Modulating the Auditory Turn-to Reflex on the Basis of Multimodal Feedback Loops: the Dynamic Weighting Model

Benjamin Cohen-Lhyver[†,⋆], Sylvain Argentieri[†,⋆] and Bruno Gas[†,⋆]

Abstract— **This paper is focused on the triggering of Spontaneous Head Movements (SHM), i.e. head movements which are supposed to be used to get additional information on a specific area of the robot environment. For that purpose, a Dynamic Weighting model (DWmod) is formulated as a low-level attention algorithm which allows an exploratory robot to drive its attention toward important items. DWmod is primarily based on auditory information, possibly coupled with visual data. These audiovisual characterizations rely on classification experts whose outputs are used by DWmod to trigger a SHM. The attention mechanism modeled by DWmod is rooted in the notion of congruence and predictability of items, allowing the robot to dynamically create its own rules about what is important or not in the current scene. This behavior is especially relevant in exploration tasks and particularly in search & rescue scenarios, where the robot has to react quickly with potentially no knowledge at all about the current environment. Within the TWO!EARS framework[1], a three-layered architecture on a simulated robot is developed, and simulation results demonstrate how the proposed model outperform basic low-level auditory-based turn-to reflexes.**

I. INTRODUCTION

In Robot Audition, a very common issue highlighted in the *cocktail party* effect [1] is the extreme difficulty to separate multiple sound sources and to make relevant classification [2], [3]. On the other hands, Humans can perform these tasks with very high performances. Indeed, besides our very powerful auditory and visual systems, head movements –be they spontaneous or self-willed– are known as a key mechanism allowing to process complex audiovisual scenes by driving our attention toward a narrow area of the environment. Audition is a key modality to achieve such movements, and can indeed serve as a trigger to turn the cameras toward specific sound sources of interest. There are numerous way to achieve such an action. Among them, lets consider two different so-called *reflex* movements. The first one would basically consists in a low-level analysis of the auditory scene, generally relying on sound source localization algorithms driven by signal-based features, for instance Time Delay of Arrival (TDOA), or Interaural Time [4] & Level Differences (ITD & ILD) [5] etc. Such an approach would result in *reflexive* movements making the system react to any sound present in the vicinity of the robot. While being effective in very basic scenarios involving only a few sound sources, such a behavior will not be consistent in realistic *cocktail party* situations. Introducing high-level information (like the speaker identity, the type of sound, etc.) could then allow the modulation of such a reflex. The resulting so-called *reflective* movements, involving high-level considerations on the scene, should then conduct to a system able to turn the head only toward specific targets. This paper is focused on these reflective movements, and aims at formulating this auditory turn-to reflex modulation in a robotics task through the definition of a Dynamic Weighting model (DWmod). Importantly, DWmod is being developed within the TWO!EARS framework[1]. Following the insight that exploration and search is a typical task for autonomous robots performing in rescue missions [6], the TWO!EARS project aims at developing an innovative system able to operate in complex environments for Search & Rescue tasks in a bioinspired fashion. The proposed DWmod will then be illustrated for such tasks, while not being limited to.

II. ATTENTION AND ACTION

This paper is focused on the modulation of spontaneous head movements (SHM). SHMs are supposed to be triggered to get additional information on an area of the environment where a lack of information exists. In our approach, SHMs are a response to unpredictable perceptual events[2]. This notion of event predictability has been taken into account in many robotic models, in particular those including the notion of attentional filtering, which will be introduced in a first subsection. Next, the roots of DWmod are presented through the notion of congruence in a second subsection. A short discussion ends this section.

A. Attentional filtering and reflective actions

The objective of the proposed DWmod is to trigger reflective actions (limited to head movements in this paper), i.e. to produce a movement which could allow a robot to focus its attention on important targets. This objective is very close to the design of attentional filtering systems, which have to be built with the following questions in mind:

- what is an important event?
- how to react with minimal knowledge about the environment?
- how to avoid the curse of dimensionality in a full world modelization approach?

