ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

The Technology of Binaural Listening & Understanding: Paper 715

Multimodal fusion and inference using binaural audition and vision

Cohen-Lhyver Benjamin[(a,b)], **Argentieri Sylvain**[(a,b)], **Gas Bruno**[(a,b)]

[(a)]Sorbonne Universités, UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005 Paris, France, name@isir.upmc.fr
[(b)]CNRS, UMR 7222, ISIR, F-75005 Paris, France

Abstract:

Hearing is a key modality on which several perceptual human processes rely on. Together with vision, these two modalities offer a 360 degrees wide, highly sensitive, quickly adaptive, and incredibly precise system of perception of the environment. In an exploratory robotics context, the concept of audiovisual objects is very relevant for a robot since it enables it to better understand its environment, and also to interact with it. However, how to face the cases when an object is out of sight, or when it does not emits sound, that is, the cases of missing information? The proposed Multimodal Fusion and Inference (MFI) system takes advantages of having (i) multimodal information and (ii) the ability to move in the environment, to implement a low-level attentional algorithm that enables a mobile robot to understand its environment in terms of audiovisual objects. In the case of a missing modality, the proposed algorithm is able to infer the missing data thus providing to the robot full information to higher cognitive stages. The MFI system is based on an online and unsupervised learning algorithm using a modified self-organizing map. Furthermore, the MFI exploits the ability to turn the robot head towards objects, thus benefiting from active perception to reinforce autonomously what the system is actually learning. Results exhibits promising performances in closed-loop scenarios involving sound and image classifiers.

Keywords: robot, audiovision, learning, head movements

Multimodal fusion and inference using binaural audition and vision

1 Introduction

The ability of a mobile robot to autonomously travel around an unknown environment is most often called *exploration*. It has to extract by itself some information originating from the environment through the use of its own sensors. The work presented in this article is focused on a specific case of exploration called *Search & Rescue scenario* (S&R). In such cases, assessment of the situation is needed to successfully launch search and rescue missions [3]. The current work is rooted in the TWO!EARS project [11], whose goal is to develop an intelligent and active computational model of auditory perception in a multimodal context. The final system is expected to assign *meaning* to acoustic events. The overall model will be implemented on a robotic platform able to actively explore environments, orientate itself in it, and move its sensors in a humanoid manner. The system embedded in this robot will be evaluated in a search & rescue mission. Let's precise that the current work will be limited to the simulation of the TWO!EARS components, since some components are not available yet. In S&R situations, **event significance** (or meaning) becomes highly relevant. Indeed, all events occurring in the environment are not equal in terms of importance to the mission to be accomplished by the robot. This problem has already been addressed through object learning [7], object-based attention [10] (see [4] for a survey), or motivation for exploration [1]. However, most of these contributions don't address the question of relevance of the object to be explored. In other words: everything in the scene is of equal importance. Another important point is the sensory modality considered. In the references cited above, most of the objects are exclusively *visual*, whereas they are in the reality mainly multimodal (and often particularly audio-visual), such as persons talking or yelling, fire crackling... Classical topographical maps of environments aim at representing physical objects like obstacles, rooms, corridors, danger zones.... Here, the goal is to build a topographical map incorporating what lies in the environment that is being explored in terms of **objects** or **events** populating it. The learning of such a map is done in an **unsupervised** way by mean of a multi-modal self-organizing map (M-SOM) that uses an **online** algorithm. By doing that, the system can autonomously learn and adapt quickly when facing new situations in unknown environments. The authors have already partially adressed the importance of an autonomous, unsupervised, audio-visual object learning in previous papers [5, 12] through the development of the Dynamic Weighting model (DW) based on the notion of *Congruence* of audiovisual events. The Multimodal Fusion and Inference model (MFI) presented here is a part of a more global *Head Turning Modulation* (HTM) model which includes both the MFI and the DW models. HTM aims at modulating head movements caused by the apparition of events during exploration of unknown environments. The main objective of the whole system is to learn the cases when triggering a head movement is relevant for the robot's exploration task. The paper is organized as follows. Important definitions are outlined in a first section, followed by the formalization of the proposed MFI system in a second section. Its effectiveness in two different conditions is then assessed in a last section. Finally, a conclusion ends the paper.

