has to be defined in view of its relations to the current environmental surroundings. To be sure, these surrounding can be variable. Thus, stimuli may change their perceptual role accordingly. They will elicit different behavioral responses from one situation to another one, from one context to another, from one place to another, and/or from one point in time to another one. Whereas the MMN has not yet been directly linked to any direct motor reaction, its strong in-

volvement in attentional reactions (Escera et al., 1998, 2003) makes it a solid basis for triggering eye, body or head movement ~~generation~~ due to incongruent stimuli or objects, especially through the notion of *Saliency*, presented in the following section.

2.4 Saliency

Saliency is a measure of how much a stimulus, such as a sound wave or the pixel of an image, differs from what surrounds it, be it temporally or spatially. In human perception, saliency has mainly been studied in vision, as it is often the case for the visual system is easier to study than the audio one. In particular, following the definition of Treisman and Gelade (1980), saliency stems from local singularities that are exhibited within a stream of perceived data. For instance, within an image composed of numerous red circles, the presence of a unique green one would present a local singularity, in terms here of color: the green circle would then be considered as salient. From this analysis of the perceptual streams, and mainly exhibited by the auditory and visual cortices, attentional reactions can be elicited, such as eyes movements towards visual stimuli of high intensity ((Wolfe, 1994; Nothdurft, 2006)).

Moreover, saliency is shaped and influenced by learning and experience. For instance, while a musician is able to detect a false note instantly without even having to focus on listening, it could go by unnoticed an untrained person. In the visual system, the primary visual cortex (V1) already has a map of visual saliency (Li, 2002). Mazer and Gallant (2003) have shown that the activity of neurons of the extrastriated visual area (V4), a structure placed higher in the hierarchy of visual signals analysis, can predict towards which particular area in space an eye saccade will be directed an ongoing visual exploration task. This observation supports the assumption of presence of a topographical map of saliency in (V4). Further, the intra-parietal lateral area (Bisley and Goldberg, 2006) and the frontal eye field (Thompson and Bichot, 2005) have been associated with the phenomenon of visual saliency as well.

The human auditory system also respond well to saliency, and potentially also triggering motor reactions, in particular head and body movements. However, the attributes that the auditory sense is sensitive to, and on which it bases its interpretation of the auditory scene, are different from those used in vision. In particular, as concerns saliency, the auditory system mainly processes spectral and temporal modulations (Yost, 1992; Alain et al., 2001) and, based on these, it is able to extract auditory entities of relevance even in noisy environments (Hall et al., 1984). Addressed acoustic attributes are predominately such as spectral contrast, temporal contrast, and intensity. These are then exploited in parallel by neurons of the auditory areas, consequently leading the formation of saliency maps dedicated to specific attributes. These maps are then merged in order to create a global map of auditory saliency of the actual acoustical environment – compare Figure 4. Be it for the visual

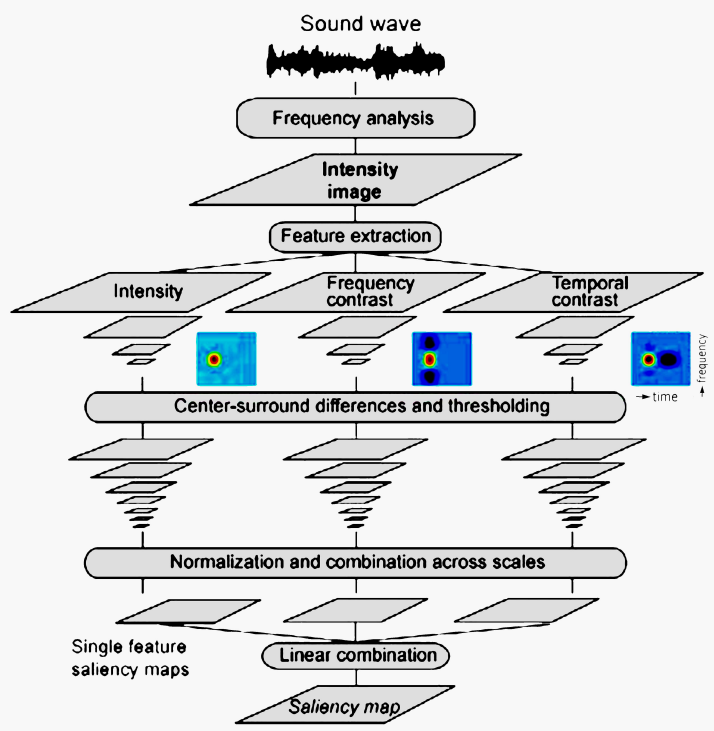

Fig. 4. Building an auditory-saliency map after Kayser et al. (2005), inspired by the work of Koch and Ullman (1985). An acoustic wave is received and then converted into a spectro-temporal representation allowing the extraction of attributes such as intensity, spectral and temporal contrasts. The resulting maps are combined into a comprehensive auditory-saliency map following a normalization step Figure processed

or the auditory system, the creation of saliency maps within dedicated sensory areas is an important step toward understanding the world models of the systems. Indeed, these maps provide potential candidates for a reorientation of attention. For instance, a person speaking at a certain azimuth outside the visual field of a listener and thus requiring a head movement, or a suddenly moving target that requires an eye saccade. These movements will lead to both a different capturing position of the audio sensors, thus refining the audio data processing, and to the bringing of the visual sensors towards the object of interest, thus allowing the capture of visual data that will greatly help the auditory system to raise potential ambiguities in the identification and/or localization of an object.

Saliency, be it visual, auditory or multimodal, has intensively been discussed, modeled and implemented in the robotics, artificial intelligence, and computational neurosciences communities (see for instance Koch and Ullman

(1985); Itti et al. (1998); Oliva et al. (2003); Kayser et al. (2005); Duangudom and Anderson (2007); Ruesch et al. (2008)). However, whereas it is a key component of the fashion in which low-level attributes shape motor reactions, it is not sufficient enough for a comprehensive understanding of attention reorientation. Indeed, by solely considering low-level attributes of the perceived stimuli, the means for including feedback from higher levels of the central nervous system are rather limited, though not impossible – compare Blauert and Brown (2018), this volume.

2.5 Conclusion of Section 2

How does the sensory areas of the brain deal with information coming from different sensors, each one having its own very particular characteristics, to the end of triggering relevant behavioral reactions to the incoming perceptual streams? In this section, this question was addressed by presenting four phenomenon that are part of the global and very complex mechanism of *attention*. These phenomena are, (i), the *Superior Colliculus* as a brain structure responsible for multimodal integration and consequent motor reactions, (ii), the *Reverse Hierarchy Theory* as an attempt to explain how the sensory areas compute stimuli differently given their level of ambiguity and the specific surroundings and, (iii), the *Mismatch Negativity* as a quick neuronal response to localize unpredictable perceptual objects, and (iv) *Saliency*, as a reaction to local singularities low-level characteristics of perceived signals are susceptible to exhibit. Each of them represents an important part of attention in multimodal perception. Integration, prediction abilities, detection of incongruences, selective in-depth analyses of the perceptual streams, and motor reactions are directly bound to each of these neuronal phenomena. The *active* component of perception is particularly relevant in this contexts. Indeed, whenever there are ambiguities in the understanding of an environment, motor reactions will enable the brain to access new information for refining its previous representation of the scene. In doing so, this additional information will help solving the previous ambiguities. At the same time, learning mechanisms will continuously increase the system's knowledge, and thus prepare the system for future similar tasks. For instance, when the position of an auditory object seems to be *odd*, that is, incongruent or unexpected, turning the head toward this object will initiated a redirection of visual sensors to an adequate position for better localization.

The next section will introduce a computational model rooted in the biological phenomena described here, and that provides to a mobile robot an attentional behavior.

3 Modulating the head movement – the HTM model

The previous section listed and described some important mechanisms playing a role in attention, perception and motor reactions to either incongruent,

salient or unpredictable events. From a technological point of view, several attentional systems have already been implemented. Most of them, however, share an important thing: they heavily rely on data that have been gathered before the robot even started its life, data for which dedicated learning systems have been specifically trained to solve very specific problems that the robot has not even encountered yet. But when considering the phenomena presented above, none of them rely on prior learning of specific skills: saliency is a property of the signals, the MMN is a very short-term reaction and is heavily adaptable, so as the RHT, and the SC computes a quick multimodal integration directly followed by a motor reaction depending on the content of the incoming multimodal information. Consequently, it should be possible to design an artificial system that implements the key features of human auditory (and visual) attention without having to priorly gather huge amount of training data that could help solve only one or few specific problem.

In this section will thus be described the Head Turning Modulation model, a model aiming at providing an answer to the central question of this chapter, which is

How can audition be used as a trigger for head movements towards objects of interest?

In this context, three different important aspects were presented with regard to the global phenomenon, *attention reorientation*, in which the aforementioned question is included. In the following, an attempt is described to provide a binaural and binocular humanoid robot with the ability to learn how to identify unpredictable auditory objects and, when appropriate, trigger head movements toward these objects for collecting supporting visual information. This model of high-level attention, recently introduced by the authors in Cohen-Lhyver et al. (2015, 2016, 2018) is mainly based on the four biological phenomena already discussed. The principal contributions of the HTM model are now outlined here with the a specific idea in mind, namely that some characteristic behavior of artificial agents can be achieved without having to deal with overly complex algorithms.

