[self.]: Realization / Art Installation / Artificial Intelligence: A Demonstration

Axel Tidemann[1] and Øyvind Brandtsegg[2]

[1] Department of Computer Science, Norwegian University of Science and Technology
Trondheim, Norway
tidemann@idi.ntnu.no
[2] Department of Music
Norwegian University of Science and Technology, Trondheim, Norway
oyvind.brandtsegg@ntnu.no

Abstract. This interactive installation paper describes [self.], an open source art installation where the people interacting with it determine its auditory and visual vocabulary. When the system starts, it knows nothing since the authors have decided that it should be without any kind of bias. However, the robot is equipped with the ability to learn and be creative with what it has internalized. In order to achieve this behaviour, biologically inspired models are implemented. The robot itself is made up of a moving head, mounted with a camera, projector, microphone and speaker. As an art installation, it has a clear robotic visual appearance, although it is designed to demonstrate life-like behaviour. This is done by making the system start in a "tabula rasa" state, forming categories and concepts as it learns through interaction. This is achieved by linking sounds, faces, video and their corresponding temporal information to form novel sentences. The robot also projects an association between sound and image; this is achieved using neural networks. This provides a visual and immediate way of seeing how the internal representations actually learn a certain concept.

Keywords: artificial intelligence, robot, interaction, art.

1 Background

The original goal was to draw attention to how AI will effect human society, and subsequently what the man-machine-interaction can be like. To achieve this goal, various AI techniques were employed. A design goal was to reach for some sort of primitive consciousness, no matter how simple or fraught with errors it might be. This is of course in itself an enormously ambitious goal, but given the constraints (i.e. it being an art installation), we were not confined to rigorous standards or metrics, such as the Turing test or running up a staircase. Instead, we could focus on the *perceived* intelligence, knowing that the AI techniques implemented on the robot were devised in order to achieve such a goal. The art installation context gives a certain freedom from this kind of scientific measurement, rather

K. Chorianopoulos et al. (Eds.): ICEC 2015, LNCS 9353, pp. 517–522, 2015.
DOI: 10.1007/978-3-319-24589-8_49

putting the emphasis on how the different aspects of the intelligence can inspire reflection in the participants.

The art installation wanted to examine the relationship between technology and humans, and relates to language, philosophy and the contemporary (over-)focus on self realization. In order to highlight this, the robot had a raw design, with no attempt at anthropomorphization at all. The user interface was therefore rather crude, with exposed motors and wires, in order to illustrate what comprises an artificial intelligence. The user experience depends squarely on the behaviour of the robot.

2 Materials and Methods

Instructions on how to build the robot as well as source code to run it has been published online[1]. The architecture of the robot can be seen in Figure 1. The robot employs well-known methods from the sound processing and AI literature. The robot was built with an off-the-shelf moving head for stage lightning which was gutted except for the motors. On top of the moving head, a projector, USB camera, microphones and a speaker was mounted. The final build of the robot is seen in Figure 2.

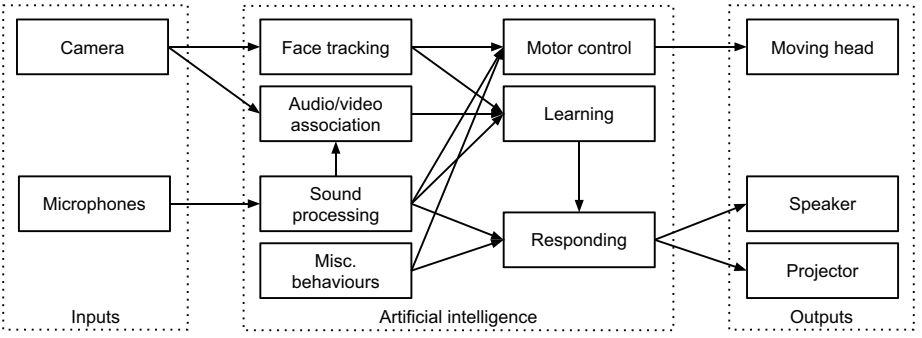

Fig. 1. The overall design of [self.] See the text for details pertaining to each module, the arrows indicate signal paths.

To enable the robot to sense the horizontal position of the sound source, two microphones were mounted on the moving head. These were mounted in a X-Y stereo configuration, allowing an amplitude comparison between the two to give information about the sound source's position. Using this information, the robot can turn towards the person talking to it. Figure 1 describes the overall architecture of [self.].

