# Developmental Learning of Audio-Visual Integration From Facial Gestures Of a Social Robot

Oriane Dermy, Sofiane Boucenna, Alexandre Pitti, Arnaud Blanchard

# Developmental Learning of Multi-modal integration for interactive and social robots

Oriane Dermy, Sofiane Boucenna, Alexandre Pitti, Philippe Gaussier
ETIS Laboratory, UMR CNRS 8051, Université de Cergy-Pontoise, ENSEA, France.
Email: firstname.lastname@u-cergy.fr

***Abstract*—This paper presents a neural network architecture allowing a robot to recognize patterns in a multimodal way (aurally and visually) following a developmental approach. Thanks to this recognition, the robot can interact socially with a partner (by means of facial expressions), as an infant would. To allow that recognition, a learning is performed by exploring several perception-action associations with a auto-supervised reinforcement learning. This network allows an imitation behavior with a partner using perception and synchronization to link the different modalities. We show that the robot can use redundant information from the different modalities to make decisions. For instance, if a sensory signal is not efficient enough, the other one compensates for it. We will show that our architecture is more robust than the unimodal ones and works in unfavourable environments where many people move and talk all the time.**

## I. Introduction

Our main goal is to understand how social skills can be acquired by a robot in an autonomous fashion, as it is for infants during development. Based on observations and findings on the developmental stages of infants, we put forward the important aspects of two cognitive mechanisms. Firstly, multimodal integration allows the robot to learn online and in a auto-supervised manner because the amount of expressions is known in advance. Secondly, enaction[1] allows it to be responsive to physical and social signals of other agents due to the fact that it is able to emit those signals itself.

On the one hand, it is known that audio-visual integration is important to recognize and locate faces [2]. Other works [3] use this same idea. On the other hand, speeches and facial expressions are important signals to mimic others, and mimicking others is an important social skill that allows, for example, to learn by imitation. These two features are important for the start of verbal and non-verbal communications. Combining them, we can build a cognitive architecture based on multi-modal integration to learn social interactions.

We propose to use a robot's head that we named Tino that combines three modalities : (1) hearing through microphones (in plastic ears), (2) sight through cameras controling eye gaze and eyelid, and (3) proprioception through a mechanic mouth and eyebrows which generate facial expressions.
Other works allow robots to have these modalities. For example, Bonnal [4] and Garcia [5] use binaural information to locate specific objects and add visual features to locate goals [6]. Regarding sound localization, Eui-Hyun Kim's works [7] concentrate on optimizing computations. Burger's study [8] is complementary to our work : he presents a companion robot that uses auditive and visual clues for multiple purposes, including speech recognition. Another interesting work using neural network is Droniou's [9], where he uses deep unsupervised learning for multimodal perception. However, their algorithm is based on a non-biological approach with important computations and requiring a semantic lexicon to work (and only allowing to recognize visual letters).
A lot of other multimodal works, based on labelled signals, exist with the aim of optimizing algorithms and performances. Instead, the aim of this study is to create a simpler neural network that may resemble the infants multimodal integration development. The redondancy across the modalities will allow our network to be efficient even in the real world.

This paper presents a neural controller that attempts to follow a plausible developmental approach to learn the redundant correlations between the different sensorimotors signals in order to have a structured multimodal interaction. Our neural network integrates both modalities through the learning of bimodal neurons, which is in agreement with the fact that deaf persons have the same activated multimodal areas than non-deaf persons. [10]
We show how audiovisual primitives are learnt in a self-organized manner with a conditional signal driven by the robot's behavior. An interesting point of this experiment is that the robot recognizes vowels even in an inadequate environment. The quality of our robot's recognition proves that multimodal integration outperforms single-modality learning. We then discuss the relevance of our developmental neural architecture to have a multimodal and social interaction.

## II. Hardware : Tino's robotic head

In our experiments, we used our hydraulic robot called Tino. This experiment only requires its head, which allows

[1]Enactivism argues that cognition arises through a dynamic interaction between an acting organism and its environment. It claims that our environment is one which we selectively create through our capacities to interact with the world.[1]

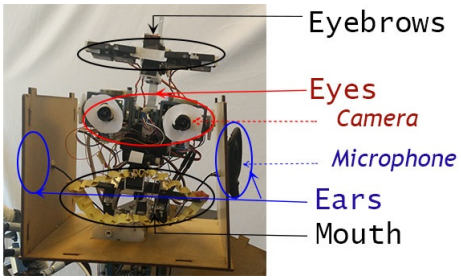


Fig. 1. Tino's head. Its eyes, composed of two pan/tilt cameras, its mouth, allowing it to display facial expressions (9 degrees of freedom) and one of its ears, created by a 3D printer and containing microphones.

to modelize three modalities, see Fig. 1.