The section is organized as follows. The first part is dedicated to the description of the concepts that the HTM relies on, that is, especially the two modules that constitute it (the Dynamic Weighting model and the Multimodal Fusion & Insoutterference module). The second part introduces aspects of algorithmic formalization of the two different modules. Finally, a third part presents some of the results obtained in simulations and on a real robot.

3.1 Concepts and global architecture

The Head Turning Modulation acts similarly to a Blackboard system (Schymura (2018), this volume) and contains two principle modules[3]:

[3] That will be called Knowledge Sources when integrated to the Two!Ears software.

- The *Dynamic Weighting* module (DW) is deciding whether an audiovisual object appearing in the environment is *incongruent*, given the other audiovisual objects already detected in the past in this environment,
- The *Multimodal Fusion & Inference* module (MFI) is in charge of providing the DW module with corrected and completed *audiovisual classes* as a basis for the computation of congruence.

As shown on Fig. 5, auditory and visual labels provided by dedicated classification experts are exploited by the HTM for emitting hypotheses (i), on the *audiovisual class* the detected sources belong to and, (ii), on which of these sources the robot should focus. Each computational expert is dedicated to the detection and the recognition of particular *auditory labels* or *visual labels* (Two!Ears et al., 2012). For instance, one expert is dedicated to the detection and recognition of the sound `speech`, another one to the sound `barking`, still another one to the visual entity `male`, and so on. On this basis, each hypothesis might potentially lead to the triggering of head movements towards audiovisual sources of interest.

Importantly, these audiovisual sources appear randomly in the environment – that is, by not following any pattern the robot either understands or not and, consequently, can predict or not. By the way, triggering head movements towards any audiovisual source would not require any form of particular intelligence. The low-level attributes of the signals are often sufficient to localize the objects for sending meaningful motor commands. However, the goal here is to *modulate head movements*, that is, to either trigger *and* inhibit them. Indeed, not all of these head movements are relevant. For instance, turning the head toward the tenth `barking dog` in a room populated with only barking dogs, is very likely redundant such as not providing any useful additional information. Thus, by inhibiting some head movements, the head of the robot can be used for other kinds of movements, as may be requested by other tasks. The two modules constituting the HTM module have been designed and implemented in a way that they are able to understand the environment being explored by the robot in terms of audiovisual objects of *importance*. Thereby, the attribute of importance is assigned to objects in the following ways.

- The DW module implements the notion of importance through the concept of *congruence*. Congruence is defined here as *semantic saliency* since it is not applied to the low-level attributes of the perceived signals, such as spectral composition, ILD, or ITD, but rather on high-level representations of these signals, namely audiovisual classes. The classes $c(a, v)$ are made by the concatenation of an auditory label, a, and a visual label, v. On this basis, and without any prior knowledge of the actual environment, the DW aims at determining whether an audiovisual source is *incongruent* or not to the environment being explored. If it is, a motor reaction is triggered toward this audiovisual object. This is motor reaction can be

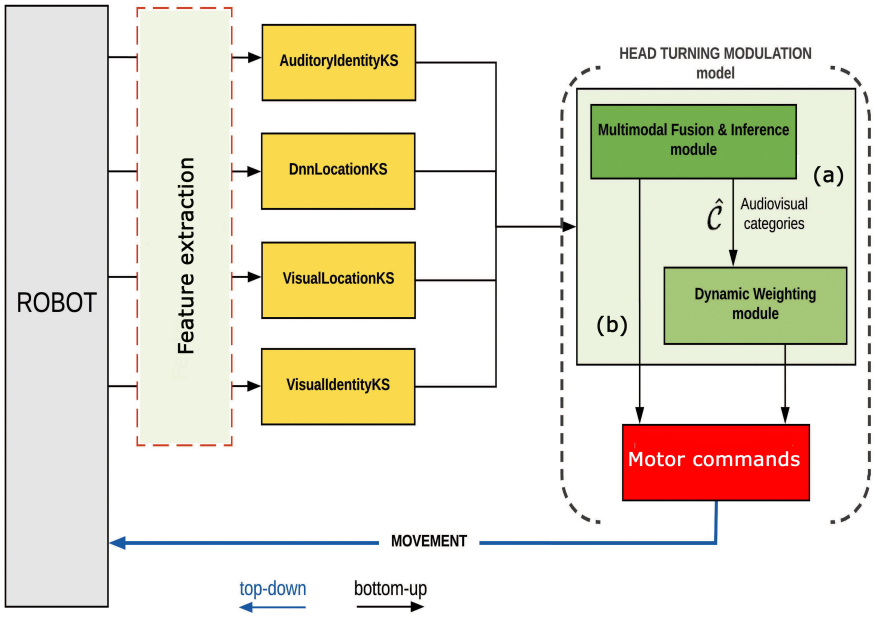

Fig. 5. Schematic architecture of the HTM model and its two main components. **(a)** The Multimodal Fusion & Inference module, in charge with providing to the DW module corrected audiovisual classes from classification experts outputs. **(b)** The Dynamic Weighting module, in charge with computing how *congruent* an audiovisual object actually is, given the environment that it is appearing in. Each of the two modules can trigger head movements separately. The **red box** depicts the computational component that realizes the combination of the different motor commands and puts them into an order to prioritize one of them, depending on the actual situation

compared to those triggered by the Superior Colliculus (see Sec. 2.1) or the MMN (see Sec. 2.3)

- The MFI module implements the notion of importance through the prism of *reduction of uncertainty* of the auditory and visual labels received from the classification experts. More precisely, the MFI module analyzes the uncertainty that it senses with regard to the combination of auditory and visual information, in particular, regarding the assignment of *audiovisual labels*, a combination that contributes to the multimodal representation of objects as used within the HTM. The MFI module is primarily based on the Reverse Hierarchy Theory (see Sec. 2.2). A further aspect originates from the principle of *Intrinsic motivation* of a person or an artificial agent to accomplish a particular action for the sake of an internal rewarding system, such as Berlyne (1950, 1954) have first described and theorized. Compare also Macedo and Cardoso (2001); Baranes and Oudeyer (2009,

2010) for examples of artificial systems furnished with such kind of motivations. In practice, the classification experts mentioned above are not unlikely to provide erroneous labels due to classification errors, or even missing labels due to occlusions, such as happens when objects are placed outside the field of view of the robot. Thus, the MFI module, being directly coupled with DW module, is in charge of providing the estimated audiovisual classes $\hat{c}$ the perceived objects might belong to. This analysis consists of a fusion of auditory and visual information as acquired by an active unsupervised learning algorithm, linked to the usage of head movements.

One potential issue arises here, that is, both modules have the ability to trigger head movements for their respective task. The MFI generates motor commands to acquire its multimodal representation, while the DW generates its commands for an attention-driven behavior directed to incongruent objects. Both head movements must then be assigned priorities – see the red box in Fig. 5. Since the DW takes decision based on congruence of audiovisual objects perceived, this information must be exempted from any classification or fusion errors. Thus, the motor commands triggered by MFI are prioritized against the ones triggered by DW. The following subsection provides details about the two modules constituting the HTM.

3.2 Algorithmic formalization

This section provides details of the algorithmic formalization of the two modules constituting the HTM, modules that respectively rely on *Congruence* and intrinsic motivation through reduction of uncertainty. As mentioned before, the HTM relies on the notion of *multimodal object* populating the environments the robot will explore. But this notion of object is not objective: it is already an interpreted notion arising from the convergence of different streams of information into a unified and coherent internal representation. Thus, considering that the environments are objectively populated with audiovisual *sources* emitting auditory, visual or audiovisual *events* Ψ_k, one of the first task of the HTM is to make emerge the notion of object, such as

$$\Psi_k = \{\theta_k, \mathtt{c}(\Psi_k)\} \longrightarrow o_j = \{\hat{\theta}_j, \widehat{\mathtt{c}}(o_j)\}, \text{ with } \widehat{\mathtt{c}}(o_j) = \{\widehat{\mathtt{c}}^a(o_j), \widehat{\mathtt{c}}^v(o_j)\}, \quad (1)$$

where $\mathtt{c}$ represents the real audiovisual class of the event Ψ_k, $\widehat{\mathtt{c}}$ depicts the estimated classes (audio, visual or both) the object o_j belongs to, θ_k the real angular position of the event, and $\hat{\theta}_j$ the estimated one by the localization expert. The estimated classes $\widehat{\mathtt{c}}$ come from the analysis performed by the MFI (see Section 3.2.2) of the data brought by the audio and visual classification experts that have been trained beforehand to identify particular sounds and images. And it is these audiovisual classes that will be utilized by the DW to compute the congruence of the concerned object. The raw data the HTM will retrieve from the Blackboard system will be organized as follows:

$$\mathbf{V}[t] = (\mathbf{P}[t], \boldsymbol{\Theta}[t]), \text{ with } \mathbf{P}[t] = (\mathbf{P}^a[t], \mathbf{P}^v[t]) \text{ and } \boldsymbol{\Theta}[t] = (\boldsymbol{\Theta}^a[t], \boldsymbol{\Theta}^v[t]). \quad (2)$$

where on the one hand $\mathbf{P}^a[t] = (p_1^a[t], \ldots, p_{N_a}^a[t])$ and $\mathbf{P}^v[t] = (p_1^v[t], \ldots, p_{N_v}^v[t])$ are the vectors of probabilities from the auditory and the visual classification experts, respectively; and on the other hand $\boldsymbol{\Theta}^a[t] = (\theta_1^a[t], \ldots, \theta_{N_\theta}^a[t])$ and $\boldsymbol{\Theta}^v[t] = (\theta_1^v[t], \ldots, \theta_{N_\theta}^v[t])$ are the vectors of probabilities from the auditory and visual localization experts, respectively. This is precisely the vectors $\mathbf{P}^a$ and $\mathbf{P}^v$ that the MFI will try to correct or, whenever one of them is missing, to infer.