The first question received a lot of attention during the last decade. In 2008, Ruesch et al. [7] successfully developed a powerful model of attentional filter based on the saliency of multimodal inputs. This algorithm provides the robot with the ability to detect what is the important event or

[†]Sorbonne Universités, UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005 Paris, France — name@isir.upmc.fr

[⋆]CNRS, UMR 7222, ISIR, F-75005 Paris, France

[1] www.twoears.eu

[2] By 'perceptual events' or 'perceptual objects', we mean every event that is acoustically or visually perceivable by an agent, be it a person walking or talking, or a glass falling.

object in a restricted environment, but does not take into account the context. More recent models take into account contextual features: Nguyen et al. in 2013 [8], and Ivaldi et al. in 2014 [9] have developed a powerful algorithm as a contribution to the cognitive architecture of the MACSi project[3]. Integrated on the iCub platform, their algorithm enables a robot to actively and interactively learn the objects populating the environment by experiencing actions on it. Behaviors relying on *Curiosity* and *Intrinsic Motivation* have thus been modelized to enable the robot to understand its environment in a more subtle and relevant way (see [10] for a proposal of a unified human motivations taxonomy). For instance, [11] defines *Uncertainty Motivation* as the attraction for novel stimuli; [12] defines *Information Gain Motivation* as the *pleasure of learning* that guides the robot to minimize the level of uncertainty of its knowledge of the environment; [13] defines *Empowerment Motivation* as a behavior that encourages the sequence of actions that will lead to the acquisition of the maximal amount of information by the robot's sensors (see [14] for a case study of intrinsically motivated robots in active exploration tasks). All the models developed on the basis of motivation show very good results on exploratory robots. However, whereas these models aim at determining a particular area of the environment to go to, or the next more relevant action to be performed in relatively simple environments, the proposed Dynamic Weighting model acts as a low-level attentional filter motivated by novelty detection. The ambition of the present work is to provide a simple *–but not simplistic–* algorithm that enables the robot to filter the environment it is experiencing *without any prior knowledge*.

B. The Notion of Congruence

As described above, DWmod aims at providing to a robot the ability to assign importance to objects in the environment it is experiencing. This is achieved by considering object apparition through the prism of *Congruence*. In Algebra, two plane figures are congruent if they share similar features (such as size and shape). From the point of view of Biology, the notion of congruence, and particularly its opposite, *incongruence*, is reflected by an electrical response called *Mismatch Negativity* (MMN, see [15] for a review). MMN occurs in every sensory area of the brain[4] when an odd stimulus arises among other predictible stimuli. MMN occurs at around hundred milliseconds after the onset of the odd stimulus. This quick reaction is considered to be an alert mechanism that leads to a quick behavioral response [16]. By extension of both these mathematical and biological definitions, congruence will thus be defined along

- the features shared by two perceptual objects, such as visual and acoustic labels;
- the links that exist between a perceptual event and a given environment.

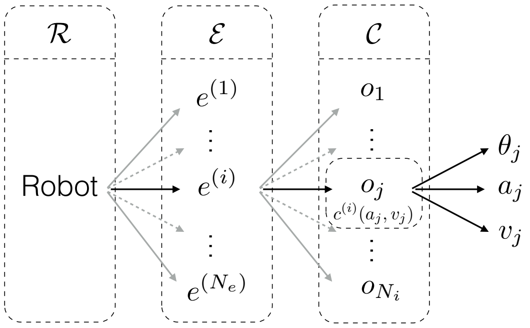

Fig. 1. Object oriented formulation of the proposed system. A robot is in an environment $e^{(i)} \in \mathscr{E}$ defined by multiple objects $o_j \in \mathscr{C}$ characterized by their position θ_j and audiovisual label a_j,v_j, see §III-B.

C. Discussion — Congruence: What For?

If an object has been detected as incongruent, a quick head movement will be triggered toward the direction of the object. This head movement will have several consequences about perception. The most striking evidence is for an audiovisual object emitting sound but which is out of sight. Making the head turning toward the object will (i) enhance the estimated position of the object, by updating its ITD and ILD; (ii) enhance its discrimination from other sound sources present in its surroundings (problem of the sound source separation, observed in the cocktail party problem); (iii) access the missing visual information of the object. This leads to a more accurate perception of the object.