[1] `www.github.com/axeltidemann/self_dot`

The robot is equipped with a Kalman filter [8] to do face tracking. Csound[2] is used for both audio input and output. Csound is an open source programming language for sound with a rich API. Transient detection is used to segment the audio, this is done based on the amplitude slope and is therefore independent of the absolute amplitude. Upon detection of a transient, audio recording starts. It is stopped when the signal drops below a certain threshold, this is typically 6 dB below the initial transient. Before passing on the audio segment to other parts of the system, silent/noisy parts are deleted.

The raw WAVE files created by Csound are then processed by a biologically inspired model of the inner ear [9], an analysis based on a bio-mimetic "cascade of asymmetric resonators with fast-acting compression". The output is referred to as a neural activation pattern (NAP). This is different from a Fourier spectrum since the NAP has features that correspond to auditory physiology. The NAP is used for learning audio concepts, as well as building a transformation from audio to video.

We wanted to enhance the expressive capabilities of the robot by projecting an association between sound and image. [self.] records video at the same time as it records audio, and this video along with the NAP can then be used to train a neural network that represents a visual memory of the interaction. Echo State Networks (ESNs) [7] are used to form this audio/video transformation. The ESN is trained as follows: when a certain segment is to be learned, the NAP is used as input to the ESN and the corresponding images are used as the output of the ESN. The ESN then learns a mapping from the auditory domain to the video domain, and provides the visual output.

Both face tracking and sound localization influence motor control. Face tracking uses the estimation of the position of the face to rotate the moving head directly in front of the face. Similarly, the localization of sound in the horizontal plane guides the moving head to position it towards the sound source. The face tracking can be thought of as a higher level cognitive function than the sound localization. The organization of these two levels are inspired from Brooks' subsumption architecture [2], a sound can inhibit the motor signals determined by the face tracking. When someone is standing right in front of [self.] but not talking to it, the robot will then turn towards someone else who starts talking.

A crucial part of [self.] is its ability to learn and form categories by clustering similar sounds together. The robot builds an episodic audio-visual memory from each interaction. Similar sounds are clustered together, based on the Hamming distance [5] between sounds of roughly the same length. In this module, face recognition is also implemented. The "Face tracking" module also outputs the face it extracts from each image. Facial recognition is done by Support Vector Machines [4], and the different learning techniques are chosen based on empirical testing.

Context analysis is performed on the recorded audio segments, and this is used to create a set of quality dimensions. They forms a multi-dimensional web of associations, where dynamic weights are applied to each quality dimension.

[2] `www.csounds.com`

This is balanced and weighed similarly to fuzzy logic, where a variable can have partial membership in several relevant contexts. The response is based around the loudest sound perceived in a sentence, and this initiates the association that [self.] uses to generate a response, by comparing the Hamming distance of the new sound (i.e. NAP) to the ones already in memory. [self.] then looks for associations to this sound by looking up the various contexts, as described above. For instance, it can look for similar sounds, sounds uttered by the same face, and sounds in the same sentence or a specific time span. This creates a chain of associations, which yields a repository of sounds that are used to build a response sentence.

The visual output comes from the corresponding audio-visual ESN that was trained earlier. This network is then fed the NAP of the new sound, and the output is sent to the projector. This distributed representation of the audio/video-transformation gives a visual output that varies in accordance with how well the network recognizes the sound. If the sound is very similar to what it has been trained on, the visual output will typically consist of clear images. However, if the sound is very different from the original training sound, the resulting sequence of images will show this difference, e.g. as grey blurs or flashing images. This provides a more "life-like" visual presence of the robot, and provides some intuition into the learning process of [self.]. The memories are not retained forever - if a memory is not recalled, it will fade away as time passes.

Although [self.] is not pre-programmed with any kind of knowledge, it is programmed to perform certain behaviours over time since it is an installation. These are mostly implemented to avoid being caught in a "catatonic" state, and exhibit more life-like behaviour. There are three such behaviours: 1) *Idling:* the robot will search for a face in a pseudo-random pattern if it does not see anyone or hear anything for a certain period of time. Upon finding a face, [self.] finds the sounds it knows this person has said earlier, and uses these to initiate an interaction. If [self.] has not said anything for a certain period of time, it will start talking to itself, i.e. say something on its own. 2) *Dream state:* each night [self.] goes through all the memories experienced throughout the day. The learning process clusters similar sounds together, forming categories. Sometimes this clustering contains errors. The similarity between sounds in the same cluster can be estimated by calculating a sparse representation of the sounds [9], and those that are too different from the others are removed from the category. This resembles what the brain does during sleep [10], and can be thought of as some sort of "mental hygiene" process. 3) *Optimization of parameters for the response mechanism:* as described above, the various associations between sounds, faces and their sequences make it possible to form a multi-dimensional web that can be used to create a response. A large number of interdependent parameters needs to be adjusted in order to achieve the desired behaviour. The parameters themselves are sensitive to the state of the memory, i.e. how many memories are stored and their inherent sequences. [self.] is able to optimize these parameters on its own by using a genetic algorithm [6]. As a consequence, [self.] writes a part of its own behavioural program by the use of evolution.