The following section introduces the DW corresponding to the highest level, that is the closest to cognition abilities, module of the HTM.

3.2.1 Congruence – the DW module

Within the DW, emphasis has been put on dealing only with high-level representations of the perceived multimodal data, namely the auditory classes that they belong to. Following this idea, the aim of a reactive robot – in terms of head movements as driven by the congruence concept – is the detection of which audiovisual object in an unknown environment is incongruent as compared to what has been observed so far. The system has neither access to the content of the multimodal objects that populate this environment nor to their time of appearance. The only tool that the HTM has when entering a new room, is a set of classification experts that have been trained beforehand (provided by the TWO!EARS project[4]. Further, the DW is designed to exploit further relevant knowledge that is available to it for usage in other later explored unknown environments.

Congruence is based on a kind of conditional pseudo-probabilities where the probability of observing a certain audiovisual class depends on the environment in which it occurs. In other words, the less an audiovisual object has been observed in the past, the less likely it is to occur again in the future[5]. On the contrary, the more an audiovisual object has been observed in the past, the more likely it is to occur again in the future.

This has been formalized by means of the posterior probability of an object o_i to belong to a class $c^{(l)}(a_i, v_k)$ in the environment $e^{(l)}$

$$p\left(o_j \in c^{(l)}(a_i, v_k) \mid e^{(l)}\right) = p\left(c^{(l)}(a_i, v_k) \mid e^{(l)}\right) = \frac{|c^{(l)}(a_i, v_k)|}{N_l}, \quad (3)$$

where $|c^{(l)}(a_i, v_k)|$ depicts the number of objects that have already been associated to the audiovisual class $c^{(l)}(a_i, v_k)$, and N_l is the total number of objects detected so far. Since no information is available about what class is more likely to occur in a given environment, the probability $p\left(o_j \in c^{(l)}(a_i, v_k) \mid e^{(l)}\right)$ will be compared to the equiprobability $K_l = 1/|\mathcal{C}^{(l)}|$ of observing any class

[4] Software freely available from `www.twoears.eu`

[5] This obviously indicates a link to Bayesian theory.

detected so far. Thus, it is possible to take a decision on the congruence of the considered object by

$$o_j \in c^{(l)}(a_i, v_k) \text{ is incongruent} \Leftrightarrow p(c^{(l)}(a_i, v_k)) \leq K_l. \tag{4}$$

Following that, and to render the notion of *importance* of the emitting objects, two functions have been designed to assign weights to them with respect to their congruence

$$w_{o_j}[n] = \begin{cases} f_\omega^\bullet[n] = 1/(1 + 100\, e^{-2n}) & \text{if } p\left(c^{(l)}(a_i, v_k)\right) \leq K_l, \\ f_\omega^\circ[n] = (1/1 + 0.01\, e^{2n}) - 1 & \text{else,} \end{cases} \tag{5}$$

where $f_\omega^\bullet$ is an increasing positive function converging to 1 and dedicated to incongruent objects (high weight equals high importance), f_ω° is its symmetrically decreasing negative function converging to -1 and dedicated to congruent objects, and where n is a temporal index that is systematically reset whenever the congruence state of the object changes. To trigger a head movement, the object with the highest weight, that is, the most incongruent, will be considered as the target of the motor reaction. And if two objects share the same weight, the one that appeared the latest will be prioritized, thus applying a form of motivation by *novelty*. Note that the computation of the motor orders, not detailed here (see Cohen-Lhyver (2017) for complete description), is conceptually and mathematically formalized by the use of a GPR model (developed by Gurney et al. (2001a) Gurney et al. (2001b)) and inspired by the basal ganglia-thalamus-cortex loop present in humans and playing an important role in motor command selection.

All of this leads to the very definition of *environment*. In the robotics community, an environment is most often defined by its physical existence, its topographical characteristics (including the size of the room), the number of access points, usable paths, zones of danger, and light conditions. In the context of the DW (and, by extension, of the HTM), however, an environment is also understood in terms a semantic approach, namely, by the audiovisual objects that are present in it. Going even a bit further, a refined definition reads as follows

> *An environment is defined by the relative congruence of all the audiovisual classes that have been perceived in it*

In the vein of this definition, two very different rooms, such as two conference rooms at different universities, will be considered as being identical, if and only they share the same set of audiovisual classes congruence values. The respective status of congruence of the audiovisual classes detected in all the already explored environments consequently constitute the *knowledge* of the world the DW creates. This knowledge is used by the DW whenever it detects that the current explored environment is similar enough to one the robot already explored in the past. Being able to transfer acquired knowledge to new

unknown environments quickens the understanding of it by taking advantage of the past experience of the robot.

But taken that congruence relies on a multimodal representation of the objects perceived in the explored environments, what happens when an object is placed behind the robot, thus hindering it from acquiring adequate visual information? Turning the head toward the object to get the full data in order to properly compute congruence would definitely be absurd since if this object would be thereafter considered as congruent, a head turn would have already been triggered... Such conflicting situation motivated the creation of a Multimodal Fusion & Inference module, as described in the following section.

3.2.2 Reduction of uncertainty – the MFI module

To circumvent deadlock situation of the mentioned kind, a second module with the ability of *inferring missing data* has been developed. The Multimodal Fusion & Inference module also constitutes a reflective feedback loop in that it uses auditory and visual data coming from the sensors (after they have been processed by the dedicated classification experts) in order to send back a motor command, as illustrated in Figure 6. This motor command will give the robot access to new data that might this time redefine the best motor action for the robot. Once again, since the system relies solely upon high-level representation of the perceived data, namely, audiovisual classes, the inference made by the MFI module will be about auditory labels in view of known visual ones, or about visual labels in view of known auditory ones. Although a functional comparison between the MFI and the DW modules would make one think that they are performing identical actions—taking multimodal input and outputing a motor command—thus both being similar to how the Superior Colliculus works (see Section 2.1), the MFI module however differs from the SC in that that it is actually triggering a head movement whenever. **A FINIR**

To achieve this, it is necessary to learn the relationships between auditory and visual labels – such as `barking dog` or a `speaking male`. In other words, every time the robot faces an object which emits sound, the MFI module will take the chance to learn the audiovisual pair that is perceived. Once this learning has been accomplished, the MFI is able to offer an inference of a missing modality. However, it has to be kept in mind that classification experts are prone to errors. In particular the auditory experts, when the acoustic conditions become ~~be~~ challenging, such as in reverberant and noise surroundings—although the use of multi-conditional training (May et al., 2011), for instance, lowers their impact—, or when the explored environment differs to much from the one used for the prior experts training. Thus, relying too much on the output of these classifiers would lead to erroneous learning of the audiovisual pairs.

The MFI has been designed around a *Self-Organizing Map* (SOM), after Kohonen (1982). Such learning algorithm performs a vector quantization of

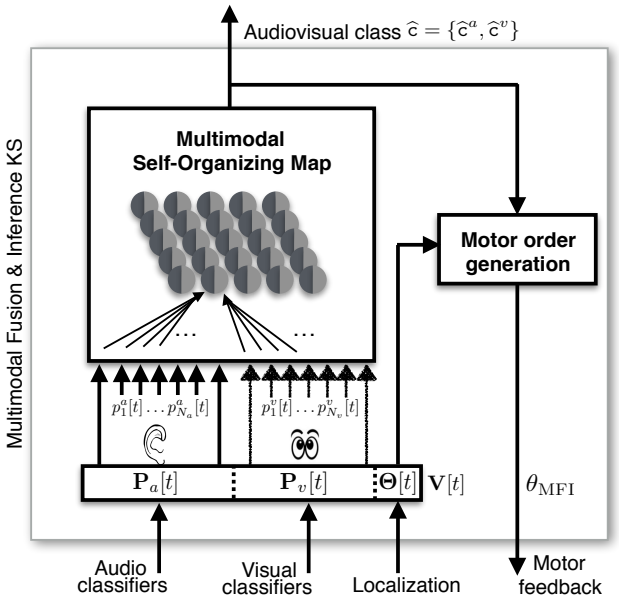

Fig. 6. Multimodal Fusion & Inference module architecture.

high-dimensional input data into a lower dimensional map (in our case, two-dimensional). Indeed, a SOM is a map composed of a certain number of nodes (or neurons) that represent the constituting vectors of the matrix of data to be processed. A SOM organizes these vectors in space by assigning them a particular node within the map. What results from this procedure is a modified representation of the input data as a map that has a lower dimension than the initial set of vectors, making it easier to process while also enabling the categorization of the input data. The SOM map is *tonotopically* organized. This means that when two regions of the SOM map are spatially close, the data that they represent are also close. The purposes of a SOM are organizing the existing data in clusters, then determining the class that a new input belongs to by localizing the node within the map which is most similar to the new vector, and finally, identifying the cluster that this node belongs to. But while the SOM algorithm provides a powerful unsupervised learning paradigm, it had to be adapted to the particular conditions in which the HTM, and the MFI in particular, has access to the data it has to process[6]

The first major change comes from the use of not only one SOM to learn the data, but of one SOM *per modality* used to define an object, thus creating the *Multimodal Self-Organizing Map* (M-SOM), as depicted in Figure 7. Here, auditory and visual data have been used to define an object. The overall M-

[6] Will only be presented here conceptually what has been changed. See Cohen-Lhyver (2017) for a thorough description of all the contributions of the M-SOM.