III. THE DYNAMIC WEIGHTING MODEL

This section introduces DWmod, which is partially inspired by biological studies on the Mismatch Negativity and based on the notion of congruence (see Sec.II-B). The global architecture of the proposed system is overviewed in a first subsection. Its formalization is then proposed in a second subsection, the evaluation of the system being proposed in §IV.

A. Global Architecture

DWmod is implemented within an object-oriented framework organized in three layers, see Fig. 1. The simulated robot creates its internal representation of the world while experiencing different environments populated with several and various entities called audiovisual objects. Everytime an object is detected by the robot, the model computes its congruence inside the current environment, given what the robot has already experienced in the past. If the object has been characterized as congruent, the robot does not turn its head toward the object. At the opposite, if the object has been characterized as incongruent, the model will trigger a quick head movement toward the direction of the perceptual object. This head movement is then expected to help the robot to get more information about this object.

The proposed architecture of the overall TWO!EARS system is shown in Figure 2. Two different data paths can be brought to the fore. First, a traditional bottom/up approach is used to feed, on the basis on extracted signal features,

[3] http://macsi.isir.upmc.fr

[4] that is the areas processing sensory information such as audition, vision, smell, touch, and taste.

some high-level experts. Among them, one can cite speech or speaker recognition systems, sound classifiers, etc. (audio experts), but also face or object recognition algorithms, etc. (visual experts). All those experts work together to provide audiovisual labels used to understand the robot surrounding space. On this basis, the proposed DWmod builds an original top/down data path which will allow to take into account high-level information (the multimodal labels) to trigger a low-level reaction (a movement) which is expected to modify the raw signals from which are extracted the audiovisual features, thus forming a closed-loop system. Importantly, this paper is only focused on the top/down data path design, so that the traditional aforementioned bottom-up path is out of the scope of the paper. Consequently, the audiovisual labels will be brought by realistic simulated classification experts and will serve as inputs to DWmod. Based on these inputs, DWmod will enable the robot to create its internal representation of the world. This world is divided in the different environments it has experienced so far. Every environment is populated with perceptual objects that will be processed by DWmod by means of congruence characterization, as formalized in the following subsection.

B. Model Formalization

1) Definitions and notations: Let $\mathscr{R}$ and $\mathscr{E}$ be respectively the robot and environment sets, with

$$\mathscr{E} = \{e^{(1)}, e^{(2)}, \ldots, e^{(N_e)}\}, \tag{1}$$

where $e^{(i)} \in \mathscr{E}$ represents the i^{th} environment explored by $\mathscr{R}$, and N_e the number of considered environments. Each environment $e^{(i)}$ is defined as a set of Objects o_j such that

$$e^{(i)} = \{o_1, o_2, \ldots, o_{N_i}\}, \tag{2}$$

with N_i the number of detected objects in the environment $e^{(i)}$. Every object o_j is defined by its relative angle to the robot θ_j, an auditory label a_j and a visual label v_j, so that

$$o_j = \{\theta_j, a_j, v_j\}. \tag{3}$$

The relative angle θ_j of the object to the robot is provided by dedicated localization experts, i.e. sound localization algorithms providing an estimation of the source location on the basis on signal features. The multimodal labels a_j and v_j are estimated by dedicated experts, and are picked among the predefined collections of labels $\mathscr{A}$ and $\mathscr{V}$ respectively. For instance, on can have $\mathscr{A} = \{walk, cry, \ldots\}$ and $\mathscr{V} = \{person, cat, glass, \ldots\}$. Importantly, the audiovisual experts are not supposed as ideal and perfect, so that classification errors can be taken into account through the introduction of their error rates $\varepsilon_j^{(a)}$ and $\varepsilon_j^{(v)}$ respectively (see §IV). Let's now define the *audiovisual categories* $c^{(i)}(a,v)$ of the i^{th} environment by

$$c^{(i)}(a,v) = \{o_j \in e^{(i)}, a_j = a, v_j = v\}. \tag{4}$$

$c^{(i)}(a,v)$ basically represents the collection of objects sharing the same auditory and visual labels a and v respectively. All categories of the i^{th} environment are gathered into a set of categories $\mathscr{C}^{(i)}$ such that $\mathscr{C}^{(i)} = \{c^{(i)}(a,v)\}$.

