Talking AIBO : First Experimentation of Verbal Interactions with an Autonomous Four-legged Robot

Frédéric Kaplan
Sony Computer Science Laboratory, Paris
6, rue Amyot 75005 Paris FRANCE
kaplan@csl.sony.fr
http://www.csl.sony.fr

Abstract

The aim of the 'Talking AIBO' project is to build a system enabling the AIBO, Sony's autonomous four-legged robot, to learn how to interact with humans using real words. We review in this article an experiment in which the robot builds a vocabulary concerning the objects it perceives visually. We discuss the results of this first prototype and the difficulties we have encountered.

Keywords: situated verbal interactions, autonomous robot, artificial pet.

1 Introduction

Autonomous robots are designed to behave in real environments without human intervention. During the last ten years, they have constituted a important research field in which a number of progress has been made. Research has mostly focused on how to manage the integration of different tasks into a single software architecture that could run on "low-cost" robots (see Steels and Brooks (1994), Arkin (1998) and Pfeiffer and Scheier (1999) for different surveys of the field). Robots are now able to walk on legs (most of them in an animal-like fashion using 4 or 6 legs), avoid obstacles, and navigate in unknown environments in a complete autonomous fashion. Some autonomous robots are used for agriculture, in the army and even in so-called "sportive" competition like robotic soccer (Asada et al., 1998).

During the same time, we have observed the emergence of a new kind of toys: the artificial pets. These apparently useless toys have been massively adopted, in Japan first, and progressively in the rest of the world. Children and also adults to a large extent (see Kusahara (2000)), have started to spend a significant part of their leisure time engaging in relationship with artificial creatures. Some were "portable" physical entities like *Tamagotchis* or *Furbies*. Others were only available as computer games like the *Norns* in the game *Creatures* (Grand et al., 1997)). We have argued elsewhere (Kaplan, 2000) that the success of the existing artificial pets relies on some clever design principles among which is the fact that they are "useless" in the sense that they do not perform any service task. They are designed to create a relationship with their owner.

The AIBO, which was commercialized by Sony in 1999, is both an autonomous robot and an artificial pet. It is the first product of a new generation of robots, designed for home entertainment. It has been created by the researchers at Sony Digital Creature Laboratory in Tokyo (Fujita and Kitano, 1998). The AIBO has a completely autonomous behavior. It has sensors corresponding to human senses (touch, audition, vision), learning ability and capacity to mature. It is not designed to be a slave as usual robots, it will not do something "useful" for its owner. For this very reason, it may actually be a companion with whom it is pleasant to interact.

Several research projects are currently conducted in various labs to explore more interesting interactions with artificial pets. *Robota* is able to learn and imitate gestures (Billard et al., 1998). *Kismet* engages in proto-conversation without semantic contents but in which a lot of non verbal

features are present (Breazeal, 2000). AIBO, in its current commercial version, uses body language and simple musical melodies to communicate with people. The aim of the 'Talking AIBO' project is to build a system enabling the AIBO to learn how to interact with humans using real words.

Figure 1: For the Talking AIBO project, we use an enhanced version of Sony's autonomous four-legged robot.

The Sony CSL Team in Paris has been investigating for more than three years the emergence of communication systems between autonomous agents. In particular, we have conducted a large scale experiment, called the 'Talking Heads', involving different robotic sites in different places of the world. This experiment has shown how a group of agents can build categories and words from scratch to name simple colored shapes on a white board (Steels and Kaplan (1999)).

With the Talking AIBO project, we are investigating how these techniques could be applied to human machine interactions and in particular to interactions with autonomous robots. We have decided to keep the original autonomous behavior of the AIBO and build our system "on top of it". The system acts as a "cognitive layer" which interferes with the current autonomous behavior, without controlling it completely. The first prototype of this system uses an external computer to perform all the computations concerning the linguistic interactions. The robot we use is an enhanced version of the commercial AIBO which will be thereafter refered as *the robot*. The computer implements speech recognition facilities which enables interactions using real words. The computer also implements a protocol for sending and receiving data between the computer and the robot through a radio connection. Through this connection, the "cognitive layer" of the robot can perceive and influence its standard autonomous behavior.

We will review here a first experiment in which the robot builds a vocabulary concerning the objects it perceives visually.