2 Architecture of the Multimodal Fusion & Inference model

The proposed active multimodal inference system is rooted in the TWO!EARS framework [11]. The framework will be implemented and evaluated on a mobile robot system that can interactively explore its environment based on audio-visual information. As stated above, since some components required to test the proposed model are not yet available in the TWO!EARS software, the evaluation of the model has been made on simulations. This also enabled us to perform more complex tests. In a first subsection, the generic TWO!EARS framework will be quickly introduced in order to allow the reader to capture the main objective of the proposed inference system with respect to its implementation within a larger framework. This section ends with a description of the proposed Multimodal Fusion and Inference system.

2.1 The TWO!EARS architecture

The TWO!EARS architecture relies on (traditional) bottom-up signal processing and top-down cognitive processes. Based on the signal captured by two microphones, a binaural auditory frontend [9] extracts auditory features. Among them, one can cite rate maps, interaural differences, interaural coherence, onsets/offsets, etc. These bottom-up processing stages are implemented as a collection of processor modules. The subsequent cues constitute the inputs of a collection of knowledge sources (KS), formulating hypotheses at different level of abstraction such as identity knowledge sources – dedicated to the identification of sound classes –, or localization knowledge sources – estimating the source event geometrical position w.r.t. the binaural sensor – etc. In this architecture, the proposed MFI system will constitute an additional KS of the architecture since it will make hypotheses about objects/events to focus on. Importantly, all these KS communicate which each others by reading and writing data on a globally-accessible structure (called *blackboard* [2,6]).

2.2 Inputs of the MFI

The MFI model relies on the outputs of three kinds of KS: (i) a Localization KS, (ii) an Auditory Identification KS, and (iii) a Visual Identification KS. The output of the KS are simulated so as to mimick the behavior of the TWO!EARS one. Thus, the Auditory Identification KS will output a vector whose dimension is equal to the number of auditory classification experts available:

$$\mathbf{P}^a[n] = (p_1^a[n], \ldots, p_{N_a}^a[n])^T, \tag{1}$$

In an identical fashion, the Visual Identification KS will output a vector whose dimension is equal to the number of visual classification experts available:

$$\mathbf{P}^v[n] = (p_1^v[n], \ldots, p_{N_v}^v[n])^T, \tag{2}$$

In equations (1) and (2), $p_i^{a/v}$ represents the probability of the current audio/video frame to belong the the i^th audio/visual category. Finally, the Localization KS outputs a vector whose components are the probabilities of a sound source to be located to a certain angle within a $360°$-wide range. These three outputs are then collected inside a time varying vector $\mathbf{V}[n]$ defined as

$$\mathbf{V}[n] = (\mathbf{P}[n]^T, \theta[n])^T, \text{ with } \mathbf{P}[n] = (\mathbf{P}^a[n]^T, \mathbf{P}^v[n]^T)^T. \tag{3}$$

In all the following, the vector $\mathbf{P}[n]$ will be referred as the n^{th} *audio-visual frame* available to the MFI.

2.3 Audio-visual category $\mathscr{C}$

We define the audio-visual category $\mathscr{C}$ of an audio-visual frame as being the concatenation of the ground true classification of the current audio and visual data. Let's consider, for instance, $N_a = 3$ auditory identification KS modeling the sound categories: $\mathscr{C}_1^a = \{\texttt{speech}\}$, $\mathscr{C}_2^a = \{\texttt{knock}\}$ and $\mathscr{C}_3^a = \{\texttt{alert}\}$. In the same vein, let's imagine $N_v = 2$ visual identification KS modeling the visual categories $\mathscr{C}_1^v = \{door\}$ and $\mathscr{C}_2^v = \{face\}$. Then, if a person is speaking in front of the robot, the audio-visual category $\widehat{\mathscr{C}}$ output by the MFI is expected to match the real category $\mathscr{C} = \{\mathscr{C}_1^a, \mathscr{C}_2^v\}$. More, the MFI output is expected to match the real category even if (i) the audio and/or visual KS produce wrong classification results, or (ii) audio or visual data are missing.

2.4 Internal structure

Fig. 1 exhibits the MFI internal structure. Two main points have to be emphasized:

- the categorization of the audio-visual frame is computed by a Multimodal Self-Organizing Map (M-SOM);
- this M-SOM is directly connected to a module responsible for triggering motor orders so as to confirm the missing data inference (*active* data inference).