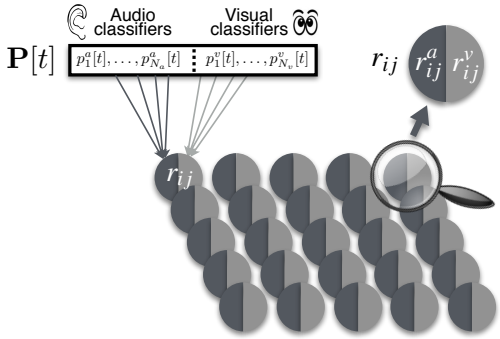

Fig. 7. Illustration of the Multimodal Self-Organizing Map which embeds two subnetworks, each of them being dedicated to coding the information from each modality used to define an object (audition and vision in our case).

SOM used in the MFI thus includes two interconnected subnetworks that will jointly participate in the creation of the internal representation of the robot's world, in terms of the audiovisual classes that have been observed during its exploration, as Figure 8 illustrates.

The second major change consists in modifying the learning process. Indeed, while the SOM is built, and usually used, to process full matrix of data, and since, as already stated before, the HTM does not have access to prior knowledge about the objects appearing in the environments, the M-SOM will only be fed with one vector of data at a time. That is, whenever a vector of data is available, the MFI has to be capable to integrate it in the M-SOM so that a learning iteration can happen. Obviously, since the goal of the MFI is to learn the relationship between the two modalities, a vector of data is sent to the M-SOM if and only if this data comes from both visual and auditory sensors and are about the same object, that is, whenever the robot faces an object emitting sound. And this is particularly here that the reflective feedback loop is present, through the triggering of head movements towards audiovisual sources of interest, in order for the robot to face these sources belonging to audiovisual classes that might need further learning.

Third, while in a traditional SOM the proofs of convergence are numerous, the problem the MFI has to solve does not imply one, or several, good solution: the robot being designed to always explore unknown environments, there is no possibility to know what are all the audiovisual classes that will be present. Consequently, the MFI implements the notion of *local convergence* of the M-SOM. In particular, the quality of the learning will be assessed by the MFI on a *class-by-class* manner: if the estimation of the audiovisual class an object is supposed to belong to is not trustworthy enough, more audiovisual data will be required in order to enhance the quality of the knowledge about this class. Such additional data is obtained by triggering a head movement towards the

concerned source. Local convergence is formalized by the implementation of an inference ratio $q(c^{(l)}(a_i, v_k))$ is used to determine whether, in an environment, $e^{(l)}$, an audiovisual class, $c^{(l)}(a_i, v_k)$, needs to be further learned by the M-SOM, or whether it has converged already to a trustworthy representation, according to:

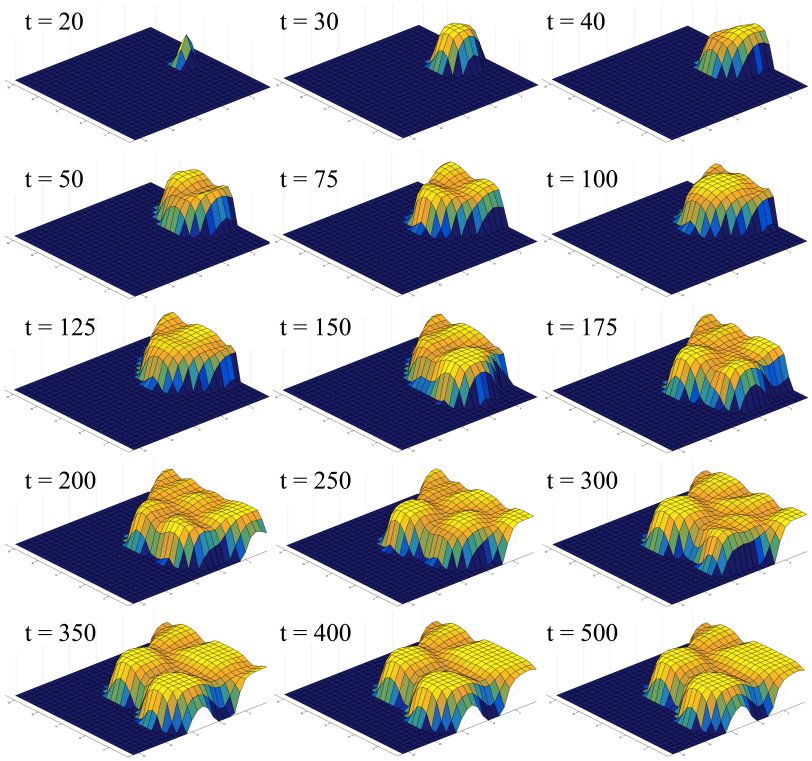

Fig. 8. *Multimodal Self-Organizing Map* (M-SOM). Each square represents a node (or a neuron) that codes a particular distribution of the input data to be analyzed. The figure shows the evolution of such a map during a 500 time steps in an experiment in simulated conditions. In the beginning the map is unorganized, and gradually, with the amount of data it is fed with, it creates clusters of neurons that represent similar categories of data. This M-SOM embeds two interconnected SOMs dedicated to each modality used to define the notion of object (audition and vision here). Four audiovisual classes have been created here, as the four highest regions of this map depict. The M-SOM is thereafter used to find the class of a new vector of classification experts data – after Cohen-Lhyver (2017).

$$q\left(c^{(l)}(a_i, v_k)\right) = \frac{\sum_{n=1}^{n=t} \delta_{i,k}^{\text{miss}}[n-1]\, \delta_{i,k}^{\text{all}}[n]}{\sum_{n=1}^{n=t} \delta_{i,k}^{\text{miss}}[n]},$$
$$\text{with } \delta_{i,k}^{\text{all/miss}} = \begin{cases} 1 \text{ if } \widehat{c}^{\text{all/miss}}(o_j) = \{a_i, v_k\}, \\ 0 \text{ else.} \end{cases} \tag{6}$$

Equation (6) describes the behavior of the MFI when it comes to setting up a hypothesis about a missing modality, and this hypothesis constitutes the reflective core of the feedback loop the MFI represents. If the ratio is too low, a command will be requested for turning the head toward the sound source in order to acquire visual data. By doing so at time $t + 1$, given that the data from the missing modality have now being available, the inference ratio $q\left(c^{(l)}(a_i, v_k)\right)$ will be updated with the new information and used to feed the M-SOM thus refining the learning. This ratio will then be compared to a dynamically changeable threshold $K_q \in \mathcal{R}^+ = [0, 1]$ to decide whether it is now high enough to accept the inference as trustworthy. If yes, no head movement will be initiated. If no, head movement will be triggered. The threshold has an effect on how quickly the MFI trusts its inference abilities.

For instance, a threshold of 0.2 would allow for eight out of ten wrong inferences on a particular class before stipulating that the inference is not trustworthy. Likewise, a threshold of 0.9 would require at least nine on ten good inferences before inhibiting head movements. The presence of such a threshold may suggests that it is solely responsible for the global performances of the MFI, but this not the case as it is explained later in this section. Extensive evaluation of the impact of the threshold value on the quality of MFI knowledge has revealed that variations are low for threshold values in the range of 0.5–0.9 (Cohen-Lhyver, 2017).

Thus, why staying with the option of setting different threshold values at all? This is the reason: The lower the threshold, the less head movements will be triggered but potentially more errors will be made. On the other hand, the higher the threshold, the more head movements will be triggered. Consequently, a suitable adaptation of the threshold can a make sense when considering the specific situation that a robot is actually exposed to. For example, in a search-and-rescue scenario the priority would be put on the search for victims, thus not requiring a full understanding of all audiovisual entities that are present in the current environment (low threshold), while in a room without any high priority task to accomplish, the robot has all the time needed for a complete exploration (high threshold).

Concerning the computation of motor orders potentially triggered by the MFI, it has been formalized similarly to the DW (see 3.2.1), that is through a GPR model enabling the selection of which object needs to be focused on.

To sum up, the main purpose of MFI is reduction of uncertainty about its knowledge using of motor reactions, hence implementing a reflective feedback loop that links information from classification experts to a motor command that will in return provoke the perception of new data, and so on. Therefore, two hypotheses are set up with regard to whether an audiovisual object

belongs to a certain class, in particular, one being based on the incoming stream of auditory labels and another one on the stream of visual labels. As an example, the robot may be facing a person and perceiving a barking sound originating from the same location: how confident is the MFI that this audiovisual source belongs to the audiovisual class `barking person`? The possible behavior of the MFI in such a case may alternatively be as follows.