2 Recognizing and naming objects

Recognizing and naming objects is a classic task that many Artificial Intelligence programs try to tackle. In the context of the Talking AIBO project, several constraints are imposed. As the robots

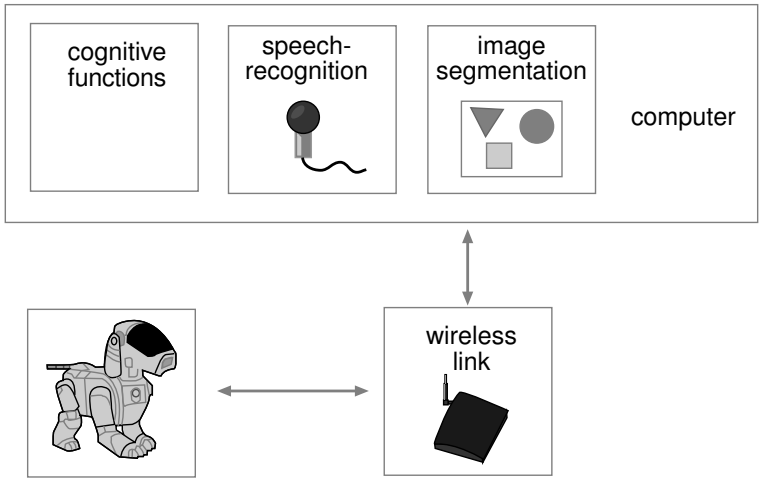

Figure 2: In this first prototype a computer implements speech recognition facilities and all the computations necessary for the verbal interactions.

is constantly moving driven by its autonomous behavior, objects can be seen under a large variety of viewing angles. The method we have chosen consists in a simplified version of the algorithm proposed by Duvdevani-Bar and Edelman (2000). The main idea is that the robot will associate with one word a collection of different views for which this word has been uttered.

2.1 Building simplified representations of the perceived objects

An image taken by the robot's camera is first segmented using a color detection algorithm. In this operation, the robot tries to separate the object in the scene from the background. We use a standard *growing regions* algorithm. Pixels are grouped into regions of similar colors. Then the regions are filtered in order to suppress too small and too large regions. The remaining regions constitute what we call *segments*.

Because the robot's camera has not a very wide view angle, only one object is generally in view. If several objects are segmented, one segment is chosen at random. The robot will consider this segment as the subject of the interaction.

Figure 3: Image perceived by the robot as analysed by the segmentation algorithm: The algorithm tries to pick out one single object in the image.

Once a segment has been selected, it is analyzed using a set of sensory channels corresponding to colour and shape properties (redness, blueness, yellowness, greenness, brightness, height, width,

rectangularity). The system is using the *opponent channels* Yellow-Blue and Green-Red instead of the standard Red-Green-Blue color decomposition. This system has been shown to be closer to human color perception (see Kaiser and Boynton (1996)). Using these sensory channels, the robot builds up a simplified representation of the perceived object, which can be summerized in a vector such as this one:

```
RED: 0.8
GREEN: 0.0
YELLOW: 0.2
BLUE: 0.0
BRIGHTNESS: 0.4
HEIGHT: 0.5
WIDTH: 0.6
RECTANGULARITY: 0.6
```

2.2 Associating words with the simplified representations

When it interacts with humans, the robot learns to associate words with its simplified representation of objects. The learning takes place during simple "language games" that the robot and its owner can play. Here is an example of a possible interaction (`H` stands for *human*, `R` for *robot*):

```
H: Listen
H: Ball.  [Showing ball]
(R looks at the ball and identifies the corresponding segment)
R: Ball ?
H: Yes.
```

When it hears "`Listen`" the robot waits for another word to be said. This second word will be associated with the segment currently identified. Several things can have already gone wrong at this point. The robot might not look to the right object or might have not segmented the image the "right" way. The word perceived by the speech recognition system might be the wrong one (e.g. It might have heard `Bad` or `Wall` instead of `Ball`). For this reason, the robot asks for a confirmation before creating any new association. After a positive confirmation, the robot associates the word with the simplified representation of the object.

As the robot plays language games, several views of the same object are associated to a single word like "`Ball`". The set of all perceptions associated with the word "`ball`" defines an implicit category (Kaplan, 1998). When a new object is perceived, the robot can try to recognize it using a *nearest neighbour* algorithm. The robot compares the segment seen with previous perceptions and utters the word associated to the ones which are the closest from the one currently perceived.