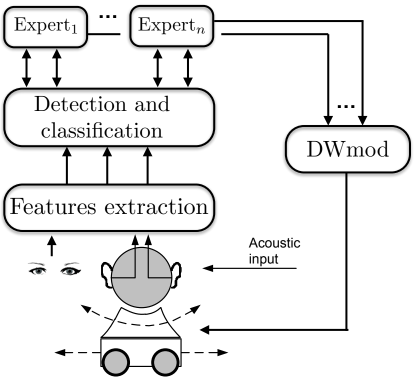

Fig. 2. Overall architecture of the TWO!EARS framework. The proposed DWmod is focused on the top/down data path.

2) Weights computation: In order to decide if an object o_j in the environment is of interest, each object is associated to a weighting function $w(o_j)$. In all the following, an audiovisual object o_j will be classified as incongruent *if other objects belonging to the same category $c^{(i)}(a_j, v_j)$ have not been detected by the system in the past*. This classification between congruent/incongruent objects will be in fact based on the object weighting function $w(o_j)$, with $w(o_j) \in$ [-1;1]. In all the following, $w(o_j) = -1$ represents a highly congruent object, while $w(o_j) = 1$ indicates a highly incongruent object. Note that the former case will not produce any movement of the robot, while the latter will trigger SHM in the direction θ_j.

First, and based on the previous definitions, lets define the pseudo-probability (i.e. the frequency) $p(c^{(i)}(a_j, v_j))$:

$$p\left(c^{(i)}(a_j, v_j)\right) = \frac{|c^{(i)}(aj, v_j)|}{N_i}, \tag{5}$$

with

$$\sum_{n=1}^{|\mathscr{C}^{(i)}|} p(c^{(i)}(a_n, v_n)) = 1, \tag{6}$$

where $|.|$ denotes the set cardinality. The pseudo-probability $p(c^{(i)}(a_j, v_j))$ can be considered as the likeliness of an object o_j belonging to category $c^{(i)}(a_j, v_j)$ to occur. On this basis, the weight $w(o_j)$ of the object o_j can then be defined as

$$w(o_j) = \begin{cases} 1 & \text{if } p(c^{(i)}(a_j, v_j)) < K_i, \\ -1 & \text{else,} \end{cases} \tag{7}$$

where K_i denotes a frequency threshold. Eq. (7) clearly shows the relation between a high object's weight $w(o_j)$ and a low probability of object's category's occurrence $p(c^{(i)}(a_j, v_j))$. Thus, if the object o_j appears in the current scene, it will be categorized as *incongruent*, and a SHM will be triggered. In all the following, the threshold K is set to

$$K_i = \frac{1}{|\mathscr{C}^{(i)}|}, \tag{8}$$

i.e. $w(o_j) = 1$ (which means that object o_j is incongruent) if the probability $p(c^{(i)}(a_j, v_j))$ is smaller than a random choice among equiprobable categories. But Eq. (7) exhibits

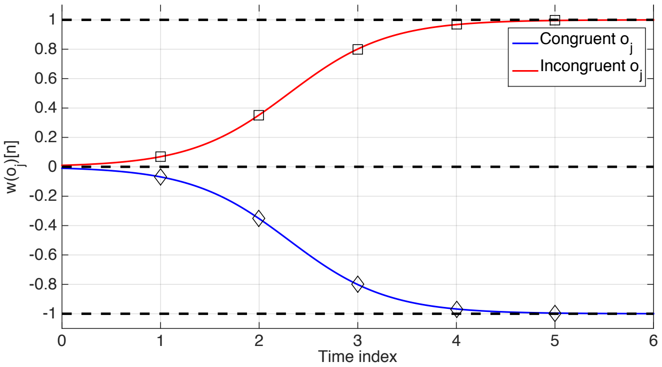

Fig. 3. Object weight $w(o_j)[n]$ as a function of time. Depending on the object congruence, one of the two functions is selected. Dots indicates the discrete time steps where values are selected.