1. The robot has encountered several (`barking person`) in the past and the MFI is now confident that it is not a classification error. The DW module can thus rely on this audiovisual fusion for computing the congruence of this audiovisual object
2. The robot has never encountered such a audiovisual class and will thus need to gather further auditory and visual data before potentially creating a new audiovisual class
3. The robot has already encountered this class but is still not confident enough whether the source does indeed belongs to it. In this case the MFI will initiate a head movement to gather more auditory and visual information.

3.2.3 Combination of the two modules

The combination of the two modules consisted mainly in dealing with which module should take the lead whenever both are triggering a head movement. Still staying within the paradigm of the GPR implementation of motor commands (see 3.2.1), the computation of the motor orders triggered by the DW has been slightly modified in order to take into account the activity of the MFI, so that, in fine, the MFI is prioritized over the DW. Indeed, the former being dedicated to provide the latter with clean data, it has to take over the DW until the MFI is confident enough in its knowledge (see 3.2.2). Combining the two modules leads to a global behavior of the HTM in three phases, as depicted in Figure 9. At first and until $t = 135$, the MFI is prioritized since it is gathering information and creating knowledge. Then, from $t = 135$ to $t = 310$, both modules trigger head movements: the MFI is confident in its knowledge about certain audiovisual classes (`speech male` for instance) but not about others (`crying female`). Finally, from $t = 310$ to the end of the simulation, the MFI does not trigger any head movement letting the DW in sole charge of deciding of the importance of the audiovisual objects present in the environment.

3.3 Experiments & results

In order to evaluate the HTM and its two modules, experiments in simulated and in realistic environments have been conducted. Simulations allow to modify the complexity of an environments and to focus only on the results of the analyses performed by the computational modules without taking into

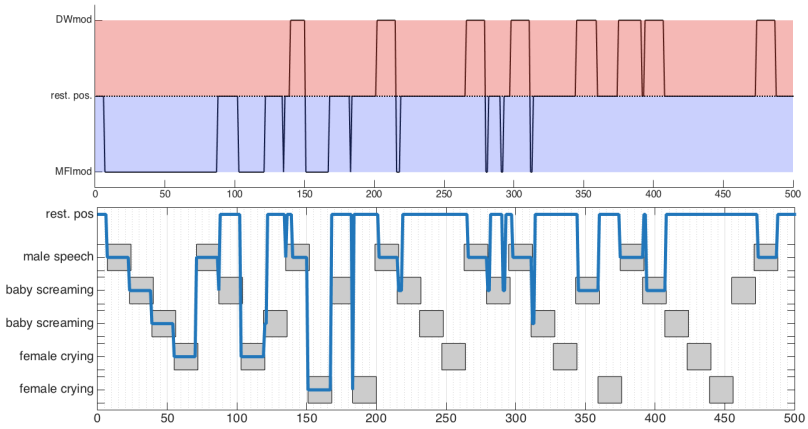

Fig. 9. Three-phase behavior resulting from the combination of the DW and the MFI. **Top:** Head movements triggered by the DW **(up, red)** or the MFI **(down, blue)**. **Bottom:** Time course of the scenario depicting which and when audiovisual objects are appearing in the environment. The blue line denotes the object to which the robot drives its attention.

account any hardware issues. Realistic environments are suitable to assessing performances of artificial system in the real world, that is, with real objects, classification and localization experts working in real-time on real data, and real robots with their mechanical limitations and imperfections.

This section briefly presents major results achieved with the HTM, first in simulated conditions and, secondly in a testing room where different environments are available. But before presenting these results, it is necessary to describe what will be evaluated, in both the simulations and in the real world.

3.3.1 The naive robot

The HTM model covers several fields of AI and robotic behavior, such as attention, learning, and perception. Moreover, robots endowed with head movements capabilities are rather rare and, as explained before, there is no *correct way* for a robot to behave – only something that could be qualified as *relevant* as compared to how human beings would behave. Thus, it was necessary to find a reference system to compare to when assessing whether a robot, when driven by the HTM, exhibits a "better" behavior than other systems. In the current study, a *"naive" robot* $\mathfrak{R}_n$ was employed for this purpose – also referred to as *naive system*. It is similar to the system that Girard et al. (2002) has used and has the following two main characteristics.

1. The naive robot does not perform any further analysis of the data that it gets from the classification experts than concatenating them, that is, the auditory and visual labels are taken from the experts *as is* without any temporal integration or deeper processing

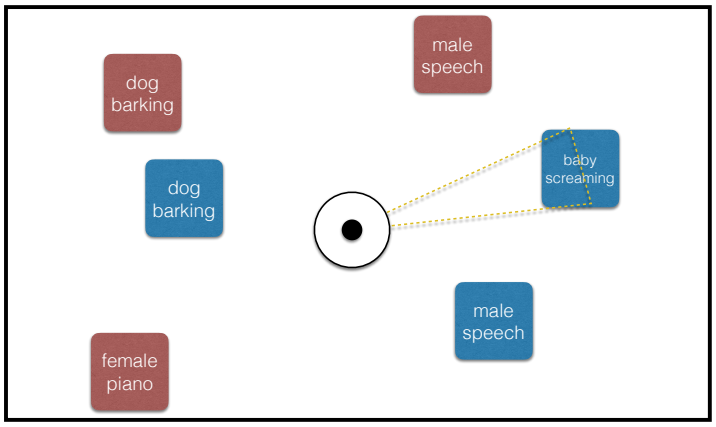

Fig. 10. *Illustration of a simulated environment.* The environments are populated with various audiovisual sources belonging to a certain audiovisual classes that the robot does not know beforehand. Some of them are emitting sound **(blue)**, others are silent, **(red)**. The ability of the robot to acquire "correct" knowledge about the semantic content of the scene is assessed on the basis of congruence of the perceived objects, either via the quality of audiovisual fusion or via the quality of head movements triggering or inhibition.

2. The naive system triggers head movements whenever there is a new audiovisual source appearing in the environment being explored. This behavior could be comparable with a simpler version of the motivation by saliency or novelty. In fact, every time a new object enters the scene, the naive system will guide the robot to focusing on it.

A robot driven by the HTM will thus be compared to this naive robot in terms of the quality of the classification and fusion of audiovisual data by the dedicated experts on the one hand, and the number of head movements triggered during the exploration of several environments on the other hand. Indeed, the HTM is a system that *modulates* the head movements by either triggering or inhibiting them. As a result of HTM employment a significant improvement of the quality of the data from the experts and a lower number of movements of the head is expected, while maintaining reasonable behavior in terms of the choices as to which objects in the scene should be focused on.

3.3.2 Simulations

The simulations consisted firstly in emulating the behavior of the classification and localization experts. The output of the simulated auditory and visual classification experts were emulated, with a certain error rate per frame included in the data generation process in order to reflect the real behavior of the real classifiers. Actually, in addition to a whole virtual environment the

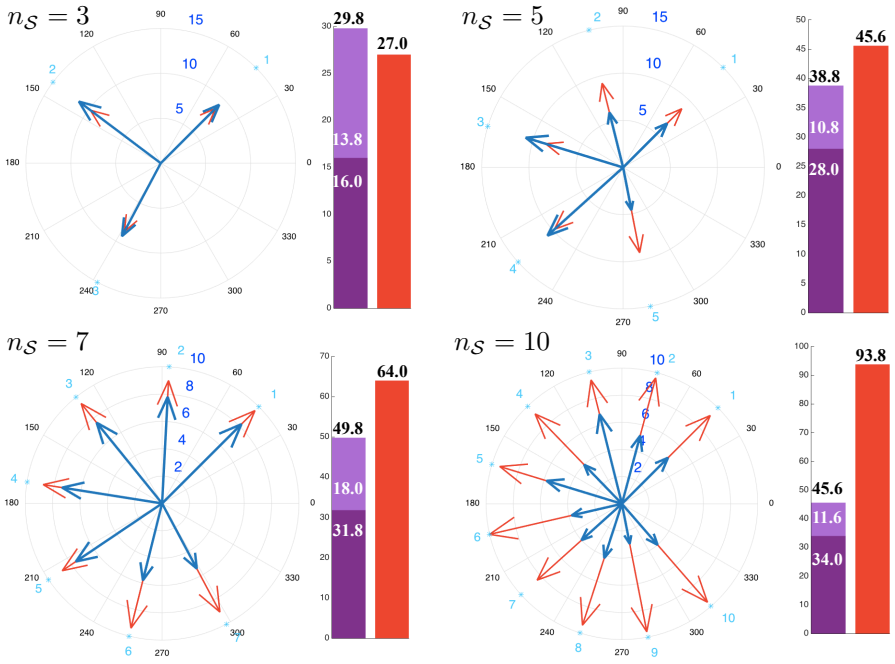

Fig. 11. Number of head movements generated in multisource scenarios. Movements generated by the HTM, **(blue)**, and by the naive robot, **(red)**. The arrows point to the positions of audiovisual sources, their length representing the number of movements toward the considered source. The histograms depict the total sum of generated movements, by the MFI, **(dark purple)**, by the DW, **(light purple)**, and by the naive robot, **(red)**. The **(white)** numbers correspond to the number of movements by module, averaged over five trials, and their sum, **(black)**.

robot explores, the simulation tool generate probabilities of an auditory/visual frame to belong to a certain auditory/visual class. Thus, one vector per modality, made of as many components as there are experts implemented, will be rendered at every time step. The simulated environments included different numbers of audiovisual sources which can appear anywhere and at any time for durations unknown to the system—see Figure 10. Two different general cases have been tested, namely, single-source scenarios with no concurring sound sources and multi-source scenarios but only the results from multi-source scenarios are presented here. These simulated environments were populated with $n_S = [3, 5, 7, 10]$ overlapping audiovisual sources, and a simulated error rate per audio and/or visual frame of $\varepsilon = 0.3$, meaning that for every audio/visual frame, there is a 30% chance that the highest audio/visual probability does not correspond to the correct audio/visual class. All numbers presented are the result of averaging over five runs for each scenario.