The first modality is *hearing* through microphones implemented into plastic ears, themselves created with a 3D printer. These ears are human like, allowing the robot to have the same auditory detection capabilities as humans. It is important to underline that wave reflections are not taken into account, as shown in a previous work using the same ears [11].

The second modality is the *sight* through analogic cameras with several motors to control the eye gaze and the eyelid. These *eyes* perform fast pan and tilt saccadic eye movements. To do the visual processing, we chose an analogic camera transmitting images which have a $640 \times 480$ resolution.

The third modality consists of *proprioception* through a mechanic mouth, which generates facial expressions and two motorised eyebrows, allowing the robot to mimic human face. Nine servomotors, corresponding to the degrees of freedom, control their movements.

The mouth is composed of five servomotors. Two of them allow the stretching of the mouth, for instance allowing to pronounce the vowel $/i/$. The three others allow the opening of the mouth : one in the center and two at the ends, that allow, for example, the pronounciation of $/a/$ and $/o/$ vowels, to smile or sulk. Each eyebrow is connected to two servomotors that allow the robot to frown, be suprised, and so on.

## III. Experimental Setup

### A. Explanation of choices

This experimental setup is based on a previous one [12], which used only one modality (vision) associated with the facial expression. Sound treatments and multimodal learning have then been added to it. It is also inspired by a previous work [11], which was based on audio-visual integration only. Figure 2 presents stages of the learning period and the recognition period.

The robot will learn to make some facial expressions that look like a human's when he pronounces the $/a/$, $/o/$ and $/i/$ vowels, as well as a neutral expression. We chose these three vowels purposefully, in order to link our research to the Kuhl and Meltzoff ones [13]. These anterious researches show that 4-month old newborns are already sensitive to

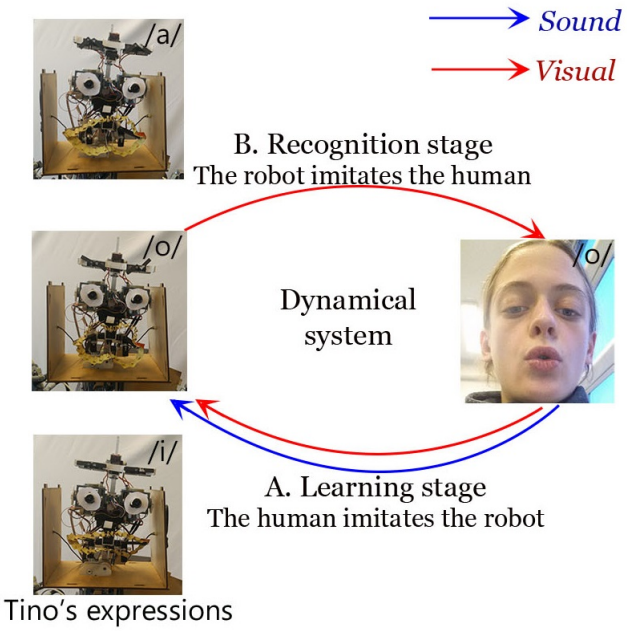


Fig. 2. Experiment stages. A. The learning stage where Tino babbles and the human partner mimics. B. The recognition stage where Tino mimics.

phonetic information, even for unknown languages, and are able to coo (where infants produce sounds that corresponds to vowels). Moreover, newborns are also able, only few hours after birth, to imitate more quickly a mouth opening if it corresponds to the $/a/$ or $/i/$ vowel [14]. In order to complexify the task, we added another vowel.

### B. Experimental method

We have performed imitation game experiments where, during the learning stage (Fig. 2.A), the human partner imitates the robot's expression, which corresponded to the vowel $/a/$, $/o/$ or $/i/$, adding audio information to the visual one. This imitation is done in a exaggerated way, to help the robot learn, as it would be for a mother with her child. The mother performs the child-directed speech named motherese [15]. The sound is pronounced continuously for three seconds. To allow the participants to catch their breath, the robot takes a *neutral expression* between two successive expressions. Finally, the robot learns to associate its own head movements with the partner's signals (audio and visual).

During the recognition stage, the partner pronounces a vowel with exaggerated audio and visual signals. The previous learning allows the robot to execute the facial movements that match the input signal. Thus, the robot can mimic the human expression.

