Robot recognizing vowels in a multimodal way

Paul Valentin, Sofiane Boucenna, Philippe Gaussier, Alexandre Pitti

Robot recognizing vowels in a multimodal way

Paul Valentin
ETIS, UMR 8051,
ENSEA, Université Cergy-Pontoise
Cergy-Pontoise, France
firstname.surname@ensea.fr

Sofiane Boucenna
ETIS, UMR 8051,
ENSEA, Université Cergy-Pontoise
Cergy-Pontoise, France

Philippe Gaussier
ETIS, UMR 8051,
ENSEA, Université Cergy-Pontoise
Cergy-Pontoise, France

Alexandre Pitti
ETIS, UMR 8051, ENSEA,
ENSEA, Université Cergy-Pontoise
Cergy-Pontoise, France

***Abstract*—This paper presents a sensory-motor architecture based on a neural network allowing a robot to recognize vowels in a multi-modal way thanks to human mimicking. The robot autonomously learns to associate its internal state to a human's vowel as an infant would to recognize vowel, and learn to associate congruent information.**

***Index Terms*—Neural network, intermodality, sound, vision, robotics, social interaction.**

I. Introduction

The purpose of this paper is to model a neural network that allows reproducing the capacity of 4 months babies to associate a vowel with a visual pattern, which will allow us to reproduce Kuhl's experiment [1].

Her experiment consists in presenting faces to babies. These faces can move in congruence or not with a sound that is emitted by a speaker. During the experiment is ongoing she will follow the attention of the baby. The result shows that the infants have a tendency to be more focused on the face that is congruent with the sound. By reproducing this experiment we want to produce a model that allows intermodal learning, that allows to be more robust than learning with only one modality.

Some experiments already use the bimodality. For example, Miura's work [2] consisted in using the prediction of the sound to produce new vowels that match those of a person, Miura uses this bimodality for restraining the space of possibilities.

II. Model

Our sensory-motor architecture is based on PerAc[1], that enables learning, recognition, and imitation. It learns sensory-motor conditioning involving two data streams associated respectively to sensation (audio and visual signal) and action. This network has been used already in various experiments for facial recognition [3] and individual [4] thanks to social interaction and more particularly imitation between two partners.

[1]PerAC is an acronyme for perception-action, that is a neural network with conditional and unconditionnal links between sensorial and motor categorisation.

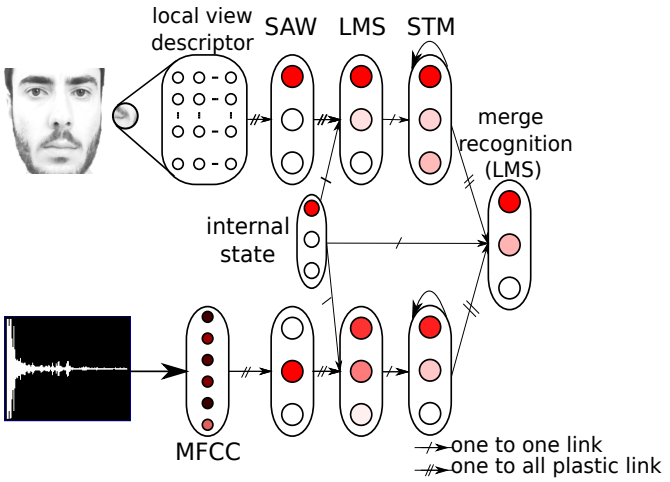

Fig. 1. The neural network architecture, we extract features from the sound and the vision, categorize them with a Self Adaptive Winner Takes All (SAW), then with a Least Mean Square (LMS) for obtaining the category desired. Then we take the most active neuron and we integrate that through the time.

The figure 1 shows the sensory-motor architecture based on our neural network. (1) In parallel a black and white image of size 320*240 pixels is processed, and the audio stream recording at 44100Hz. Visual processing enables the extraction of local views and the audio processing enables the extraction of Mel-Frequency Cepstral Coefficient (MFCC), (2) Each audio and visual feature is learned by the SAW (Self Adaptive Winner), an unsupervised system of clustering. It allows us to classify the features in one-shot, as described in [3], [4]. Each time the SAW receives an input vector, it compares this input vector to the prototypes previously learned. If the input is close enough to a prototype it averages the prototype with

the input, otherwise, it recruits a new neuron. (3) the motor production predictor [2] learns the association between the visual and audio features and motor production (a vowel that the infant performs during the learning phase). The robot interacts with the human partner to learn autonomously to recognize vowel. The partner is considered as a mirror: firstly the robot produces one vowel, and move its face to express the vowel according to its motor production. Then the partner imitates the robots vowel, thereby enabling the robot to link what it is seeing and hearing with its motor production.