Several kinds of languages games can be played using this recognition technique. The simplest one is simply to ask the robot what object is within its vision range:

```
H: What is it ?
R: Ball
H: Good.
```

In case of failure, the user can teach the robot the right answer:

```
H: What is it ?
R: Peperoni
R: No, listen.  Ball.
R: Ball ?
H: Yes.
```

In order to make the interactions more natural, we have implemented several variants based on the same sort of scheme:

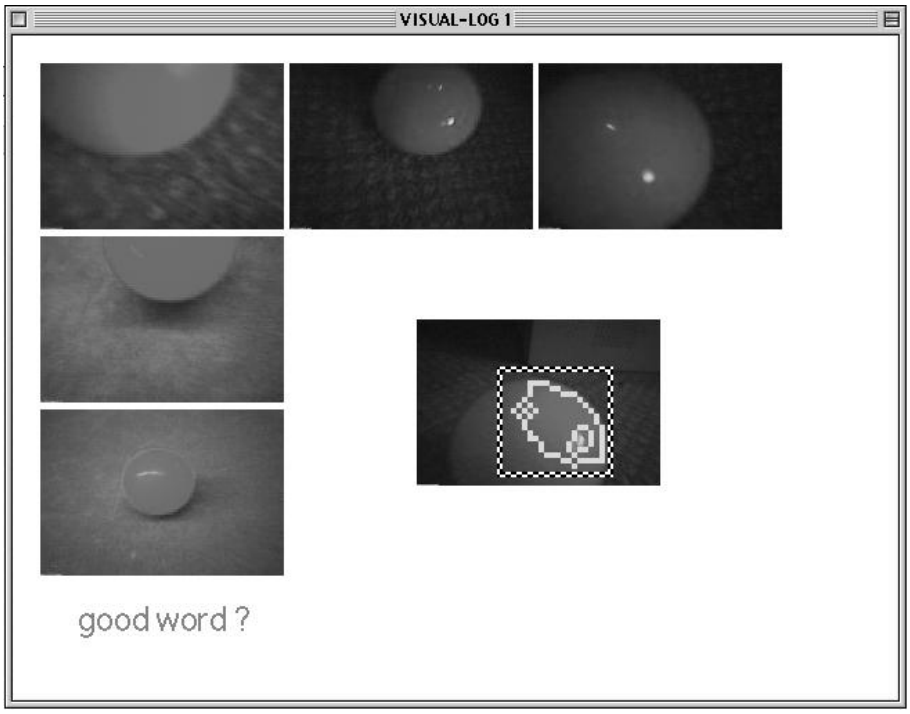

Figure 4: The segment currently perceived (in the center) is compared to previously stored segments. The closest segments correspond to interactions in which the word `ball` was uttered. The robot decides to call this current perception : `ball`.

```
H: Is it a Peperoni ?
R: No, it is Ball.
H: Good.
```

3 First results, first difficulties

Evaluation of such systems can only be obtained through the experimentation of repeated situated interactions. This first Talking robot has been trained and tested during several weeks. Interactions involving different kinds of colorful 'toys' (usually used for real pets) have been conducted with different teachers and under different lighting conditions. The robot has been able to recognize and name properly a dozen of different objects. During the experimentation, we have encountered a number of difficulties which are likely to reappear under more complex situated interactions.

3.1 Sharing attention, sharing topic

One of the main difficulties was to share attention with the robot. For instance, in order to indicate the name of an object, the robot had to look in the direction of the object and to pick the right 'segment' in the image perceived by the camera. We had already encountered similar problems during the Talking Heads experiment where the two robots had to point to indicate the meaning they were naming (we call this problem the 'Gavagai' problem in reference to the thought experiment described by the linguist Quine (Quine (1960))). Driven by its own autonomous behavior, the robot may have completely different goals which are in conflict with the aim of showing interest in the objects the teacher is presenting. As a consequence, it is sometimes very difficult to have it learn the word for some objects.

We have currently tackled this difficulty in the following way. First, we have adapted our teaching program to the robot 'natural tastes' by showing very colorful objects that we were

agitating in front of its camera (the robot is equipped with color and motion tracking capabilities). Then, in order to avoid too big effects of the errors in the topic choice, the robot was programmed to ask systematically for a verbal confirmation before associating a word with a new object. These two actions have increased significantly the quality of the lexicon learned by the robot. But the problems of sharing attention and sharing topic remain two critical issues. We are currently working on some more general solutions to address them.

