Cognitive Architecture for Joint Attentional Learning of word-object mapping with a Humanoid Robot

Jonas Gonzalez-Billandon[1,2*], Lukas Grasse[3*], Alessandra Sciutti[4], Matthew Tata[3], Francesco Rea[1]

***Abstract*— Since infancy humans can learn from social context to associate words with their meanings, for example associating names with objects. The open-question is which computational framework could replicate the abilities of toddlers in developing language and its meaning in robots. We propose a computational framework in this paper to be implemented on a robotics platform to replicate the early learning process of humans for the specific task of word-object mapping.**

I. INTRODUCTION

The success of the application of deep learning methods in robotics has enabled state-of-the-art results for visual and speech recognition and for learning complex behaviours. However, these deep networks rely on supervised learning with large and annotated datasets. The creation of such datasets requires time and human supervision. Robots can, thanks to their embodiment, gather a significant amount of data from different modalities (cameras, microphones). These data, if annotated, could be used to train deep networks in a supervised way. For example, for the task of object localisation the different objects recorded from the robot camera must be annotated and labelled manually. The challenge is to find a way to let robots learn these labels autonomously, associating an object with a semantic symbol to categorize it. Human babies, in their early development easily learn to associate semantic meaning to their experience through the development of language. Language can create a mapping from words to objects, actions, and emotions [1]. Human infants decompose speech into approximately 40 phonemic bins based on the languages they are exposed to [2]. An important component of early language learning is the link between heard acoustic patterns of speech and produced speech. This is shown by the fact that, even early on, infants begin to produce babble that is unique to the phonemes in their native language [3].

We set out to study how this learned link between hearing and reproducing phonemes might be implemented in robotic perception. Social interaction plays an important role in the development of language. The specific ability to associate objects to words in toddlers has been found to emerge from joint attentional mechanisms [4] during social interaction. Visual attention for example, allows a baby to focus on an object and retain its relevant information.

In this paper, we propose a cognitive architecture that allows a robot to learn to associate visual and audio signals and to create a basic language driven by joint attention in a natural social interaction.

We propose to use a VQ-VAE network [5] to extract a discrete latent representation from the audio signal, which is then used to train a deep network model for object localisation. The latent vector is used as label in a supervised fashion and is associated with an object during social interaction. The localisation of the object results from our work on integration of visual and stereo-vision attentional mechanism inspired by early vision in humans for a robotic platform (PROVISION) [6]. We leverage also on our work on a bio-inspired auditory attentional mechanism for the robot iCub [7] to extract the relevant audio. The sensorial matching of auditory and visual signals is performed by making the robot point to the attended object similarly to what happens in learning infants. In fact, the proposed architecture intends to replicate the natural interaction between infants and adults with a robot. Specifically, it will replicate the scenario where an adult plays with objects in front of a baby and teaches the name of the objects in the environment.

II. ARCHITECTURE

The architecture of our proposed system consists of five components:

1) Visual Attention and Segmentation
2) Robot Non-verbal Response
3) Auditory Attention and Acoustic Feature Extraction
4) Object-Name Association Learning
5) Object Identification and Speech Production

as shown in Figure 1.

A. *Visual Attention and Segmentation*

The first component of the system is a model of bottom-up visual attention based on [6]. This component allows a human teacher to direct the robot's attention to an object for training. The assumption is that the human teacher will naturally redirect the infant‘s attention to the object by moving it. Next, through an elementary computation chain leveraging on vergence and zero-disparity the object of interest is segmented from the background [8].

B. *Robot Non-verbal Response*

After the robot has located and segmented the object, the robot communicates to the human its focus of attention with implicit signals such as gaze and pointing behaviour, as infants typically interact. For example, the action of pointing has been demonstrated to occur early on by infants and has also been shown to be important for learning [2].