This experiment has first been tested and validated in the laboratory with researchers. In order to validate it with a more diverse public, we have collected data during a robot

exhibition named "Futur en Seine"[2] from both children and adults volonteers. In order to do that, the robot presented some facial expressions and we took audio-visual data from the 24 participants, half of whom were younger than 15. The recognition stage was done as follows :
First, the robot took an expression corresponding to $/a/$, $/o/$ or $/i/$ vowels. Then, the partner imitated visually and orally the robot's facial expression.
We saved 200 images and recorded 1024*20 bytes of sounds per person to do our statistical study offline. During this learning stage, we labelled each event (expression change) and sorted the database according to each vowel's label.
The data are used for the learning stage in a random way : an image is first drawn. Its label ($/a/$, $/o/$, $/i/$, or $neutral$) allowed to take a snip of sonor signals from the file corresponding to the same label.
We limited the number of facial expressions to four ($/a/$,$/o/$, $/i/$, or *neutral*), although we could have handled all vowels.

## IV. Methods

Fig.3 represents an overview of our experimental paradigm.
The robot learns to associate its internal state (its facial expression) with the audio and visual signals from its human partner.
In this section, we show how the audio and visual signal processing are performed. Then, we present the two learning mecanisms : (1) categorization mechanism : the input signal is clusterized, (2) recognition mechanism : the clusters from categorization mechanism are associated to vowels.

### A. Sound processing

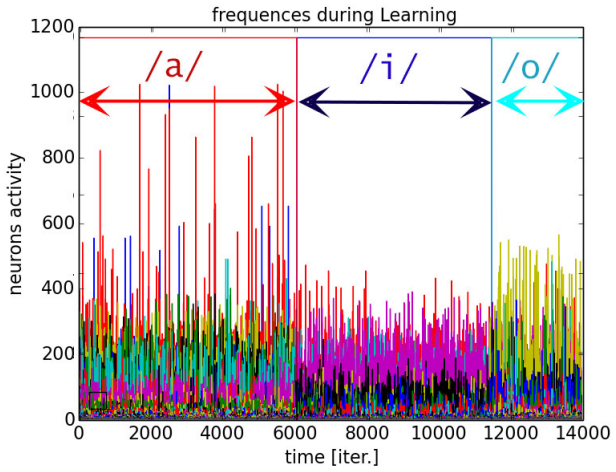


Fig. 4. Frequence variations during the pronunciation of $/a//o/$ and $/i/$ vowels

To begin, we present the main speech process researches. These studies deals with specific information such as tone and spectral envelope (with the spectogram)[16]. To retrieve these information, some items are often used : gammatone filters, the Fast Fourier Transform (FFT[17]) on which we often apply the Mel Frequency Cepstral Coefficient (MFCC) that considers only the most important frequences[17][18].
In this study, we want to recognize vowels in a developmental way, with as simple a neural network as possible, using only the FFT. This FFT replaces the natural filters that exist in human ears, done by the cochlea. This FFT is compelling because the signal amplitude cannot categorize sounds.
First of all, the audio signal is received from two microphones inserted in the ears. Thus, there are two channels, with *interleaved* information. The sample rate is 44100 Hz. 1024 samples of sound intensity are taken per iteration.
Then, the FFT is computed on the collected samples and its result is 513 coefficient of frequencies.

Fig. 4 presents the result of this treatment when someone is pronouncing the vowel $/a/$ (0 to 6000 iterations), $/i/$ (6000 to 12000 iterations) and $/o/$ (1200 to 14000 iterations). We can see that the coefficient of frequencies changes with the pronunciated vowel. This first result shows that the robot can categorize sounds as the result of a low-level process.

### B. Vision processing

The visual processing is performed on small images($320 \times 240$) to accelerate the computation time. Points of interest of these images are spotted to categorize them. We only consider the grayscales of these images, so that we can treat them with a Difference Of Gaussian filter. The size of the mask of this filter is $15 \times 15$, with $\sigma_1 = 6$ (positive coefficient of the first Gaussian) and $\sigma_2 = 3$ (negative coefficient of the second Gaussian). By doing so, important details are conserved (edges and shapes).
After this treatment, we extract the 10 most important points of interest. Then, a thumbnail centralized on each of these points is taken.
Since this exact processing has already be done in previous researches[19][20], we won't present any test here.