a very naive weighting strategy. The proposed binary decision, while very simple, might indeed lead to inconsistent behavior when dealing with multiple simultaneous objects in the environment. Additionally, classification errors might introduce spurious temporary categories $c^{(i)}(.)$ which could be detected as incongruent while not relevant. Time filtering of the decision is thus introduced in the proposed weighting function so as to increase the robustness of the approach. Inspired by the N100 cortical wave pattern arising at around 100 milliseconds after the onset of an odd stimulus (see § II-B), a modified weighting function is proposed. Two smooth symmetric sigmoid functions lying in the range $[-1;1]$ are used to introduce a dynamic weighting $w(o_j)[n]$ of an object o_j, with

$$w(o_j)[n] = \begin{cases} 1/(1+100\,e^{-2n}) & \text{if } p(c^{(i)}(a_j,v_j)) < K, \\ 1/(1+0.01\,e^{2n})-1 & \text{else}, \end{cases} \tag{9}$$

where n represents the time frame index. $w(o_j)[n]$ is plotted in Figure 3 as a function of time index. For a frame length $T_w = 20$ms (this choice will be justified in the next section), then $w(o_j)[n] \approx 1$ at time $t = 100$ms, i.e. $w(o_j)[n]$ has been selected to mimic the dynamic of the biological Mismatch Negativity phenomenon (see §II-B).

3) Head movement decision: As a last step, once the weights $w(.)$ of a new object has been computed, one have to decide if a SHM has to be triggered. A motor order $\theta_m[n]$ is produced to turn the robot's head to an angle θ_j at time index n according to

$$\theta_m[n] = \begin{cases} \theta_j & \text{if } w(o_j)[n] > 0.98, \\ \theta_m[n-1] & \text{else}. \end{cases} \tag{10}$$

Threshold value 0.98 has been selected thanks to the weighting function $w(.)$ dynamic of 100ms, see (9).

The resulting pseudo-code of the proposed Dynamic Weighting model is shown in Alg.1.

IV. SIMULATIONS AND RESULTS

This section is devoted to the evaluation of the approach. Simulations are proposed to assess the proposed model in some specific scenarios related to the TWO!EARS framework, i.e. basic search and rescue tasks.

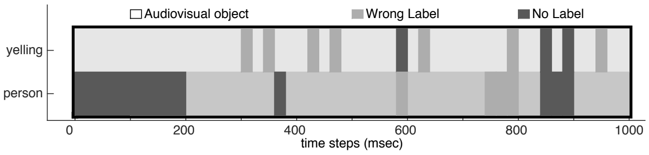

Fig. 4. Representation of an object o_j by the GEPE. Audiovisual labels are indicated, and classification errors can be introduced in the simulation.

A. Generator of Environments and Perceptual Events (GEPE)

GEPE is used to simulate environments populated with different (possibly randomly distributed) perceptual audio-visual objects along time. A standard perceptual object o_j is depicted in Fig. 4. This object is present in the simulated environment from $n = 0$ to $n = 1000$ time steps. As outlined in the previous section, each object is associated to two audio-visual labels a_j and v_j. Those simulated labels appear in the top and bottom of the object respectively: for instance, the object in Fig.4 is a yelling (auditory label) person (visual label). Considering that the visual label can only be estimated by the corresponding experts only if the object o_j is simulated in the field of view of the system, the corresponding visual label will generally not be available when the object appears in the scene (visual occlusion). Consequently, a visual label v_j will only be associated to the simulated object o_j if the system turns the head toward the object position θ_j. Otherwise, a dedicated empty label is associated to the object. This is highlighted in Fig. 4 by $v_j =$ "no label", which becomes $v_j =$ "person" at time index $n = 200$, under the assumption that the robot now see the object, i.e. a motor order has been triggered by the model toward θ_j. Visual occlusions can also easily be simulated by inserting $v_j =$ "no label" during the simulation, see Fig. 4. The same idea is exploited for the auditory label