*Co-Authors - E-mail: jonas.gonzalez@iit.it, lukas.grasse@uleth.ca

[1]Istituto Italiano di Tecnologia, RBCS dept, Genova, Italy

[2]The University of Genova, Genova, DIBRIS dept, Italy

[3]The University of Lethbridge, Neuroscience/CCBN dept, Canada

[4]Istituto Italiano di Tecnologia, CONTACT dept, Genova, Italy

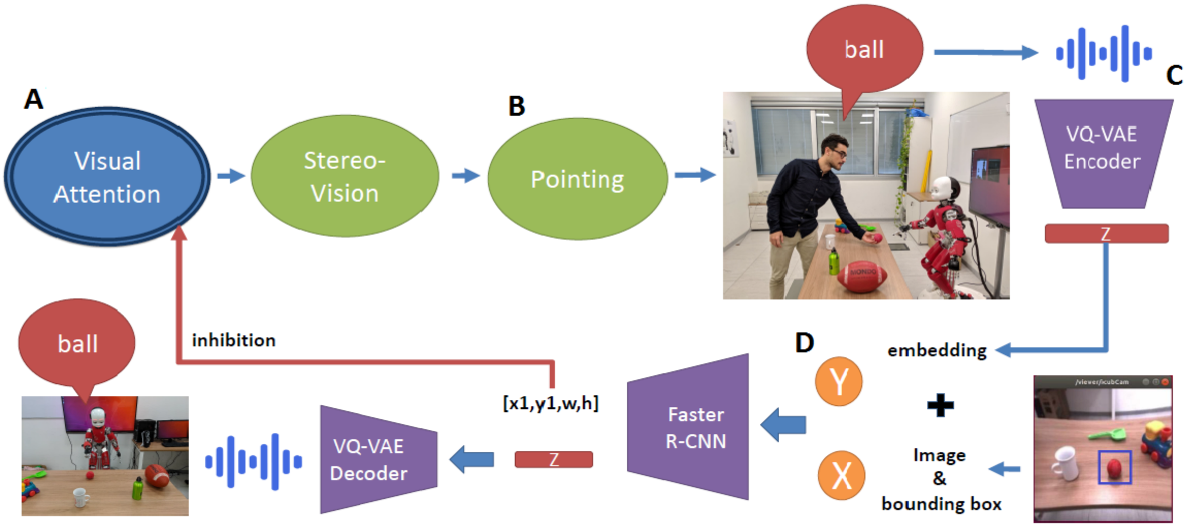

Fig. 1. **Cognitive architecture:** Illustration of a natural interaction between the iCub robot and a human partner with the different processes involved to replicate language grounding .

C. *Auditory Attention and Acoustic Feature Extraction*

Once the robot has communicated to the human teacher that it is attending the target object, the teacher says the name of the object. The auditory attention model [7] coupled with an voice activity detection system [9] allow to capture the relevant audio signal associated with the object. We convert the audio signal to an embedding space where embeddings for the same words are closer than different words, regardless of speaker. To do this we use a Vector Quantized-Variational Autoencoder (VQ-VAE) network to convert the audio into an embedding. When VQ-VAE networks are trained on audio, their embeddings have been shown to be similar to phonemes [5]. This means they can be used to model the content of speech while not being influenced by the lower-level differences in speech from different talkers.

D. *Object-Name Association Learning*

From the previous processes the meaningful embedding of the spoken name of an object (Fig 1: Y symbol) and its visuospatial information (Fig 1: X symbol) can be used to train a state-of-the-art object localisation network (YOLO, Faster-RCNN). This object-name association network is what enables the robot to learn the connection between objects with their spoken names.

E. *Object Identification and Speech Production*

Once the robot learned through interaction after gathering more data to train on, the network will generalize and allow the robot to localise and recognise objects. As the label used to train the architecture is the latent representation learnt from the encoder of the VQ-VAE it can be used in the decoder network to generate new audio. The robot can then finally pronounces the name of the object by playing the generated audio.

This process allows the robot to engage in an active learning procedure where it can point to a recognized object and name it and gather social signals to validate the learned association.

III. CONCLUSION

We propose in this work a bio-inspired cognitive architecture to replicate the autonomous mapping infants made between words and objects. We combine state-of-the-art supervised and unsupervised algorithms to guide the learning of the robot by merging auditory and visual signals in a natural interaction scenario.

ACKNOWLEDGMENT

Alessanda Sciutti was supported by a Starting Grant from the European Research Council (ERC) , G.A. No 804388, WHiSPER. Matthew Tata and Lukas Grasse were supported by an NSERC Canada Discovery Grant and Mitacs Globalink Research Award.

REFERENCES

[1] D. A. Baldwin and E. M. Markman, “Establishing word-object relations: A first step,” *Child development*, pp. 381–398, 1989.

[2] P. K. Kuhl, “Brain mechanisms in early language acquisition,” *Neuron*, vol. 67, no. 5, pp. 713–727, 2010.

[3] B. de Boysson-Bardies, “Ontogeny of language-specific syllabic productions,” in *Developmental neurocognition: Speech and face processing in the first year of life*. Springer, 1993, pp. 353–363.

[4] D. A. Baldwin, “Understanding the link between joint attention and language,” *Joint attention: Its origins and role in development*, pp. 131–158, 1995.

[5] A. van den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[6] F. Rea, G. Sandini, and G. Metta, “Motor biases in visual attention for a humanoid robot,” 2015.

[7] M. Mosadeghzad, F. Rea, M. S. Tata, L. Brayda, and G. Sandini, “Saliency based sensor fusion of broadband sound localizer for humanoids,” 2015.

[8] A. Dankers, N. Barnes, and A. Zelinsky, “MAP ZDF Segmentation and Tracking using Active Stereo Vision: Hand Tracking Case Study.”

[9] “Google webrtc.” [Online]. Available: https://webrtc.org/