Ask Alice; an Artificial Retrieval of Information Agent

Michel Valstar
University of Nottingham

Catherine Pelachaud
CNRS

Dirk Heylen
Twente University

Angelo Cafaro
CNRS

Soumia Dermouche
CNRS

Alexandru Ghitulescu
University of Nottingham

ABSTRACT

We present a demonstration of the ARIA framework, a modular approach for rapid development of virtual humans for information retrieval that have linguistic, emotional, and social skills and a strong personality. We demonstrate the capabilities of our framework in a scenario where a popular book from the English literature, 'Alice in Wonderland', is embodied by a virtual human called Alice. During the presentation one can engage in an information exchange dialogue, where Alice acts as the expert on the book, and the user as an interested novice. Besides speech recognition, sophisticated audio-visual behaviour analysis is used to inform the core agent dialogue module about the user's state and intentions, so that it can go beyond simple chat-bot dialogue. The behaviour generation module features a unique new capability of being able to deal gracefully with interruptions of the agent.

1. INTRODUCTION

Task-specific AI is attaining super-human performance in an increasing number of domains. In the near future, virtual humans (VHs) will be the human-like interface for increasingly capable AI systems, in particular information retrieval systems. However, there remains a large gap in the smoothness of interaction between either a current VH or another human being. In the Horizon 2020 project ARIA-VALUSPA we aim to drastically reduce this gap.

This means first and foremost that interacting with the ARIA-agents should be engaging and entertaining. They should display interactive believable behaviour that feels real. They should be adaptive to the user at various levels, from adapting to a user's appearance, age, gender, and voice, to sudden changes in the dialogue initiated by the user. As part of ARIA-VALUSPA, we have developed an interactive virtual human that can hold a prolonged dialogue about the book 'Alice in Wonderland', by Lewis Caroll.

Some particular challenges that we have addressed in the project and that we wish to demonstrate here are to deliver a reusable framework that can be used to create VHs with different personalities, behaviours, and underpinning knowledge bases. The framework is in principal independent of the language spoken by the user. Another important challenge that we set ourselves and will demonstrate here is to be able to deal with unexpected situations, in particular interruptions initiated by the user. This is a hard problem that has not been addressed previously.

ICMI 2016 Tokyo, Japan

DOI: 10.475/123_4

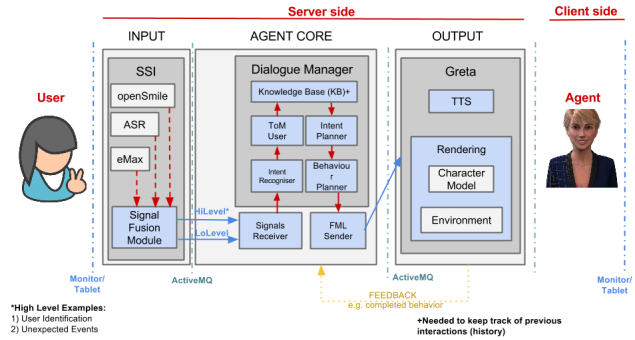

Figure 1: ARIA framework architecture, composed of modular Input, Agent Core, and Output blocks.

2. ARIA FRAMEWORK

The 'Ask Alice' demonstration is built on top of the ARIA Framework. This is a modular architecture with three major blocks running as independent binaries whilst communicating using ActiveMQ (see Fig. 1). Each block is in turn modular at a source-code level. The Input block processes audio and video integrated in SSI [2] to analyse the user's expressive and interactive behaviour and does speech recognition. The Core Agent block maintains the agent's information state, including its goals and world representation. It is responsible for making queries to its domain-knowledge database to answer questions. Once all goals and states are taken into account it decides on what agent behaviour should be generated. The Output block generates the agent behaviour, that is, it synthesises speech and visual appearance of the virtual human. The ARIA Framework makes use of communication and representation standards wherever possible. For example, by adhering to FML we are able to plug in two different visual behaviour generators, either CNRS' Greta [1] or Cantoche's Living Actor technology.