Figure 11 depicts the results obtained under multisource conditions. The histograms are the most interesting data to look at. They depict how many movements were triggered by the DW & MFI modules versus the naive robot. Interestingly, the more complex the environment gets, the more impact the HTM system has on the number of head movements. Indeed, in the scenario with ten audiovisual sources all emitting at the same time, the naive robot triggers up to 93.8 head movements while the HTM triggers only 45.6, that is less than half of them. Importantly, given the temporal dynamic of the HTM (the MFI module triggers head movements first, then, once it converged, lets the DW module take the lead), and the absence for now of an habituation mechanism for the DW module, the number of head movements triggered by this module will increase linearly and constantly with time. More precisely, this means that as long as an incongruent object keeps popping up in the scene, the DW module will continue requesting the behavioral head movement reaction towards it, despite the number of times it already requested this reaction. On the other hand, the MFI module, having a convergence criterion, see Equation (6), will not trigger any additional head movement as long as no new object belonging to a new audiovisual class appears in the environment. In Figure 11, and for the $n_{\mathcal{S}} = 10$ case, this dynamic implies that if the environment stays the same, in terms of audiovisual objects composition, the number of head movements triggered by the MFI module will cap at 11.6 (on average), whereas the DW module will continue triggering them as time goes on (and similarly for the naive robot).

3.3.3 Realistic environments

The experiments performed in realistic environments were conducted with the real robot in a pseudo-anechoic room where several audiovisual sources were placed. The auditory data were emitted by different loudspeakers with QR codes attached to them to identify them as visual objects. Three environments were tested as listed in ~~the~~ Table 1. The following audiovisual sources were employed: `barking dog`, `screaming baby`, `piano female`, `speaking male`. Moreover, with this model being dedicated to modulation of head movements, the scenarios did not include any whole-body movements but head movements only. Additionally, even though realistic experiments in conditions similar to the ones created for the simulations conducted above would have been ideal, the specific and numerous constraints of real robotic systems (and their underlying software components) make it impossible to recreate such complex scenarios. Especially in the case of audition, dealing with multi-source localization and/or recognition has shown to be particularly difficult in rooms that are different from the one where the training has been priorly achieved (which is our case here). That is why simpler environments have been set up, so as to enable relevant tests of the HTM system.

Importantly, one of the major role of the MFI is to *clean up* the data coming from the experts for they exhibit a certain amount of error per frame.

Test-scenario characteristics				
$e^{(i)}$	$n_{\mathcal{S}}$	n_{sim}^{max}	Present audiovisual classes	Angular position $\theta^{(a\|v)}$
1	3	1	barking dog #1	320°
			barking dog #2	35°
			speaking male	70°
2	3	1	crying baby #1	70°
			crying baby #2	35°
			piano female	320°
3	3	1	crying baby #1	70°
			crying baby #2	35°
			barking dog	320°
			speaking male	280°

Table 1. Specification of three scenarios created under real conditions for evaluating the HTM on a real robot.

Figure 12 shows the average number of good classifications versus fusion rate of the incoming audiovisual data as delivered by the experts, (red), and after the MFI computations, (blue). At the end of the exploration, the MFI provides an improvement of about 183.6% in the good classification rate, that is, from 37.9% to 69.6%. Taking only the labels as assigned by the classification experts would lead to the creation of multiple different audiovisual classes – as illustrated by Figure 13. This figure illustrates how the MFI considerably narrows the ensemble of possible audiovisual classes: from 22 detected by the experts, the MFI converges to only 5, that is a $\approx 78\%$ diminution. Within these five possible audiovisual classes, the `piano female` one is erroneous and should have been classified as `piano` **male**. But this wrong classification was output only two times, representing only 7.6% of the total number of time frames during which the correct `piano male` objects emitted sound. If the DW module had worked directly on the experts output, the results of congruence analysis would definitely be seriously corrupted. As an example, the `femaleSpeech alarm` audiovisual pairing has been positively detected six times by the naive robot, that is, by the identification experts. This association however never occurred in the explored environment. Its repeated detection, if no MFI module was involved in the prior data analysis and learning phase, would have led to a definitely erroneous behavioral reaction triggered by the DW module and exhibited by the robot's head movement towards the source. The usefulness of the MFI for the DW module and for the robot's internal representation at large, is thus convincingly demonstrated.

3.3.4 Discussion and conclusions of Section 3

The results presented here for simulated and realistic environments show that the HTM is able to drastically lower the number of head movements toward unpredictable audiovisual sources on the basis of *congruence* and *reduction of*

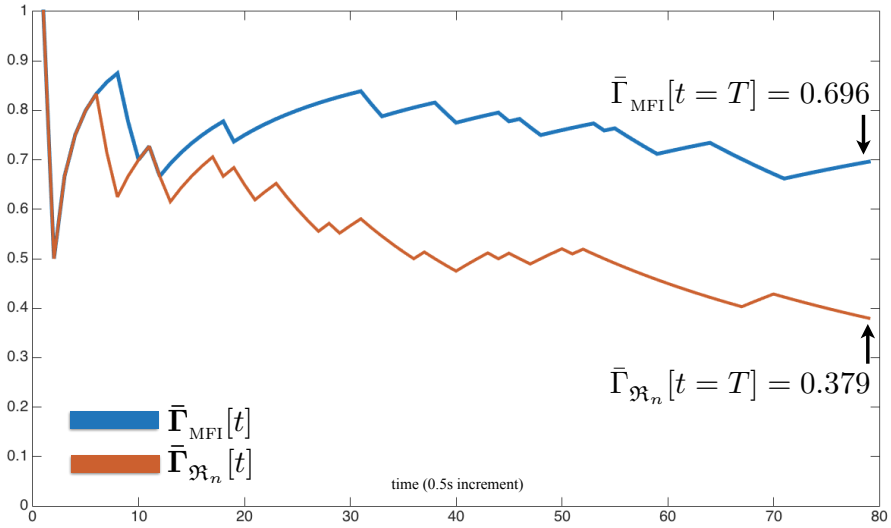

Fig. 12. Average good audiovisual classification rate as a function of time and computed on a rectangular sliding window (that is, at each time frame t, from frame 0 to frame t) for the MFI ($\bar{\Gamma}_{\text{MFI}}$), **(blue)**, and the naive robot ($\bar{\Gamma}_{\mathfrak{R}_n}$), **(red)**, that is directly at the classifiers output. The two numbers denote the final results at the end of exploration. Figure adapted from Cohen-Lhyver et al. (2018).

uncertainty based on computational analysis by the DW and the MFI modules. Modulating the generation of such head movements is of importance for achieving suitable means of behavior regarding the detection of what is of importance and what is not. These two modules enable mobile robots endowed with human like audiovisual perception to explore unknown environments and to react quickly and without prior knowledge to incoming audiovisual objects. The "How", "Where" and "When" of the objects to appear in the environments is unknown to the system – and thus to the robot. In combination, these two modules form the *Head Turning Modulation model* and constitute a complete system, which is working closely together with several experts (classification and localization) in order to establish a form of endogeneous attentional behavior in humanoid robots.