Algorithm 1 Pseudo-code of DWmod

```
loop                                          ▷ % For each time index
    % Compute proba. of categ. and objects weights
    for all $c^{(i)}(a_j,v_j) \in \mathscr{C}^{(i)}$ do
        Compute $p(c^{(i)}(a_j,v_j))$ according to (5)
        Compute $w(o_j)$ according to (7) or (9)
    end for
    % Find the object associated to the maximal weight
    $[W, idx] = \max_j(w(o_j))$
    % Head turns toward object idx if $w(o_{idx}) > 0.98$
    Compute motor order $\theta_m$ according to (10).
    % Add current object category in category list
    if $c^{(i)}(a_j,v_j) \notin \mathscr{C}^{(i)}$ then
        $\mathscr{C}^{(i)} = \{\mathscr{C}^{(i)}, c^{(i)}(a_j,v_j)\}$
    end if
end loop
```

a_j, simulated by inserting $a_j =$ “no label” (depicting low energy signals, or silence frames in the speech, etc).

Additionally to those “no label” areas, some classification errors can occur. Indeed, the auditory and visual experts are supposed to be characterized by their error rates $\varepsilon_j^{(a)}$ and $\varepsilon_j^{(v)}$ respectively. GEPE simulates these errors by randomly introducing wrong labels picked among $\mathscr{A}$ and $\mathscr{V}$, so that

$$\begin{cases} p(a = a_j) = 1 - \varepsilon_j^{(a)}, \\ p(v = v_j) = 1 - \varepsilon_j^{(v)}. \end{cases} \tag{11}$$

Fig. 4 exhibits these classification errors by randomly assigning “wrong” auditory or visual labels to the object o_j (i.e. labels $a \neq a_j$ and $v \neq v_j$ respectively) according to Eq. (11). In all the following, a time step is supposed to correspond to a time frame length $T_w = 20$ms. This is usually the frame length used in speech/speaker/sound recognition systems and in visual processing. It is also coherent with weight computation dynamics (set to 100 ms, see §III-B.2): it will thus take five time steps to compute the decision of triggering a motor order. Consequently, $\varepsilon_j^{(a)}$ is analog to the frame classification error.

The next section is devoted to the evaluation of DWmod. However, we ran this evaluation on several other scenarios, with more or less complexity. Identical results were obtained. We thus chose the scenario described thereafter because of its simplicity, while being enough complex to highlight the behavior of the DWmod.

B. Evaluation

Let’s consider the following scenario, inspired by a search and rescue task. A robot endowed with auditory and visual sensing capabilities is placed in room with two persons talking together. Then, a faulty wiring (or anything else!) ignites some papers. One of the two persons yells, and then comes a third person in the room. In such a basic scenario, the unexpected event is of course the fire appearance, on which the robot must focus its attention to detect danger and possibly to alert fireman. Let’s precise that the robot has not learned anything at the moment: this scenario aims at illustrating how the system reacts during the very first steps of its exploration. Thus, the robot has no prior knowledge about the environment nor any congruence rules implemented. It will create these rules while observing this new environment and will immediately react according to them. For this scenario, the following sets $\mathscr{A}$ and $\mathscr{V}$ of audiovisual experts are used:

$$\begin{aligned} \mathscr{A} &= \{\text{talking}, \text{crackling}, \text{yelling}, \varnothing\}, \\ \mathscr{V} &= \{\text{person}, \text{fire}, \varnothing\}. \end{aligned} \tag{12}$$

The proposed scenario is depicted in Fig. 5, together with the head movements produced by (i) the binary weighting strategy (red curve, see Eq. (7)), (ii) the proposed DWmod strategy (green curve, see Eq. (9)) (see §III-B.2). Error rates of classifiers outputs are set to $\varepsilon_j^{(a)} = 40\%$ and $\varepsilon_j^{(v)} = 40\%$. The two weighting functions Eq. (7) and Eq. (9) are compared in terms of robustness. The resulting mean head movements are depicted in Fig. 5, computed after 1000 random runs of the scenario.

