A Social Robot as an Aloud Reader: Putting together Recognition and Synthesis of Voice and Gestures for HRI Experimentation

Arnaud Ramey
RobotiscLab, Univ. Carlos III
of Madrid
Leganés, Spain
arnaud.ramey@m4x.org

Javier F. Gorostiza
RobotiscLab, Univ. Carlos III
of Madrid
Leganés, Spain
jgorosti@ing.uc3m.es

Miguel A. Salichs
RobotiscLab, Univ. Carlos III
of Madrid
Leganés, Spain
salichs@ing.uc3m.es

ABSTRACT

Advances in voice recognition have made possible applications in robotics controlled by voice only. However, user input through gestures and robot output gestures both create a more vivid interaction experience. In this article, we present an aloud reading application offering all these interaction methods for the HRI-research robot Maggie. It gives us a testbed for user studies investigating the effect of these additional interaction methods.

Categories and Subject Descriptors

H.5.2 [**User interfaces**]: Voice I/O; I.2.9 [**Robotics**]: Operator interfaces; H.5.2 [**User interfaces**]: Theory and methods

General Terms

Design, Experimentation, Human Factors

Keywords

Gesture Recognition, ETTS, Multimodal Interaction

1. INTRODUCTION

Over the last few years, both voice recognition and synthesis have made great advances, which offers new possibilities in robotics, and especially makes possible creating voice-only controlled robots, such as the flower robot presented in [2]. In [4], it is made use of a robot with optional arms and eyes to test the attention of the users to the robot instructions. It underlines the gain of interest from the users when the robot performs, in addition to the voice, human-like gestures. It also stresses the difficulties of older users to use dialogue-only based interfaces, which suggests the use of additional interaction methods.

In this article, we present an application for the social robot Maggie converting it into an aloud reader: it can utter aloud some content such as books or news feeds. We couple both approaches: a voice-enabled interface with additional support for gesture-based input and output.

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.
ACM 978-1-4503-1063-5/12/03.

2. DESCRIPTION OF THE SYSTEM

The robot Maggie hardware and software architectures are presented in [1]. The latter is made of parallel modules synchronizing with one another thanks to short messages called *events*. A detailed description of the developed modules follows.

2.1 Voice recognition and synthesis

Speech recognition (ASR) in Maggie is based on the commercial software suite Loquendo ASR. For voice synthesis, the robot presents an Emotional Text-To-Speech ($ETTS$) architecture: it enables the conversion of a written text into sound, but with a voice depending on the emotional state of the robot. This affects a number of parameters, as the voice pitch, its speed, its volume, etc. It is based on the Loquendo ETTS software.

2.2 Gesture recognition

A gesture recognition system has been previously developed for Maggie [5]. It makes use of the Kinect time-of-flight camera. So-called *primitives*, which correspond to quick moves of the user's hand into a given direction, are detected. For instance, a sweeping move to the right (along the x axis) is a primitive. A full gesture is then represented as a sequence of these primitives. Each gesture is associated to an user-defined action, which is executed when the gesture is recognized.

2.3 Gesture synthesis

The robot is able to express her mood, if is paying attention, and other linguistic gestures as greetings, affirmation, agreement, pointing, etc.

Most of the gesture systems on Social Robotics are based on a *gestionary*, that is, a dictionary or a set of discrete gestures ([3], and many others). The system developed here also include the feature of generating gestures as "attitudes": they can be also derived on the fly as a dynamic movement based on sensory-motor feedback functions. For example, the robot can exaggerate a happiness attitude when the user is closer.

2.4 Core application: the aloud reader

The robot will read aloud some user-chosen content, while obeying user instructions and performing output gestures.

2.4.1 Input sources

The texts supplied as input can currently be of two types. Use can be made of English and Spanish literature books provided by the Gutenberg project, a project aiming at making available online books with an expired copyright. The robot can also subscribe to RSS feeds from important national and international newspapers, such as El Mundo.
The raw text can be extended thanks to metadata, which goal follows.

2.4.2 Application output

The main expected outputis the proper aloud reading, that is, is the vocal pronunciation of the wanted text. The text fetched as presented in 2.4.1, is cut into short chunks sent to the ETTS module.
Besides, [6] for instance showed that an ETTS system that does not take into account the emotional state of the content strongly lacks of expressivity. When encountered, the metadata tags, which come within the raw text, but are indicated via a simple markup language. affect the robot current emotion, resulting in changes in both its way of moving and speaking. as seen in 2.1 and 2.3. For instance, an excerpt, if it comes with a "sad" metadata tag, will be read with a sad voice and gestures showing a sad behavior.