The ARIA framework's Input block includes state of the

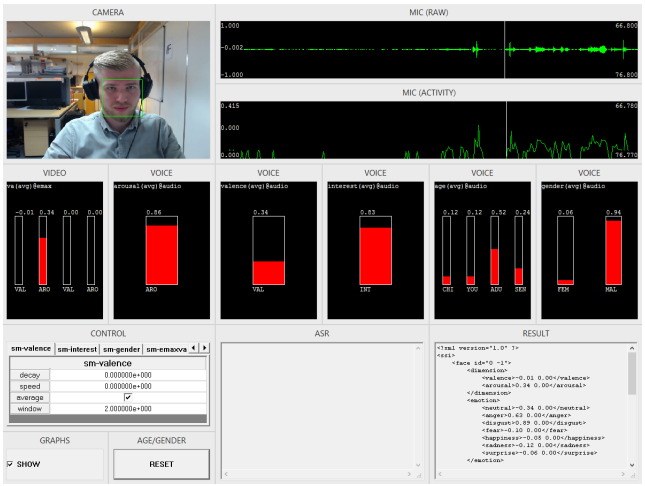

Figure 2: Input Block Visualisation

art behaviour sensing, many components of which have been specially developed as part of the project. From Audio, we can recognise gender, age, emotion , speech activity and turn taking [7, 6], and a separate module provides speech recognition [4]. Speech recognition is available for the three languages targeted by the project. From Video, we have implemented face recognition, emotion recognition [3], detailed face and facial point localisation [5], and head pose estimation. Fig. 2 shows a visualisation of the behaviour analysis.

3. DEMONSTRATION

In the demonstration, a single user will be invited to face Alice, who is displayed on a large screen. The invitation will either be done by a researcher or by Alice herself, if it detects the presence of a new face and isn't already engaged in an interaction with someone else. Once Alice has detected that the user is engaging with her, she will initiate a greeting process, and then introduce the topic of the book, Alice in Wonderland. Alice will first try to establish if the user has any domain knowledge, for example by determining whether they've read the book, seen the original animation film or the later hollywood film of the book. Depending on this domain knowledge, she will either elaborate on some more background information about the book or will dive straight into offering her views on the book, and allowing the user to ask questions and provide their own opinion. This interaction lasts until the user is satisfied with the interaction, or until Alice gets bored with the user. Fig. 3 shows the Living Actor output of the demo, i.e. Alice, and Fig. 2 shows the user interacting with Alice.

Because of the framework's ability to adapt to people who interact with it, Alice will be able to recognise a user and pick up a conversation with someone who previously visited the demo. She will also be able to deal with common issues related to interruptions that occur during a typical conference demo setting, i.e. multiple people interacting with it, or a user suddenly addressing one of their colleagues instead of Alice, or a user simply walking away from the interaction mid-interaction.

Figure 3: Samples of Alice expressive posing during an interaction with the user.

4. ACKNOWLEDGMENTS

Funded by European Union Horizon 2020 research and innovation programme, grant agreement No 645378.

5. ADDITIONAL AUTHORS

Additional authors: Elisabeth André, Tobias Bauer, Johannes Wagner (University of Augsburg), Laurent Durieu (Cantoche), Matthew Aylett, Pascal Blaise (Cereproc), Eduardo Coutinho, Björn Schuller, Yue Zhang (Imperial College London), Mariet Theune, Jelte van Waterschoot (Twente University).

6. REFERENCES

[1] F. De Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. De Carolis. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International journal of human-computer studies*, 59(1):81–118, 2003.

[2] S. Flutura, J. Wagner, F. Lingenfelser, and E. André. Mobilessi - asynchronous fusion for social signal interpretation in the wild. In *Proc. of ICMI*. ACM, 2016.

[3] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8, March 2016.

[4] A. E.-D. Mousa and B. Schuller. Deep bidirectional long short-term memory recurrent neural networks for grapheme-to-phoneme conversion utilizing complex many-to many alignments. In *Proc. Interspeech*, 2016.

[5] E. Sánchez-Lozano, B. Martinez, and M. Valstar. Cascaded regression with sparsified feature covariance matrix for facial landmark detection. *Pattern Recognition Letters*, 73, 2016.

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, S. Zafeiriou, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. ICASSP*, pages 5200–5204. IEEE, 2016.

[7] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, et al. Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5185–5189. IEEE, 2016.