Each subsystem will be formalized in the following subsections. Again, the reader has to keep in mind that the MFI performs active data inference *on the basis on the classifier outputs*, and not on the audio or visual cues extracted from the raw signals. Thus, the MFI can be understood as a classifier fusion system that estimates the audio-visual category $\widehat{\mathscr{C}}[n]$ of a perceived object. The exploitation of this estimated audio-visual category is out of the scope of the paper.

2.5 M-SOM

The M-SOM is directly based on a traditional SOM which is an artificial neural network which provides a discretized representation of an input space in an unsupervised way [8]. In other words, a classical SOM provides a way to visualize high-dimensional data through a low-dimension projection preserving the input data topology. In practice, a two-dimensional map is often used to represent the input data through a 2D arrangement of nodes r_{ij}, each of them being associated with (i) a weight vector $\mathbf{w}_{ij}$ of the same dimension as the input data vectors, (ii) a position (i, j) in the map space. Learning such a map traditionally requires two steps:

1. detection of the Best Matching Unit (BMU), i.e. the node $r_{\text{BMU}}[t]$ whose associated weight vector is the most similar to the input vector being $\mathbf{P}[n]$ learned at iteration t, i.e.

$$r_{\text{BMU}}[t] = r_{IJ}[t], \text{ with } (I,J) = \arg\min_{(i,j)}\{\|\mathbf{P}[n] - \mathbf{w}_{ij}[t]\|\} \tag{4}$$

   where $\|.\|$ represents the Euclidean distance, and $(i, j) \in [1,\ldots,N_a] \times [1,\ldots,N_v]$;

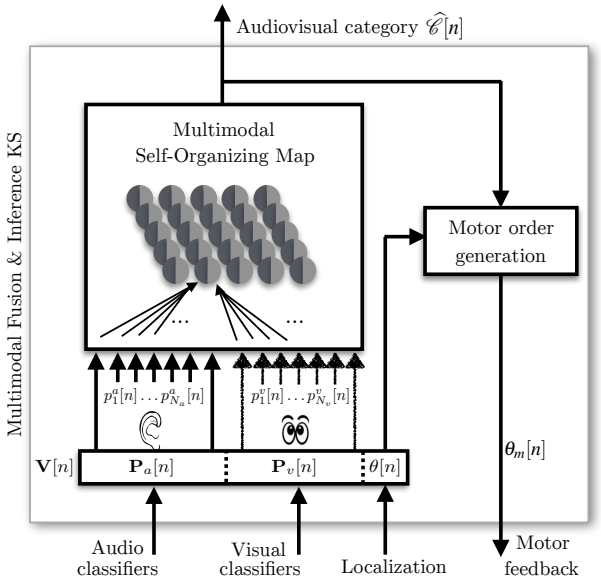

Figure 1: Global architecture of the Multimodal Fusion and Inference model.

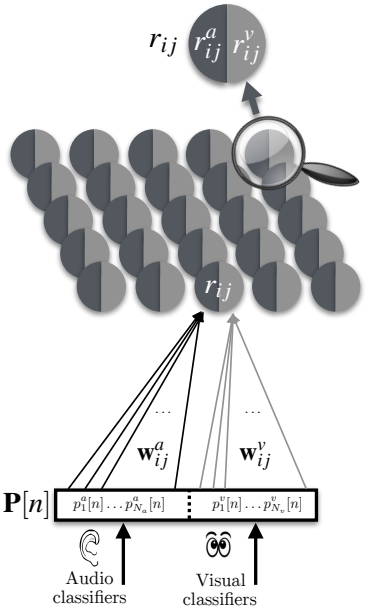

Figure 2: Structure of the M-SOM. Each node r_{ij} carries two weight vectors $\mathbf{w}_{ij}^a$ and $\mathbf{w}_{ij}^v$.