2.4.3 Application control: input methods

The software architecture allows the application control with *events*, which can easily be emitted from any other process. Thus, it is easy to turn an existing program into a control module of the aloud reader. The current structure is visible in figure 1.
Voice orders can be also transmitted to the application thanks to the ASR system seen in 2.1. If the user utters the order *"go to next headline"*, the ASR module will convert that information into the corresponding event, which will make the robot perform the wanted action.
The gesture recognition system presented in 2.2 is used to give orders to the robot. In a similar way to the voice control, a sweeping gesture will be converted into the appropriate event to go to the next headline or next chapter.

3. RESULTS AND CONCLUSIONS

The aloud reader application has been implemented and tested in the robot Maggie. A sample picture of an end-user using it is visible in figure 2 The gesture recognizer lightweight workload allows it to run in the background with no delay on the recognition, as does the ASR and coupled with the agile handling of the queue of ETTS utterances, a fluent reading without pauses is generated.
We now aim at measuring how both input and output gestures change the user experience. A user study would allow us to quantify if they help maintaining the users' attention and how.

4. ACKNOWLEDGMENTS

The authors gratefully acknowledge the funds provided by the Spanish MICINN (Ministry of Science and Innovation) through the project "*A new Approach to Social Robotics (AROS)*".

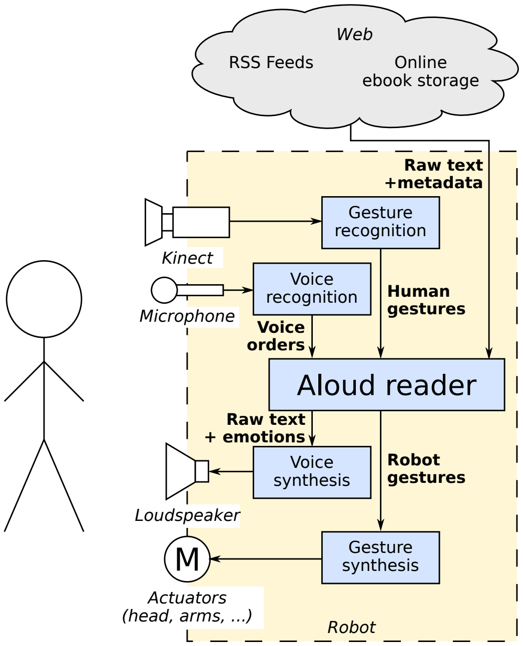

Figure 1: The current architecture of the aloud reader

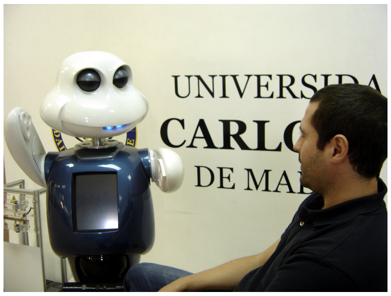

Figure 2: An user listening to the robot Maggie as an aloud reader, which gestures show its current (happy) emotion.

5. REFERENCES

[1] R. Barber and M. Salichs. *A new human based architecture for intelligent autonomous robots*, pages 85–90. Elsevier, 2002.

[2] S. Chang, S. Ham, and D. Suh. Rohini: A robotic flower system for intuitive smart home interface. In *Control Automation and Systems (ICCAS), 2010 International Conference on*, pages 1773 –1776, oct. 2010.

[3] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu. Robovie: an interactive humanoid robot. *Industrial Robotics*, 28:498–503, 2001.

[4] H. Osawa, J. Orszulak, K. Godfrey, M. Imai, and J. Coughlin. Improving voice interaction for older people using an attachable gesture robot. In *RO-MAN, 2010 IEEE*, pages 179 –184, sept. 2010.

[5] A. Ramey, V. González-Pacheco, and M. A. Salichs. Integration of a low-cost rgb-d sensor in a social robot for gesture recognition. In *Proceedings of the 6th international conference on Human-robot interaction*, HRI '11, pages 229–230, New York, NY, USA, 2011. ACM.

[6] S. Roekhaut, J.-P. Goldman, and A. C. Simon. A model for varying speaking style in tts systems. *Proceedings of the 5th International Conference on Speech Prosody SP2010*, (Table 1):4–7, 2010.