### C. Model of the sensori-motor neural network

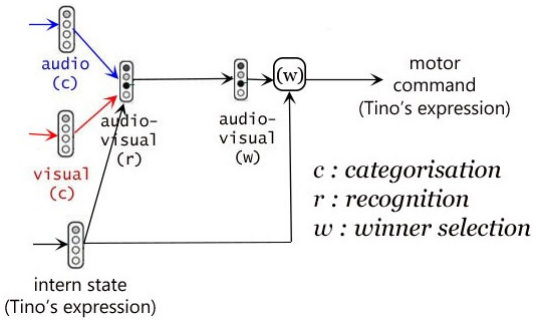


Fig. 5. Neural model of audio-visual integration

[2]This event took place inside "Visages du monde" at Cergy Le Haut

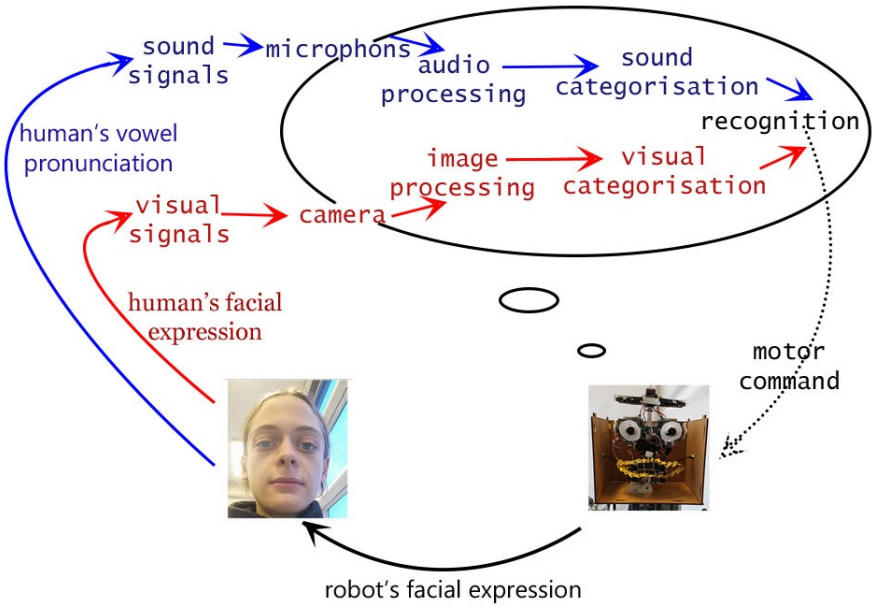


Fig. 3. Overall architecture

This architecture follows the idea that perception is the coupling between sensation and action. The action can be triggered by internal state as well as the sensation (sound and visual). This idea is an important principle in AI, allowing the robot to react in a more human way. This is the PerAc (Perception-action) architecture [21], created in our laboratory.
This idea is an agreement with the fact that babies are more sensitive to perception that matches their own motor movements. Indeed, studies as [14] show that newborns mimic movements more easily if those match audio-visual information.

The integration of modalities can be done in several ways. Due to space conditions, only one of the possible models is explained here.
All of these models use two learning algorithms presented in next sections. After being treated by these algorithms, a winner take all algorithm filters neuronal activities to maintain only the most active one. Then, an action that is linked to this recognition is chosen .

Fig. 5 presents one of these models.
Once the signals are pretreated, the categorization (c) is performed modality by modality, in a separate way thanks to a neural network name Selective Adaptative Winner (SAW IV-C).
Then comes the recognition process, annoted (r), that corresponds to the labelization of the previous categorizations. During this recognition, the robot learned to assimilate its internal states (its facial expressions) with categories activated by the SAW. The recognition uses two modalities at the same time. The Least Mean Square (LMS IV-C) mecanism allows this recognition.
After the learning stage, the robot will select the most recognized label to trigger the associated expression. As a result, it can mimic the human partner.

**Selective Adaptative Winner (SAW)**

The SAW is an ART-like neural network [22] that allows to create categories that are related to inputs. When a new input is presented, the neural network selects the category that is the best match. If it is close enough to the input, the robot adapts this category to be more similar to the input. If not, the robot creates a new category that perfectly matches the input. The SAW has the following principle :
If $e_j$ are the input values and if the importance of this input for the output neurons $k$ is given by the weight $w_{kj}$, then the activity of this $k^{th}$ neuron of the SAW output equals : $A_k = 1 - \frac{1}{N} \times \sum_j (|w_{kj} - e_j|)$.
The weight $w_{kj}$ is updated with : $\Delta(w_{kj}) = \epsilon^{SAW}(e_j - w_{kj})$.
Here, $\epsilon^{SAW}$ is the learning rate :

$$\epsilon^{SAW} = \begin{cases} 1 & \text{if } A_k > vigilence^3 \\ 0 & \text{else} \end{cases}$$

**Least Mean Square (LMS)**

This learning algorithm called Least Mean Square (LMS) links the robot internal state to categories that are activated by the SAW algorithm. The LMS follows this principle :
If $e_k$ are the inputs coming from the SAW, $W_{ik}$ the weights between the $k^{th}$ input of the SAW and the $i^{th}$ output of the LMS, then the activity of the $i^{th}$ neuron of the LMS output equals : $O_i = \sum_k (W_{ik} \times e_k)$
The weight $w_{ik}$ is updated with :
$\Delta(w_{ik}) = \epsilon^{LMS}(actual\_expression - O_i) \times e_k$.