2. weight adaptation, including a neighborhood function h_{ij} which allows the propagation of input topology around the BMU, i.e.

$$\mathbf{w}_{ij}[t+1] = \mathbf{w}_{ij}[t] + \alpha[t]\ h_{ij}[t]\ \|\mathbf{P}[n] - \mathbf{w}_{ij}[t]\|, \tag{5}$$

with h being defined as a Gaussian neighborhood function, with

$$h_{ij}[t] = \exp\left(-\frac{\|r_{\text{BMU}}[t] - r_{ij}\|^2}{2\sigma[t]^2}\right). \tag{6}$$

Once the learning phase is over, the SOM can be used for clusterization. Given an input vector $\mathbf{P}[n]$, the BMU is first localized in the map by using Eq. (4). Its corresponding weight vector $\mathbf{w}_{\text{BMU}} = \mathbf{w}_{IJ}$ then carries information about both audio and visual modalities since $\mathbf{w}_{\text{BMU}} = (\mathbf{w}_{\text{BMU}}^a{}^T, \mathbf{w}_{\text{BMU}}^v{}^T)^T$, with $\mathbf{w}_{\text{BMU}}^a = (w_1^a, \dots, w_{N_a}^a)^T$ and $\mathbf{w}_{\text{BMU}}^v = (w_1^v, \dots, w_{N_v}^v)^T$. Then, the audio-visual category $\widehat{\mathscr{C}}[n]$ of the input $\mathbf{P}[n]$ is estimated along

$$\begin{aligned} \widehat{\mathscr{C}}[n] &= (\widehat{\mathscr{C}}_A^a[n], \widehat{\mathscr{C}}_V^v[n]), \text{ with} \\ A &= \arg\max_k w_k^a \text{ and } V = \arg\max_l w_l^v. \end{aligned} \tag{7}$$

While the SOM has proven to be a very efficient way to compact and represent high-dimensional data, this tool is not able to cope with missing data. Unlike the traditional SOM, the proposed Multimodal Self-Organizing Map (M-SOM) uses *two* weight vectors $\mathbf{w}_{ij}^a$ and $\mathbf{w}_{ij}^v$ per node r_{ij}. Each weight vector is thus dedicated to a given modality. The two maps are then fused together under the constraint $r_{ij}^a = r_{ij}^v = r_{ij}$.

2.5.1 If all modalities are available

In such a case, the system will be able to learn the relationship between the audio and visual components, but also to possibly correct the wrong classification results from the audio or visual KS.

Learning step An *audio-visual* BMU r^{av}_{BMU} is defined along

$$\begin{aligned} r^{av}_{\text{BMU}} &= r_{IJ}[t], \text{ where} \\ (I,J) &= \arg\min_{i,j} \left(\|\mathbf{P}^a[n] - \mathbf{w}^a_{ij}\| \|\mathbf{P}^v[n] - \mathbf{w}^v_{ij}\| \right), \end{aligned} \tag{8}$$

Once this multimodal BMU is found, the rest of the learning algorithm remains the same and follows Eq. (5) and (6).

Category estimation Audiovisual categories can still be estimated thanks to Eq. (7), but by using the components w^a_i and w^v_i of the weight vector of the audio-visual BMU $\mathbf{w}^{av}_{\text{BMU}}$ found with Eq. (8). This will result in an estimated category $\widehat{\mathscr{C}}^{(\text{all})}[n]$ (where $^{(\text{all})}$ indicates that both audio and visual data was available). Importantly, even if some classification errors occur the M-SOM should be able to *correct* those errors. This will be illustrated in §3.

2.5.2 If one modality is missing

In such a case, there is no learning phase. Instead, the current state of the M-SOM is used to *infer* the missing data. Let's consider, as an example, the case where visual data is not available:

1. audition alone is used to derive the audio BMU r^a_{BMU} in the audio map, whose associate weight $\mathbf{w}^a_{\text{BMU}}$ can be used to decide the audio category $\widehat{\mathscr{C}}^a_A[n]$, with $A = \arg\max_k w^a_k$;
2. the visual BMU is directly derived from the audio one with $r^v_{\text{BMU}} = r^a_{\text{BMU}}$. *This is the step where the link between audio and visual data built during the learning step is exploited*;
3. then, the weight $\mathbf{w}^v_{\text{BMU}}$ associated with the visual BMU r^v_{BMU} can now be used to decide the visual category $\widehat{\mathscr{C}}^v_V[n]$, with $V = \arg\max_l w^v_l$.

At the end, the system is then able to provide an estimated audio-visual category $\widehat{\mathscr{C}}^{(\text{miss})}[n] = (\widehat{\mathscr{C}}^a_A[n], \widehat{\mathscr{C}}^v_V[n])$ (where $^{(\text{miss})}$ indicates that there was a missing modality), even if no visual data is available. A reciprocal approach can be used when audio data is missing.