Our experimental protocol consists of a subject who is sitting in front of a robot's face. The robot will firstly produce vowels: [a], [ə], [i], [o]. A neutral expression is done between each vowel to avoid human misinterpretation of the robot facial expression, that is performed by a set of 8 servomotors that represent: eyebrows, eyes, and mouth. It tells the vowels using a speaker, and pre-recorded sound file, then it will wait for the partner to repeat the vowel for 1.6 sec. The robot learns to link the sound and the image that the human produces. No fixed instruction is given to the person that imitate the robot face, we let them use their most natural posture in front of the robot, and speak as in their everyday life.

After this phase, where the robot learns vowels, we have a test phase. The subject produces a vowel's sound that the robot must imitate. If the robot has learned correctly then it should reproduce what the human is saying.

In order to test the model, we created a database that contains 6 persons. Each subject passed the test as described in the two previous paragraphs. The database is composed of 5 sound files which lasting 6.5 sec each and 40 images per vowel captured at a frequency of 6 Hz. The database is learned offline, under the same condition as online learning. Then our model is tested on the same database.

Our results are obtained on 200 iterations (the sensory-motor architecture must predict the correct vowel 20 times during period of 10 iterations), a mean of 26% bad recognition for the visual recognition, there was especially on the vowels [ə] and [o], that are extremely similar, 12% errors for the sound recognition, and 12% errors for the bimodality. The visual recognition has a bad recognition rate and doesn't help very often, except for a few particular cases. This is the consequence of two things: the freedom of the subjects to move or to have a bad positioning during the learning phase, and the fact that [ə] and [o] are visually very similar.

But, even with this difficulty, we can see on figure 2 some advantages of using bimodality rather than using only the sound if we plot the response of the neuron over time we can see that bimodality can avoid some errors, or let us recognizes vowels with higher certainty. In some cases, especially with vowels [a] and [i], vision can avoid errors.

[2]The recognition of a specific vowel induces a specific motor configuration, that is hard coded. The fundamental question of how the robot produces sounds is not addressed here.

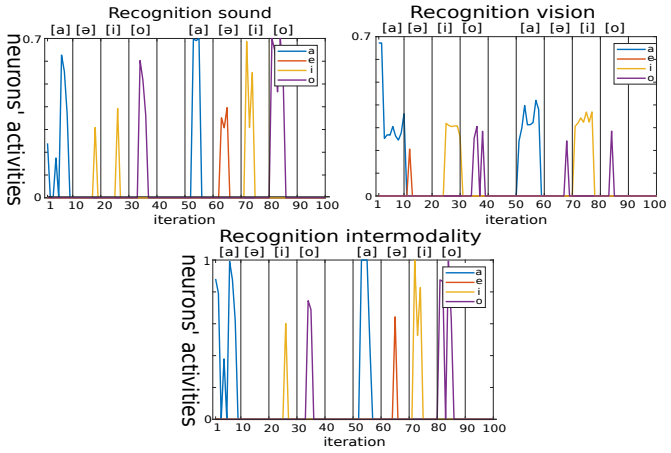

Fig. 2. This figure represents the activities of the neurons that code each vowel through the time. The data represented here was taken on one subject, during two cycles were the robot should imitate the human.

III. Conclusion

In this paper, we have presented a model, that allows a robot to learn to recognize vowel thanks to two modalities and social interaction with a human. Our neural network allows recognizing visual and sound stimuli, to gain more robustness.

Thanks to an imitation between the robot and one caregiver, we can learn to associate visual stimuli to a sound and reciprocally a sound to an image, and so report on Kuhl's experiment.

In future work, we will assess the effect of the type of the partners i.e.: adult typically developing, or children with autism spectrum disorder, on robot learning during the imitation. In the long-term, we want to link sound and image of a full word to allow the robot to predict the image of the object by the sound, or reciprocally. This will allow us to have a first communication way with a robot. For instance, that would allow us to just name an object that belongs to a scene, and let the robot find it, or show an object and name it.

Acknowledgment

We want to thanks Université de Cergy-Pontoise and Paris//Seine, and the doctoral school EM2PSI to support this research.

References

[1] Patricia K Kuhl and Andrew N Meltzoff. The bimodal perception of speech in infancy. Science, 218(4577):11381141, 1982.

[2] Yuichiro Yoshikawa, Minoru Asada, Koh Hosoda, and Junpei Koga. A constructivist approach to infants vowel acquisition through mother-infant interaction. l three experiments we used developmental robotics and computer modeling to implement a test of the idea that preverbal mutual imitation of actions between infant and caretaker may sConnection Science, 15(4):245258, 2003.

[3] Boucenna, S., Gaussier, P., Andry, P., & Hafemeister, L. (2014). A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game. International Journal of Social Robotics, 6(4), 633-652.

[4] Boucenna, S., Cohen, D., Meltzoff, A. N., Gaussier, P., & Chetouani, M. (2016). Robots learn to recognize individuals from imitative encounters with people and avatars. Scientific reports, 6, 19908.