Here, $\epsilon^{LMS}$ is the learning rate.
Thus, this learning algorithm adapts weights $W_{ik}$ in order to match the *actual_expression* variable (that is the internal state of the robot) to the computed expression $O_i$.

### D. Results

First of all, our experiment took place in our laboratory (University of Cergy Pontoise / ETIS).

The database is composed of 200x24 images that corresponds to the 20 changes of vowels produced by Tino during its babbling, when 10 pictures per change were taken for each of the 24 participants. This same database is used for the learning stage and the recognition one.
Figure 6 attests directly that the multimodal learning allows more robustess than the unimodal ones.

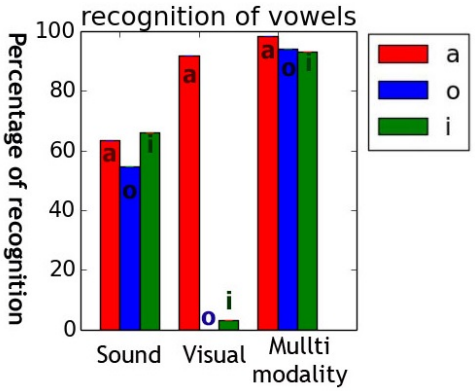


Fig. 6. Correct recognition percentages for all vowels and all learning models

This figure shows that the visual envionnement wasn't adequate, with a weak percentages of recognition for all vowels (22% in average). Even with that bad visual input, the multimodal learning allows good performance, with an average recognition percentage above 90%. This percentage is even higher than the audio-modality one, that has a recognition average of 60%. It is interesting that audio-modality seems to be sufficient to recognize expression even if the environment was very noisy.
We will now go into details.

All the next figures are about neuronal activity. The following explanation is necessary to undersand these figures.
Each vowel is represented by a control (centroid of all the same labelled signals). We use this control to determine what each new signals should be labelled. You can imagine that each label (vowel) is associated with a neuron, and that the neuronal activity of each label increases as the signal is close to its control.
To help you to understand, the next two figures are focused on the neuronal activities during the pronunciation of the $/o/$ vowel.

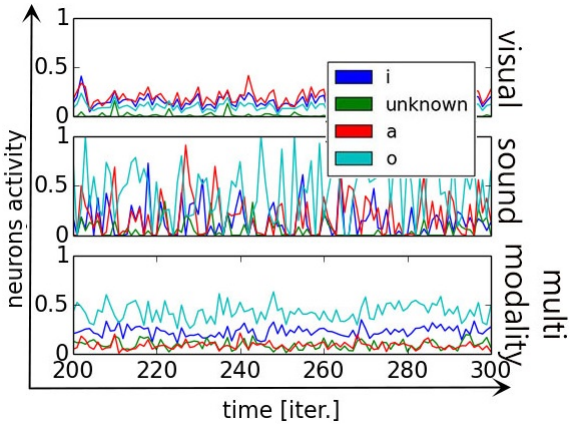


Fig. 7. Activity of neurons specialized in the recognition of vowel $/a/$ in red, $/o/$ in light blue and $/i/$ in darkblue during the pronunciation of the $/o/$ vowel

Fig. 7 shows the neuron's activity in function of iterations (time) during the pronunciation of the $/o/$ vowel.
The most activated neuron corresponds to the vowel that the robot recognizes.
Three activities are presented : the vowel $/a/$ in red, $/i/$ in darkblue and $/o/$ (the pronounced one) in lightblue.
Unimodal recognitions are weak (as we expected), due to the environmental perturbations during the experiment. Indeed, for these two learnings, it recognizes each neurons at the same proportion (the variance inter-class is very weak). Moreover, for the visual modality, it is even the $/a/$ instead of the $/o/$ that is recognized.
The multimodal learning is more powerful since the recognized vowel is always correct (in this sample) and it offers a bigger inter-classe variance for the recognition than any unimodal learning.

Now, Fig. 8 presents the recognition rate using a histogram of neuronal activities during the pronunciation of vowel $/o/$.

This figure also proves previous observations, such as the weak visual recognition. Moreover, it specifies the bound values and average neuronal activities. The visual recognition is totally deficient, with the $/a/$ vowel recognized (instead of the $/o/$ vowel). Slightly better, the sound recognition recognizes the right vowel, but with an unstable recognition rate : the neuron's activity can go below 0.2 even though its maximum activity is equal to 1. Its average activity is equal to 0.55.
At the bottom, the multi-modal learning allows to recognize the pronounced vowel with more stability than any unimodal learning recognition. Indeed, there are less recognition variations, with an average equal 0.5.
As expected, the maximum rate of the multimodal recognition is weaker than that of the hearing recognition because of poor visual data that influenced the robot's