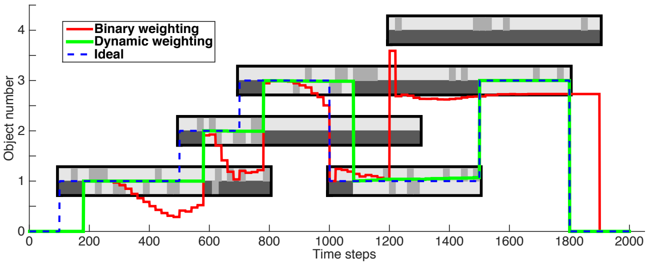

Fig. 5. Scenario 2. Mean head movements computed on 1000 random runs with classification errors. Red line denotes binary weighting strategy, green line denotes dynamic weighting one. Ideal movements are also indicated in blue for reference.

Two behaviors can be exhibited by the binary weighting function: *instability of focus* and *erratic detection of incongruent events.* Instability of focus is particularly highlighted in Fig. 5 from $n = 300$ to $n = 600$. The red solid line for these time steps depicts the oscillating behavior between object o_1 and the resting state (formalized by the object number 0). At the opposite, the green solid line depicts a highly stable behavior: the focus of the robot stays robustly focused on object o_1. Secondly, the erratic detection of incongruent events can be particularly observed between $n = 1200$ and $n = 1900$. At $n = 1200$, the binary weighting function induces the detection of o_4 (a *person talking*) as an incongruent event, while it should be detected as congruent. Then, from $n = 1500$ to $n = 1900$, the binary weighting function induces oscillations of focus between o_1 (a *person yelling*) and o_3 (a *fire crackling*) with a preference for the latter. At the opposite, the dynamic weighting function forces the robot to stay focused on o_1, which is the object to be targeted. The two behaviors –instability and erratic detection– are caused by the introduction of error rates $\varepsilon_j^{(a)}$ and $\varepsilon_j^{(v)}$, at a relatively high 40% level. As the binary weighting function makes the robot react each time a classifier produces an output, it is extremely sensitive to every change in classification of the perceptual events. At the opposite, the dynamic weighting function leads to a far more robust behavior, that is, not sensitive to quick and transitory classification errors.

Furthermore, an “ideal” behavior has been set up as a reference (depicted by the blue dashed line in Fig. 5). According to this targeted behavior, the robot should make exactly 6 head movements. Over the 1000 simulations, the binary weighting function induced 10.3 head movements whereas the dynamic weighting induced 6.1, together with great differences in variances (4.69 and 0.23 respectively). This also reflects the oscillations that appears with the binary weighting function and not with the dynamic weighting one.

C. Discussion

The proposed simulations, considering simple tasks in a limited simulated environment, clearly show how an audiovisual-based attentional filter can help to close the loop between traditional bottom-up signal processing approaches and more prospective top-down systems. In order to assess

DWmod in more complex realistic situations, a robotic simulation tool is being developed within the Two!Ears framework to evaluate all the other feedback loops that would be potentially included in the overall system shown in Fig. 2. A first version of this simulation tool, the Bochum Experimental Feedback Testbed (BEFT), is presented in [17] by the authors. The coupling between DWmod and BEFT is an ongoing work. Since BEFT provides a powerful system of environment simulation, including obstacles, path finding, changes in visual and auditory labels over time etc., the evaluation of SHMs triggered by DWmod will then be possible in more complex scenarios. Next, the coupling between auditory and visual labels is being investigated. Indeed, auditory labels can be used to form hypothesis on the visual labels, hypothesis which could then be confirmed by the visual experts once the head has turned toward the source of interest. This will allow to define the notion of object label congruence.