3.2 Spontaneous verbal interventions

As soon as an autonomous agent is capable of spontaneous verbal interactions, a new problem arises: measuring what the human listener can cope with and what it cannot. Many robotic pets among the ones commercially available tend to finish their life with their batteries removed because they talk too much and in "inappropriate" circumstances. During our experiments, as the robot was asking too many times whether its use of a word was correct or not, it became clear that spontaneous speech should be avoided unless there is a special mechanism to manage it in an adapted fashion. We currently study how an autonomous agent can evaluate the level of irritation or tediousness it is generating and learn to stop talking when its clear that no one is interested anylonger in interacting with it.

4 Future work

As these first experiments gave encouraging results, we are now working on some complex behaviors involving verbal situated interactions. The goal of "finding an object" has been chosen as a first research topic. Several strategies can be built: moving around, asking directions, crying for help. We will study how the robot can develop linguistic competence in order to tackle this particular goal. For instance, in order to be guided towards the object the robot will have to learn words like "`walk`", "`turn left`", "`turn right`". The robot will test each strategy in interaction with the user, and choose the one that has proved to be the most successful. Apart from increasing the verbal capacities of the robot, future work will also include robot-robot grounded interactions and robot teleportation into other robotic bodies (Kaplan and Oudeyer, 2000).

5 Acknowlegments

The help and the comments of Luc Steels, Angus McIntyre, Pierre-Yves Oudeyer and Masahiro Fujita have been very precious to build this first prototype.

References

Arkin, R. (1998). *Behavior-Based Robotics*. MIT Press, Cambridge, MA.

Asada, M., Kuniyoshi, M., Drogoul, A., Asama, H., Mataric, M. Duhaut, D., Stone, P., and Kitano, H. (1998). The robocup physical agent challenge: Phase 1. *Applied Artificial Intelligence Journal*, 2–3(12):251–265.

Billard, A., Dautenhahn, K., and Hayes, G. (1998). Experiments on human-robot communication with robota, an interactive learning and communicating doll robot. In Edmonds, B. and Dautenhan, K., editors, *Socially situated intelligence workshop (SAB 98)*, pages 4–16.

Breazeal, C. (2000). Proto-conversations with an anthropomorphic robot. In *Proceedings of IEEE-ROMAN 2000 Workshop on Anthropomorphic Interactive Communication*.

Duvdevani-Bar, S. and Edelman, S. (2000). Visual recognition and categorisation on the basis of similarities to multiple class prototypes. *Int. J. Computer Vision*.

Fujita, M. and Kitano, H. (1998). Development of an autonomous quadruped robot for robot entertainment. *Autonomous Robots*, 5.

Grand, S., Cliff, D., and Malhotra, A. (1997). Creatures: Artificial life autonomous software agents for home entertainment. In *Proceedings of the First International Conference on Autonomous Agents,*, pages 22–29, New York. ACM Press.

Kaiser, P. and Boynton, R. M. (1996). *Human colour vision.* Optical Society of America, Washington, DC.

Kaplan, F. (1998). A new approach to class formation in multi-agent simulations of language evolution. In Demazeau, Y., editor, *Proceedings of the third international conference on multi-agent systems (ICMAS 98)*, pages 158–165, Los Alamitos, CA. IEEE Computer Society.

Kaplan, F. (2000). Free creatures: The role of uselessness in the design of artificial pets. In *Proceedings of the 1st Edutainement workshop.*

Kaplan, F. and Oudeyer, P.-Y. (2000). From talking aibo to talking head: Morphing sensory spaces between two robotic bodies. submitted to SRF 2000.

Kusahara, M. (2000). The art of creating subjective reality: an analysis of japanese digital pets. In C., M. and Boudreau, E., editors, *Artificial life VII Workshop Proceedings*, pages 141–144.

Pfeiffer, R. and Scheier, C. (1999). *Understanding intelligence.* MIT Press, Cambridge, MA.

Quine, W. (1960). *Word and Object.* The MIT Press, Cambridge, MA.

Steels, L. and Brooks, R. (1994). *The 'artificial life' route to 'artificial intelligence'. Building Situated Embodied Agents.* Lawrence Erlbaum Ass, New Haven.

Steels, L. and Kaplan, F. (1999). Situated grounded word semantics. In Dean, T., editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence IJCAI'99*, pages 862–867, San Francisco, CA. Morgan Kaufmann Publishers.