Fig. 2. A person interacting with [self.]

Sound output comes from the speaker mounted on top of the projector, using Csound as the audio engine. In order to make the sound output more life-like and with a character corresponding to the person who taught the given word, the sound is transformed using granular [1] and spectral[3] transformation techniques. These transformations are subtle, with the goal of providing a gentle coloring and personalization of the sound, similar to what the ESN does in the video domain.

The primary moving head constitutes the motor output of [self.], which it uses to locate people. This is seen in figure 2. A secondary moving head (not pictured) is placed a few metres from the main moving head. This is similar to the main head, except it only contains an ultrasonic speaker. The purpose of an ultrasonic speaker is to create a tight sound beam[4], and we use the motorized head to direct this beam around the room. This speaker is used to play back the secondary associations of [self.] to the sounds it perceives, somewhat akin to "what you think about while speaking". These sounds are time stretched using granular techniques and the sound beam reflects on the wall of the room. This is intended to create an impression of "being inside the mind" of [self.] since its secondary associations are projected as sound all around the room, and the time stretching creates a dreamy slow moving representation of the secondary sounds.

3 Performance

When creating the art installation, the focus has been to avoid pre-programming the robot with knowledge, and letting the robot build up its own knowledge base from scratch through interaction. When seeing the system in action, the authors observe how this design choice encourages interactivity, since people interacting

[3] `www.csounds.com/manual/html/SpectralRealTime.html`
[4] `www.holosonics.com/technology.html`

with [self.] find it rewarding to recognize themselves or someone else they know. To achieve this goal, biologically inspired models are implemented, as a means of implementing some form of cognition in the robot. By employing the web of contexts, it can also answer creatively, something that is further enabled by the use of evolution to write parts of the behavioural programme. This is discussed more in [11]. Since the art installation is completely open in terms of what it receives as input, the user experience is dependent on what it learns throughout the installation period, which gives it an organic feel.

Since [self.] is completely open source, we envision that it can be used also as a research platform for the interplay between people and artificial intelligence. However, the authors are very aware that the system is far from being a truly sentient AI. Even though the robot learns like a child, it currently does not have the possibility to grow into an adult in terms of reasoning power and deeper knowledge of its surroundings, since AI as a field has not progressed to this level yet. On the other hand, this serves as a motivation to continue implementing models of human cognition to get closer to this goal. This kind of *cognitive incrementalism* has been regarded as a way of achieving full-blown human cognition by gradually adding cognitive bells and whistles to an entity [3].

References

1. Brandtsegg, Ø., Saue, S., Johansen, T.: Particle synthesis – a unified model for granular synthesis. In: Linux Audio Conference (2011)
2. Brooks, R.: A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation 2(1), 14–23 (1986)
3. Clark, A.: Mindware. Mindware (2001)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
5. Hamming, R.: Error detecting and error correcting codes. The Bell System Technical Journal 29(2), 147–160 (1950)
6. Holland, J.H.: Adaptation in Neural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
7. Jaeger, H., Haas, H.: Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. Science 304(5667), 78–80 (2004)
8. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Fluids Engineering 82(1), 35–45 (1960)
9. Lyon, R.F., Rehn, M., Bengio, S., Walters, T.C., Chechik, G.: Sound retrieval and ranking using sparse auditory representations. Neural Computation 22(9), 2390–2416 (2010)
10. Stickgold, R., Hobson, J.A., Fosse, R., Fosse, M.: Sleep, learning, and dreams: Off-line memory reprocessing. Science 294(5544), 1052–1057 (2001)
11. Tidemann, A., Brandtsegg, Ø.: [self.]: an Interactive Art Installation that Embodies Artificial Intelligence and Creativity. ACM Cognition + Creativity (to appear)