4 Final discussion and conclusion

Audition and vision are two major senses used by most mammals and humans in particular. Both exhibit incredible performances in perceiving and processing the world in their own way. The data that they uses are often very complex, be it spatially or temporally, and can change dynamically. The system of very sensitive sensors (eyes and ears) coupled to incredibly powerful means of analysis, such as dedicated sensory areas in the auditory and visual

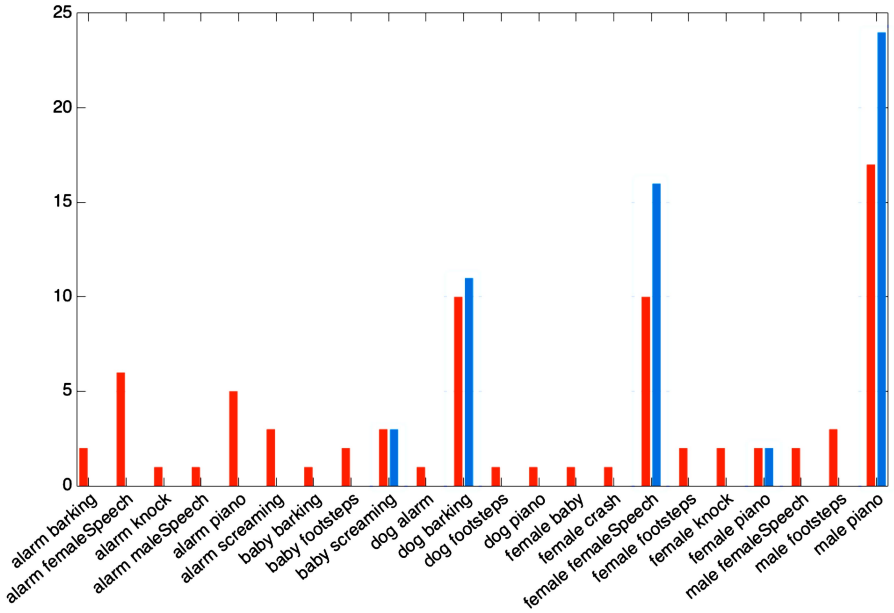

Fig. 13. Audiovisual classes created by the naive robot, **(red)**, and by the MFI, **(blue)**. The height of histograms depicts the number of frames for which the corresponding audiovisual class has been assigned. Figure adapted from Cohen-Lhyver et al. (2018). ~~The **light blue rectangles** highlight the audiovisual classes that are common amongst the two fusion systems~~.

cortices, make us understand the real world without too much of an effort. However, when trying to "simulate" such systems, as human-like robotics aim to do, audition and vision are often considered as two separate information channels. Moreover, it is rather rare to see artificial systems with an additional "cognitive" layer of multimodal integration allowing the robot to build a deeper internal representation of the world than just a collection of object labels. Also behavioral rules are often pre-determined by the experimenter leading to "if-else" statements kind of reactions, such as *If a baby is crying, go to the baby*. These kind of rules might be useful in simple scenarios and for robots with a short lifespan but whenever the robotic agent is put in more complex and varying environments, which have to be explored for longer periods of time (weeks, months, years...) the binary priorly defined rules cannot anticipate all the different objects prone to occur. In particular, relevant and comprehensive behavioral rules for guiding the exploration properly will not readily be available.

Thus, the idea of letting the robot create its own behavioral rules was central to the HTM model the proposed and described here. Inspired by several biological phenomena as are involved in the understanding of the audiovisual perceptual worlds, the HTM model is an example of how audition can be used

as a trigger for head movements towards particular audiovisual sources of interest, thus enable requisition of data from the visual modality for refining the perception of audiovisual sources of importance. In particular, the results presented in Sec. 3.3.3 provide evidence for the usefulness of multimodal integration of auditory and visual information for humanoid robot to explore unknown environments when prior knowledge of their audiovisual content is sparse. Moreover, the time needed for the robot to behave adequately and meaningful in unknown environments becomes significantly shorter in this way. Actually, only a few examples are enough for the robot to create its first behavioral rules, thus undermining the widespread misconception that real-time learning and the inability to quickly react in unknown conditions come in couples.

Certainly, the HTM model is far from being the only computational model that integrates several modalities in order to enrich the representation of the world models of robots (see Noda et al. (2014) for instance). But most current models rely on strong a priori knowledge gained from off-line learning in controlled environments, or such as are available in the form of "if-else" statements. Such paradigms often prohibit the robots from either learning more from what they experience, or from quickly adapting to situations that have not encountered before. Yet, the ability of doing so is one of the most powerful competences that human brains have, that is, to quickly adapt to odd situations, be they odd because of their unpredictability or because of their novelty.

Acknowledgment

This work has been supported by the European FP7 TWO!EARS project, ICT-618075, `www.twoears.eu`. We also thank an anonymous reviewer for their previous comments on this work.

References

Ahissar, M., and Hochstein, S. (**2004**). "The Reverse Hierarchy Theory of Visual Perceptual Learning," Trends in Cognitive Sciences **8**(10), 457–64, doi: `10.1016/j.tics.2004.08.011`.

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (**2001**). "'What' and 'Where' in the Human Auditory System," Proceedings of the National Academy of Sciences of the United States of America **98**(21), 12301–6, doi: `10.1073/pnas.211209098`.

Alho, K. (**1995**). "Cerebral Generators of Mismatch Negativity (MMN) and Its Magnetic Counterpart (MMNm) Elicited by Sound Changes," Ear and Hearing **16**(1), 38–51, doi: `10.1097/00003446-199502000-00004`.

Arnal, L. H., and Giraud, A.-L. (**2012**). "Cortical Oscillations and Sensory Predictions," Trends in cognitive sciences **16**(7), 390–8, doi: `10.1016/j.tics.2012.05.003`.

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., and Bizley, J. K. (**2018**). "Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding," Neuron **97**(3), 640–655.e4, https://doi.org/10.1016/j.neuron.2017.12.034, doi: 10.1016/j.neuron.2017.12.034.

Baranes, A., and Oudeyer, P.-y. (**2009**). "R-IAC : Robust Intrinsically Motivated Active Learning," IEEE Transactions on Autonomous Mental Development **1**(3), 155–169.

Baranes, A., and Oudeyer, P.-Y. (**2010**). "Intrinsically Motivated Goal Exploration for Active Motor Learning in Robots: A Case Study," in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Iros, Ieee, pp. 1766–1773, doi: 10.1109/IROS.2010.5651385.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (**2000**). "Voice-selective areas in human auditory cortex.," Nature **403**(6767), 309–312, doi: 10.1038/35002078.

Berlyne, D. E. (**1950**). "Novelty and Curiosity as Determinants of Exploratory Behavior," British Journal of Psychology **41**(1-2), 68–80.

Berlyne, D. E. (**1954**). "A Theory of Human Curiosity," British Journal of Psychology **45**(3), 180–191.

Bisley, J. W., and Goldberg, M. E. (**2006**). "Neural Correlates of Attention and Distractibility in the Lateral Intraparietal Area," Journal of Neurophysology **95**, 1696–1717, doi: 10.1152/jn.00848.2005.

Blauert, J., and Brown, G. (**2018**). The Technology of Binaural Understanding, Chap. Reflexive and reflective auditory feedback, 0–0 (Springer).

Cherry, E. C. (**1953**). "Some Experiments Upon the Recognition of Speech With One and Two Ears," J. Acoust. Soc. Am. **25**, 975–979.

Cherry, E. C., and Taylor, W. K. (**1954**). "Some Further Experiments upon the Recognition of Speech, with One and with Two Ears," J. Acoust. Soc. Am. **26**(4), 554–559.

Cohen-Lhyver, B. (**2017**). "Modulation de Mouvements de Tête pour l'Analyse Multimodale d'un Environnement Inconnu," Ph.D. thesis, University Pierre and Marie Curie.

Cohen-Lhyver, B., Argentieri, S., and Gas, B. (**2015**). "Modulating the Auditory Turn-to Reflex on the Basis of Multimodal Feedback Loops : the Dynamic Weighting Model," in IEEE Robio.

Cohen-Lhyver, B., Argentieri, S., and Gas, B. (**2016**). "Multimodal Fusion and Inference Using Binaural Audition and Vision," in International Congress on Acoustics.

Cohen-Lhyver, B., Argentieri, S., and Gas, B. (**2018**). "The head turning modulation system: an active multimodal paradigm for intrinsically motivated exploration of unknown environments," Frontiers in Neurorobotics doi: 10.3389/fnbot.2018.00060.

Corbetta, M., Patel, G., and Shulman, G. L. (**2008**). "Review The Reorienting System of the Human Brain : From Environment to Theory of Mind," 306–324, doi: 10.1016/j.neuron.2008.04.017.

Duangudom, V., and Anderson, D. V. (**2007**). "Using Auditory Saliency to Understand Complex Auditory Scenes," in European Signal Processing Conference, 15th.

Escera, C., Alho, K., Winkler, I., and Naatanen, R. (**1998**). "Neural Mechanisms of Involuntary Attention," J Cogn Neurosci **10**(5), 590–604, doi: 10.1162/089892998562997.

Escera, C., Yago, E., Corral, M. J., Corbera, S., and Nuñez, M. I. (**2003**). "Attention capture by auditory significant stimuli: Semantic analysis follows attention switching," European Journal of Neuroscience **18**(8), 2408–2412, doi: 10.1046/j.1460-9568.2003.02937.x.

Fendrich, R., and Corballis, P. M. (**2001**). "The temporal cross-capture of audition and vision," Perception & Psychophysics **63**(4), 719–725, doi: 10.3758/BF03194432.

Finney, E. M., Fine, I., and Dobkins, K. R. (**2001**). "Visual stimuli activate auditory cortex in the deaf," Nature Neuroscience **4**(12), 1171–1173, doi: 10.1038/nn763.

Friston, K. (**2005**). "A Theory of Cortical Responses," Philosophical Transactions: Biological Sciences **360**(1456), 815–836, doi: 10.1080/00222935708693955.

Gebhard, J. W., and Mowbray, G. H. (**1959**). "On discriminating the rate of visual flicker and auditory flutter," The American journal of psychology **72**(4), 521–529.

Girard, B., Cuzin, V., Guillot, A., Gurney, K. N., and Prescott, T. J. (**2002**). "Comparing a Brain-Inspired Robot Action Selection Mechanism With 'Winner-Takes-All'," in From Animals to Animats 7: Proceedings of the seventh international conference on simulation of adaptive behavio MIT Press, Vol. 7, p. 75.

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), 401–410.