V. Conclusion

This paper was focused on the trigger or inhibition of spontaneous head movements, guided by high-level audiovisual experts continuously analyzing the scene. These head movements are of high importance for robotics systems operating in complex environments where a lot of perceptual events can occur. The proposed approach allows to close the loop with traditional bottom-up signal processing techniques by inserting a Dynamic Weighting model in the top-down data path. Results show that the resulting triggered head movements allow the robot to be focused on incongruent events by maintaining its head toward their directions. Importantly, DWmod dynamically updates its decision by computing object weighting functions, which are not defined with an a priori knowledge on the scene, thus making the approach fitted for unknown environments exploration, and especially for Search & Rescue scenarios.

In §II-A, three main questions were raised considering attentional filter models:

- what is an important event?
- how to react with minimal knowledge about the environment?
- how to avoid the complexity of a full world modelization?

DWmod addresses the first question by defining an important event as a rare event. Based on probabilities of occurrences of perceptual events, DWmod addresses the second question by enabling the robot to react with very few information. Finally, the third question is addressed by the modularity of DWmod. Currently, DWmod takes two objects characteristics into account: a visual label and an auditory label. With only these two data, the simulated robot can react properly to its environment. However, it is possible to take more characteristics into account, such as the gender of the person, the action of the object (running, walking, falling, ...), its color, its shape, its material, the emotions detected on persons' faces... Thus, the Dynamic Weighting model constitutes an adaptive and powerful tool that enables a robot to reduce the flow of information by selectively turning its head toward an event of interest.

Acknowledgment

This work is supported by the European FP7 Two!Ears project, ICT-618075, `www.twoears.eu`.

References

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, pp. 1875–1902, Sept. 2005.

[2] J. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *Robotics, IEEE Transactions on*, vol. 23, pp. 742–752, Aug 2007.

[3] M. C. Woelfel and J. McDonough, *Distant speech recognition*. Chichester: Wiley, 2009.

[4] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source {TDOA} estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950 – 1960, 2012. Latent Variable Analysis and Signal Separation.

[5] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1–13, Jan 2011.

[6] D. Calisi, A. Farinelli, L. Locci, and D. Nardi, "Multi-objective Exploration and Search for Autonomous Rescue Robots," *Journal of Field Robotics*, vol. 24, no. 8/9, pp. 763–777, 2007.

[7] J. Ruesch, M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 962–967, 2008.

[8] S. M. Nguyen, S. Ivaldi, N. Lyubova, A. Droniou, D. Gerardeaux-Viret, D. Filliat, V. Padois, O. Sigaud, and P. Y. Oudeyer, "Learning to recognize objects through curiosity-driven manipulation with the iCub humanoid robot," *2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013 - Electronic Conference Proceedings*, 2013.

[9] S. Ivaldi, S. M. Nguyen, N. Lyubova, A. Droniou, V. Padois, D. Filliat, P. Y. Oudeyer, and O. Sigaud, "Object learning through active exploration," *IEEE Transactions on Autonomous Mental Development*, vol. 6, pp. 56–72, 2014.

[10] R. M. Ryan and E. L. Deci, "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions.," *Contemporary Educational Psychology*, vol. 25, pp. 54–67, Jan. 2000.

[11] X. Huang and J. Weng, "Novelty and Reinforcement Learning in the Value System of Developmental Robots.," 2002.

[12] N. Roy, A. Mccallum, and M. W. Com, "Toward Optimal Active Learning through Monte Carlo Estimation of Error Reduction," in *international conference on Machine Learning*, 2001.

[13] P. Capdepuy, D. Polani, and C. L. Nehaniv, "Maximization of Potential Information Flow as a Universal Utility for Collective Behaviour," in *IEEE Symposium on Artificial Life*, pp. 207–213, Ieee, Apr. 2007.

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

[15] R. Näätänen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (MMN) in basic research of central auditory processing: a review.," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 118, pp. 2544–90, Dec. 2007.

[16] L. H. Arnal and A.-L. Giraud, "Cortical oscillations and sensory predictions.," *Trends in cognitive sciences*, vol. 16, pp. 390–8, July 2012.

[17] T. Walther and B. Cohen-Lhyver, "Multimodal Feedback in Auditory-Based Active Scene Exploration," in *Forum Acusticum*, 2014.