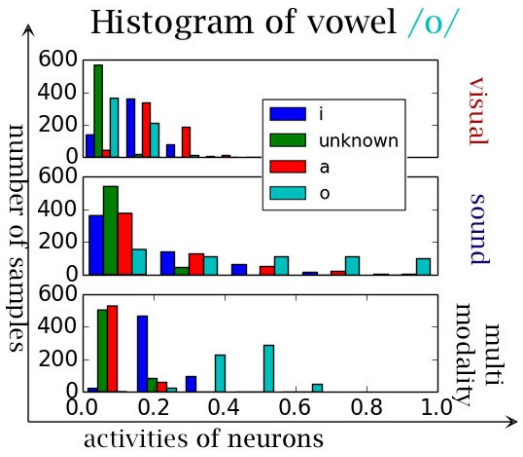


Fig. 8. Neurons' activity rate for all vowels and for all types of learning during the pronunciation of the vowel /o/

perception (0.65 vs 1.0). Indeed, the robot takes both modalities into account to recognize the facial expressions. However, it is obvious that the multimodal learning is more robust than the auditive one thanks to three things : (1) the minimum activity of the multimodal recognition is higher than that of the auditive one, (2) the intra-class variance of the multimodal recognition is lower than the auditive one, (3) the inter-class variance (the variance between all neurons' activity) of multimodal recognition is higher than that of the auditive recognition.

Fig.9 shows the activity rate of the recognition, but this time when the recognition works. This is done for each learning (visual only, auditive only or multimodal) and for each of the three vowels.

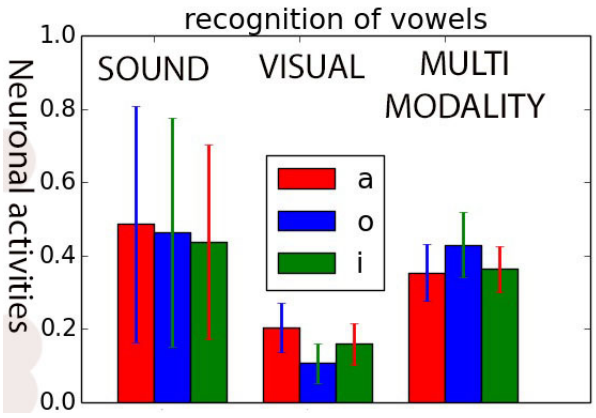


Fig. 9. Neuronal activity rates for each vowel.

Thinking that the visual recognition works better than the multimodality recognition would be a mistake. Indeed, even if the recognition rates are more important in this learning than in the multimodal one, remember that the other vowel recognition rates (that mismatch the pronounced one) are also more important in the visual learning. It is the intra-class variance that allows the robot to choose the expression. Thus, comparing the activity recognition rates between the different learnings is pointless.

A revelant information is the stability in the recognition : the multimodal recognition is more stable. Thus, inter-class variance is really weak. Moreover, when we saw the minimum rate, represented by the bottom extremity of the thin bars, in all these three learnings, we could see that the multimodal one is the most activated. Thus, our architecture is really more robust than an architecture based on one modality only.

To present this result in a simpler way, the figure10 presents the mean difference of recognition between the pronunced vowel and the average of other vowels. It is this information that the robot uses to select an action (face expression).

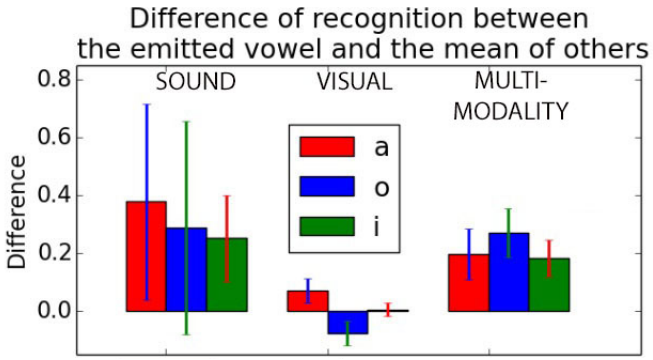


Fig. 10. Difference of recognition between the pronunced vowel and the average of the two others

Vowel /a/ is presented in red, /o/ in blue and /i/ in green. The thick bars stand for the mean difference. For example, if the vowel /a/ is produced, red bars are computed as follows : $\frac{1}{N}\sum(a.activity - \frac{o.activity+i.activity}{2})$.

The figure also shows the variance of each data (by means of thin lines).

The result highlights the fact that there are a lot of mistakes in the visual-only recognition, with the /o/ vowel unrecognized. Sound-only is correct most of the time, but its variance for the /o/ vowel shows that it can make mistakes too.

Finally, the bars corresponding to the multimodality learning show that the mean difference is always positive, with little variance.