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), 411–423.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis," The Journal of the Acoustical Society of America **76**(1), 50–56.

Hay, J. C., Pick, H. L., and Ikeda, K. (**1965**). "Visual capture produced by prism spectacles.," Psychonomic science .

Hochstein, S., and Ahissar, M. (**2002**). "View from the Top : Hierarchies and Reverse Hierarchies Review," Neuron **36**(3), 791–804.

Itti, L., Koch, C., and Niebur, E. (**1998**). "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259, doi: 10.1109/34.730558.

Iurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., Tucci, V., Benfenati, F., and Medini, P. (**2012**). "Sound-Driven Synaptic Inhibition in Primary Visual Cortex," Neuron **73**(4), 814–828, doi: 10.1016/j.neuron.2011.12.026.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (**2005**). "Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map," Current Biology **15**, 1943–1947, doi: 10.1016/j.cub.2005.09.040.

Koch, C., and Ullman, S. (**1985**). "Shifts in Selective Visual Attention: Towards the Underlying Nneural Circuitry," Human neurobiology **4**(4), 219–227.

Kohonen, T. (**1982**). "Self-Organized Formation of Popologically Correct Feature Maps," Biological Cybernetics **43**(1), 59–69, doi: 10.1007/BF00337288.

Li, Z. (**2002**). "A Saliency Map in Primary Visual Cortex," Trends in Cognitive Sciences **6**(1), 9–16.

Lochmann, T., and Deneve, S. (**2011**). "Neural processing as causal inference," Current Opinion in Neurobiology **21**(5), 774–781, doi: 10.1016/j.conb.2011.05.018.

Macedo, L., and Cardoso, A. (**2001**). "Modeling Forms of Surprise in an Artificial Agent," in Proceedings of the Cognitive Science Society, Vol. 23.

May, P. J. (**2006**). "The mammalian superior colliculus: laminar structure and connections," in Progress in Brain Research, pp. 321–378, doi: 10.1016/S0079-6123(05)51011-2.

May, T., van de Par, S., and Kohlrausch, A. (**2011**). "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," Audio, Speech, and Language Processing, IEEE Transactions on **19**(1), 1–13, doi: 10.1109/TASL.2010.2042128.

Mazer, J. A., and Gallant, J. L. (**2003**). "Goal-Related Activity in V4 during Free Viewing Visual Search : Evidence for a Ventral Stream Visual Salience Map," Neuron **40**, 1241–1250.

Meredith, M. A., and Stein, B. E. (**1986**). "Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration," Journal of Neurophysiology **56**(3), 640–662, doi: citeulike-article-id:844215.

Molholm, S., Martinez, A., Ritter, W., Javitt, D. C., and Foxe, J. J. (**2005**). "The neural circuitry of pre-attentive auditory change-detection: An fMRI study of pitch and duration mismatch negativity generators," Cerebral Cortex **15**(5), 545–551, doi: 10.1093/cercor/bhh155.

Moschovakis, A. K. (**1996**). "The superior colliculus and eye movement control," Current opinion in neurobiology **6**(6), 811–816.

Näätänen, R., and Alho, K. (**1995**). "Generators of Electrical and Magnetic Mismatch Responses in Humans," Brain topography **7**(4), 315–320.

Näätänen, R., Gaillard, A., and Mäntysalo, S. (**1978**). "Early Selective-Attention Effect on Evoked Potential Reinterpreted," Acta Psychologica **42**, 313–329.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (**2007**). "The Mismatch Negativity (MMN) in Basic Research of Central Auditory Processing: A Review," Clinical neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology **118**(12), 2544–90, doi: 10.1016/j.clinph.2007.04.026.

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), e126, doi: 10.1371/journal.pbio.0060126.

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 5–12.

Noda, K., Arie, H., Suga, Y., and Ogata, T. (**2014**). "Multimodal Integration Learning of Robot Behavior Using Deep Neural Networks," Robotics and Autonomous Systems **62**(6), 721–736, doi: 10.1016/j.robot.2014.03.003.

Nothdurft, H.-C. (**2006**). "Salience and Target Selection in Visual Search," Visual Cognition **14**(4-8), 514–542.

Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. (**2003**). "Top-Down Control of Visual Attention in Object Detection," IEEE International

Conference on Image Processing, September 14-17 **1**, 1–4, doi: 10.1109/ICIP.2003.1246946.

Pick, H. L., Warren, D. H., and Hay, J. C. (**1969**). "Sensory conflict in judgments of spatial direction," Attention, Perception, & Psychophysics **6**(4), 203–205.

Posner, M. I., Nissen, M. J., and Klein, R. M. (**1976**). "Visual Dominance: An Information-Processing Account of its Origins and Significance," Psychological Review **83**(2), 157–171, doi: 10.1037/0033-295X.83.2.157.

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," Proceedings - IEEE International Conference on Robotics and Automation 962–967, doi: 10.1109/ROBOT.2008.4543329.

Saldana, H. M., and Rosenblum, L. D. (**1993**). "Visual influences on auditory pluck and bow judgments," **54**(3), 406–416.

Scheier, C. R., Nijhawan, R., and Shimojo, S. (**1999**). "Sound Alters Visual Temporal Resolution," in Investigative Ophthalmology & Visual Science, Vol. 40, pp. S792—-S792.

Schymura, C. (**2018**). "Blackboard systems for modeling binaural understanding," in The technology of binaural understanding, edited by J. Blauert and J. Braasch (Springer and ASA Press, Heidelberg–New York–London), p. XXX, this volume.

Shamma, S. (**2008**). "On the Emergence and Awareness of Auditory Objects," PLoS biology **6**(6), e155, doi: 10.1371/journal.pbio.0060155.

Shams, L., Iwaki, S., Chawla, A., and Bhattacharya, J. (**2005**). "Early modulation of visual cortex by sound: An MEG study," Neuroscience Letters **378**(2), 76–81, doi: 10.1016/j.neulet.2004.12.035.

Shams, L., Kamitani, C. A. Y., Thompson, S., and Shimojo, S. (**2001**). "Sound Alters Visual Evoked Potentials in Humans," Cognitive Neuroscience and Neuropsychology **12**(17), 3849–3852.

Shams, L., Kamitani, Y., and Shimojo, S. (**2002**). "Visual Illusion Induced by Sound," Cognitive Brain Research **14**, 147–152.

Sharma, J., Angelucci, A., and Sur, M. (**2000**). "Induction of visual orientation modules in auditory cortex," Nature **404**(6780), 841–847, doi: 10.1038/35009043.

Spence, C., and Driver, J. (**1996**). "Audiovisual Links in Endogenous Covert Spatial Attention," Journal of Experimental Psychology: Human Perception and Performance **22**(4), 1005.

Spence, C., and Driver, J. (**1997**a). "Audiovisual Links in Exogenous Covert Spatial Orienting," Perception & Psychophysics **59**(1), 1–22, doi: 10.3758/BF03206843.

Spence, C., and Driver, J. (**1997**b). "On Measuring Selective Attention to an Expected Sensory Modality," Perception & Psychophysics **59**(3), 389–403, doi: 10.3758/BF03211906.

Spence, C. J., and Driver, J. (**1994**). "Covert Spatial Orienting in Audition: Exogenous and Endogenous Mechanisms," Journal of Experimental Psychology: Human Perception and Performance **20**(3), 555.

Stein, B. E., Jiang, W., and Stanford, T. R. (**2004**). "Multisensory integration in single neurons of the midbrain," The handbook of multisensory processes **15**, 243–264.

Thompson, K. G., and Bichot, N. P. (**2005**). "A Visual Salience Map in the Primate Frontal Eye Field," Progress in Brain Research **147**.

Treisman, A. M., and Gelade, G. (**1980**). "A Feature-Integration Theory of Attention," Cognitive psychology **12**(1), 97–136, doi: 10.1016/0010-0285(80)90005-5.

Turatto, M., Benso, F., Galfano, G., and Umiltà, C. (**2002**). "Nonspatial Attentional Shifts Between Audition and Vision," Journal of Experimental Psychology: Human Perception and Performance **28**(3), 628–639, doi: 10.1037//0096-1523.28.3.628.

Two!Ears, Ma, N., Trowitzsch, I., Kashef, Y., Mohr, J., Obermayer, K., Schymura, C., Kolossa, D., Walther, T., Wierstorf, H., May, T., Brown, G., Cohen-Lhyver, B., Danès, P., Devy, M., Forgue, T., Podlubne, A., and Vandeportaele, B. (**2012**). "Report on Evaluation of the Two!Ears Expert System," Technical Report .

Vetter, P., Smith, F. W., and Muckli, L. (**2014**). "Decoding sound and imagery content in early visual cortex," Current Biology **24**(11), 1256–1262, doi: 10.1016/j.cub.2014.04.020.

Welch, R. B., and Warren, D. H. (**1980**). "Immediate perceptual response to intersensory discrepancy.," Psychological bulletin **88**(3), 638.

Wolfe, J. M. (**1994**). "Guided Search 2.0 - A Revised Model of Visual Search," Psychonomic Bulletin & Review **1**(2), 202–238, doi: 10.3758/BF03200774.

Yost, W. A. (**1992**). "Auditory Perception and Sound Source Determination," Current Directions in Psychological Science **1**(6), 179–184.