2.6 Motor order generation

As outlined above, no learning phase occur if a modality is missing while inference is made to estimate the category of current object. Benefiting from having a mobile robot, a motor action could allow to *actively* catch this missing data. In this paper, the active behavior will be

restricted to head movements only, while not being conceptually limited to. Let's consider the Kronecker delta $\delta_{ij}^{(k)}[n]$ defined along

$$\delta_{ij}^{(k)}[n] = \begin{cases} 1 \text{ if } \widehat{\mathscr{C}}^{(k)}[n] = (\mathscr{C}_i^a[n], \mathscr{C}_j^v[n]), \\ 0 \text{ else,} \end{cases} \tag{9}$$

where $k = \{\text{all}, \text{miss}\}$ denoting if the the category has been obtained without or with missing data. We can then define the *inference ratio* $q_{ij}[n]$ of the audio-visual category $(\mathscr{C}_i^a, \mathscr{C}_j^v)$ with

$$q_{ij}[n] = \frac{\sum_{k=1}^{n} \delta_{ij}^{(\text{miss})}[k-1]\, \delta_{ij}^{(\text{all})}[k]}{\sum_{k=1}^{n} \delta_{ij}^{(\text{miss})}[k]}. \tag{10}$$

q_{ij} captures the ratio between the number of confirmed inferences and the number of time an inference has been made through the M-SOM for a given audio-visual category. On this basis, a head motor command $\theta_m[n]$ is generated according to

$$\theta_m[n] = \begin{cases} \theta[n] \text{ if } K_{\text{head}} \leq q_{ij}[n] < 1, \\ \theta_m[n-1] \text{ else.} \end{cases} \tag{11}$$

The angle $\theta[n]$, given by the Localization KS, is exploited to turn the head towards the estimated sound position at time n. $K_{\text{head}} \in [0,1]$ represents in Eq. (11) a threshold allowing to tune the active behavior. A low threshold will make the system quickly trust in the inferences (and thus will inhibit the head movements), while a high threshold will trigger a lot of head movements so as to verify often if the inferred audio-visual category is correct.

3 Simulations and Results

The Multimodal Fusion & Inference will be evaluated along a simulation of 13 audio classifiers and 9 visual classifiers for a total of 14 possible audiovisual pairs (since some audio labels can be associated to two or more visual labels). This will enable the statistical evaluation of the MFI performances, in terms of category estimation, data inference/correction, but also in terms of head movement modulation

3.1 M-SOM convergence and classification rate

A first evaluation of the M-SOM performances, in terms of frame categorization, consists in comparing its category estimation output $\widehat{\mathscr{C}}$ to a decision taken directly at the outputs of the audio and visual classifiers $\overline{\mathscr{C}}$, with $\overline{\mathscr{C}} = (\overline{\mathscr{C}}_{\tilde{A}}^a, \overline{\mathscr{C}}_{\tilde{V}}^v)$ where $\tilde{A} = \arg\max_k p_k^a$ and $\tilde{V} = \arg\max_l p_l^v$. Fig. 3 shows the results of a 5000 steps simulation of 125 audiovisual objects. The black and gray curves exhibit the *mean good frame categorization rate* for the proposed M-SOM and the decision at the KS outputs, respectively. The M-SOM systematically outperforms the naive decision on the classifiers *a priori* probabilities by 7.4% during the first 500 steps, to 15.8% on the last 500 steps. This demonstrates the efficiency of the audio-visual fusion performed by the proposed M-SOM. Fig. 3 also represents the generated objects as rectangular boxes. The