These results prove that our architecture allows an important robustness since the experiment's environment was definitely unadapted (there were people talking, laughing, trying to get into Tino's sight area, and there was even a piano playing music near our robot). We have shown that bimodal neurons can recover information from a degraded sound signal with the use of visual signal, and the other way around. Thanks to the bimodal neurons in our system, we show that the

robot's performance is better than if it was using only one modality.

## V. Discussions

In this paper, we have presented a developmental scenario for a robot to learn social interaction with the help of partners. Our neural network allows to recognize multimodal stimuli patterns even if one of the two modalities (sound or vision) is missing. We show that multimodal neurons are robust enough to recognize the inputs better than to unimodal ones, even though our signal preprocessing was rather fast and easy. Moreover, the results show that our sensory-motor neural network allows the multimodal integration to work even in an unadapted environment.
We will then check the performances of our model with more complex audio-visual signals such as syllables or all vowels.

This study has many interesting perspectives and discussions. For instance, we could create a totally unsupervised model. Indeed, we expect the robot will be able to learn, in a self-organized manner and without indication, visual and sound repertoiries associated with the robot's proprioception. Thus, proprioception would be an input signal like the other ones, but directly linked tto the robot internal state.
In our neural network, the three modalities have been learnt, in the same mulimodal neurons. Many other models for multimodal integration exist, such as a model that integrates visual-proprioception and audio-proprioception modalities in a separate way.
Another interesting work would be to add gestual modality. Indeed, babies use gestual clues in addition to acoustic and facial clues to understand langages [23]. Our ongoing work raises the question of how to combine multimodal learning with the mechanisms of intentionality in a unified model, based on complementary works that perfomed a sensory-motor neural network to understand intentionality [24][25].
An interesting experiment will be to show the advantages of developmental robotics over probabilistic robotics by comparing our results with those observed in cognitive development researches. (1) It is known that infants begin to learn sound information in utero, where he is differentiating $/ba/$ and $/ga/$ sounds. A hypothesis suggests that newborns match these auditive information with visual ones [26]. We could test a corresponding model and see if it is efficient enough. (2) It is also well known that babies are already sensitive to the McGurk effect. We will test our robot's sensitivity to this effect. (3) In the same conditions than in [27], we will do a mismatch negativity or a contingency detection measure to see if our robot is sensitive to contingent information. To do that, we will have our robot focus on the objects that are the best match in term of auditory and visual modalities. (4) It would be interesting to check that the robot uses visual clues which are localized around the mouth, eyebrow, menton and jown because these clues are used by infants to understand langages [28], [29], [30], [31].


## Acknowledgements

We would like to thank the *Université de Cergy-Pontoise* for financing the research award "bourse d'excellence Master Recherche", Orange which has financed the "bourse Tremplin", the research grant "chaire d'excellence CNRS-UCP" and the ROBOTEX Equipex-Ile-de-France for the robotic equipment.


## References


[1] F. Varela, H. Maturana, and R. Uribe, "Autopoiesis: The organization of living systems, its characterization and a model," *Biosystems*, vol. 5, no. 4, pp. 187 – 196, 1974. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0303264774900318

[2] M. Vaillant-Molina, L. Newell, I. Castellanos, L. E. Bahrick, and R. Lickliter, "Intersensory redundancy impairs face perception in early development," in *Poster presented at the International Conference on Infant Studies, Kyoto, Japan*, 2006.

[3] G. Fanelli, J. Gall, and L. V. Gool, "Hough transform-based mouth localization for audio-visual speech recognition," in *British Machine Vision Conference*, September 2009.

[4] J. Bonnal, S. Argentieri, P. Dans, and J. Manhs, "Speaker localization and speech extraction with the ear sensor." in *IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS)*, St. Louis, MO, USA, October 2009, pp. 670 – 675.

[5] B. Garcia, M. Bernard, S. Argentieri, and B. Gas, "Sensorimotor learning of sound localization for an autonomous robot," in *Forum Acusticum*, Sept. 2014.

[6] A. Laflaquiere, S. Argentieri, B. Gas, and E. Castillo-Castenada, "Space dimension perception from the multimodal sensorimotor flow of a naive robotic agent," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 1520–1525.

[7] K. Ui-Hyun and O. Hiroshi G., "Improved binaural sound localization and tracking for unknown time-varying number of speakers," *Advanced Robotics*, vol. 27, no. 15, pp. 1161–1173, 2013. [Online]. Available: http://dx.doi.org/10.1080/01691864.2013.812177

[8] B. Burger, I. Ferrané, and F. Lerasle, "Multimodal interaction abilities for a robot companion," in *Computer Vision Systems*. Springer, 2008, pp. 549–558.

[9] A. Droniou, S. Ivaldi, and O. Sigaud, "Deep unsupervised network for multimodal perception, representation and classification," *Robotics and Autonomous Systems*, vol. 71, no. 0, pp. 83 – 98, 2015, emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889014002474

[10] A. Giraud, E. Truy, and R. Frackowiak, "Imaging plasticity in cochlear implant patients," *Audiol Neurootol*, vol. 6, pp. 381–393, 2001.

[11] A. Pitti, A. Blanchard, M. Cardinaux, and P. Gaussier, "Gain-field modulation mechanism in multimodal networks for spatial perception," in *Humanoid Robots (Humanoids), 2012 12th IEEE-RAS International Conference on*. IEEE, 2012, pp. 297–302.

[12] S. Boucenna, P. Gaussier, P. Andry, and L. Hafemeister, "Imitation as a communication tool for online facial expression learning and recognition," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 5323–5328.

[13] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: Vocal imitation and developmental change," *The Journal of the Acoustical Society of America*, vol. 100, pp. 2425–2438, 1996.

[14] A. Streri, M. Coulon, and B. Guella, "The foundations of social cognition: Studies on face/voice integration in newborn infants," *International Journal of Behavioral Development*, vol. 1-5, 2012.

[15] C. Kitamura, B. Guella, and J. Kim, "Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads," *PLoS ONE*, vol. 9, no. 10, p. e111467, 10 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0111467

[16] J. Sanchez-Riera, "Capacits audiovisuelles en robot humanode nao," Ph.D. dissertation, 2013, thse de doctorat dirige par Horaud, Radu Mathmatiques et Informatique Grenoble 2013. [Online]. Available: http://www.theses.fr/2013GRENM009

[17] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Beyond the query-by-example paradigm: New query interfaces for music information retrieval," in *In Proc. Int. Computer Music Conference*, 2002, pp. 177–183.

[18] P. D., "Content-based methods for the management of digital music," *International Conference on Acoustics, Speech, and Signal Processing*, vol. IV, pp. 2437–2440, 2000.

[19] S. Boucenna, S. Anzalone, E. Tilmont, D. Cohen, and M. Chetouani, "Learning of social signatures through imitation game between a robot and a human partner," *Autonomous Mental Development, IEEE Transactions on*, vol. 6, no. 3, pp. 213–225, 2014.

[20] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *International Journal of Social Robotics*, 2015, to appear.

[21] P. Gaussier and S. Zrehen, "Perac: A neural architecture to control artificial animals," *Robotics and Autonomous Systems*, vol. 16, no. 24, pp. 291 – 320, 1995, moving the Frontiers between Robotics and Biology. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0921889095000526

[22] R. P. Lippmann, "An introduction to computing with neural nets," *ASSP Magazine, IEEE*, vol. 4, no. 2, pp. 4–22, 1987.

[23] D. McNeil, "hand and mind : What gestures reveal about thought," *University of hicago Press*, 2005.

[24] D. Lewkowicz, Y. Delevoye-Turrell, D. Bailly, P. Andry, and P. Gaussier, "Reading motor intention through mental imagery," *Adaptive Behavior*, p. 1059712313501347, 2013.

[25] L. Cohen, W. Abbassi, M. Chetouani, and S. Boucenna, "Intention inference learning through the interaction with a caregiver," in *Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on*, Oct 2014, pp. 153–154.

[26] E. Partanen, T. Kujala, R. Ntnen, A. Liitola, A. Sambeth, and M. Huotilainen, "Learning-induced neural plasticity of speech processing before birth," *Proceedings of the National Academy of Sciences*, vol. 110, no. 37, pp. 15 145–15 150, 2013. [Online]. Available: http://www.pnas.org/content/110/37/15145.abstract

[27] M. Cheour, P. Leppnen, and N. Kraus, "Mismatch negativity (mmn) as a tool for inverstigating auditory discrimination and sensory memory in infants and children," *Clin Neurophysiol*, vol. 111, pp. 4–16, 2001.

[28] L. Bernstein, S. Eberhardt, and M. Demorest, "Single-channel vibrotactile supplements to visual perception of intonation and stress." *J Acoust Soc Am.*, vol. 85, 1989.

[29] P. H. Graf, E. Cosatto, V. Strom, and H. F. J., "Visual prosody: Facial movements accompanying speech." *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[30] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: head movement improves auditory speech perception." *Psychol Sci.*, vol. 15(2), pp. 133–137, 2004.

[31] E. J. Krahmer and M. Swerts, "More about brows: a cross-linguistic analysis-by-synthesis study, in from brows to trust: Evaluating embodied conversational agents." *Kluwer Academic Publishers*, vol. 15(2), pp. 191–216, 2004.