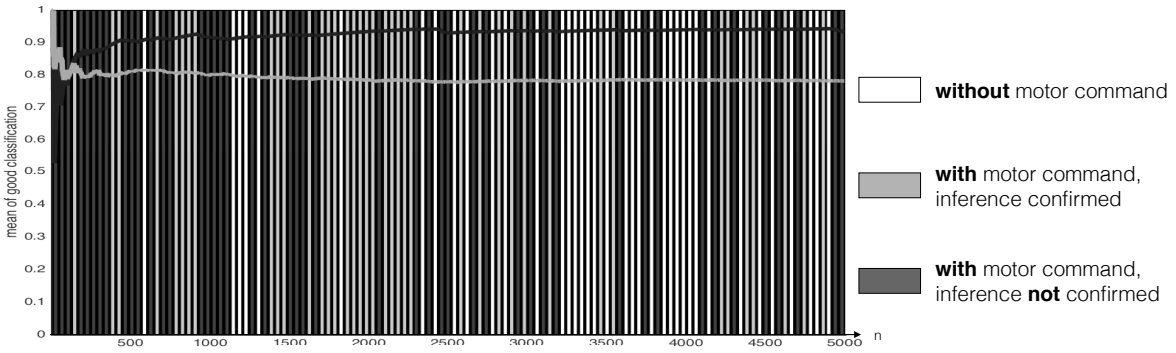

Figure 3: 5000 steps simulation of 125 audiovisual objects. Mean good frame categorization rate of the MFI (black), and at the KS output (grey)

color of these boxes indicates whether a motor command has been triggered to confirm the inference by the M-SOM (gray and dark gray boxes) or not (white boxes). It appears that head movements are mainly triggered at the beginning of the simulation (few white boxes), while being less necessary later (more white boxes). This dynamic follows directly the progressive learning of the MFI together with the growing confidence the system has in its inferences.

3.2 Head movement modulation

In order to quantitatively observe how the head movements can be modulated, the MFI system is compared with a naive robot that would turn its head every time a new object appears in the environment. Fig. 4 shows the result of this comparison by plotting the number of time the MFI has produced a head movement towards a source (red line), versus the naive robot (black line). During the whole experiment, 125 objects have been observed by the simulated robot, thus triggering 125 head movements by the naive robot, against only 89 for the proposed MFI system.

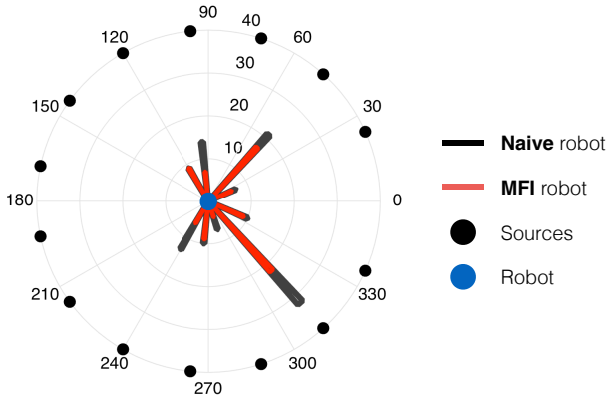

Figure 4: Number of head movements triggered by the MFI (red line) and by a naive robot (black line)

4 Conclusion

This paper has presented an original Multimodal Fusion and Inference system, based on an unsupervised real-time learning algorithm that works without any *a priori* knowledge about the environment. It enables an exploratory robot (i) to infer missing information on the sole basis of observation, and (ii) to drive its attention by inhibiting some spontaneous head movements. Ongoing work is now focused on the integration of the system into a real robot and on experiments in realistic conditions.

References

[1] 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*, pages 1766–1773, 2010.

[2] G. Brown, R. Decorsière, D. Kolossa, N. Ma, T. May, C. Schymura, and I. Trowitzsch. The TWO!EARS Software Architecture. Technical Report Deliverable 3.1, May 2014.

[3] D. Calisi, A. Farinelli, L. Locci, and D. Nardi. Multi-objective Exploration and Search for Autonomous Rescue Robots. *Journal of Field Robotics*, 24(8/9):763–777, 2007.

[4] Z. Chen. Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, 74(5):784–802, 2012.

[5] B. Cohen-Lhyver, S. Argentieri, and B. Gas. Modulating the Auditory Turn-to Reflex on the Basis of Multimodal Feedback Loops: the Dynamic Weighting Model. In *Robotics and Biomimetics (ROBIO), 2015 IEEE International Conference on*, page in press, 2015.

[6] L. Erman, F. Hayes-Roth, V. Lesser, and D. Reddy. The HEARSAY-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty. *Blackboard Systems*, pages 31–86, 1988.

[7] 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*, 6:56–72, 2014.

[8] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[9] T. May, R. Decorsière, C. Kim, and A. Kohlrausch. The auditory front-end framework user manuel. Technical Report Supplement to TWO!EARS Deliverable 2.2, November 2014.

[10] 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.

[11] Two!Ears Team. Two!ears auditory model, 2015. https://zenodo.org/record/35492.

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