Metalogue: A Multimodal Learning Journey

Dimitris Koryzis
Hellenic Parliament,
Athens, Greece
dkoryzis@parliament.gr

Vasileios Svolopoulos
Hellenic Parliament,
Athens, Greece
v.svolopoulos@parliament.gr

Dimitris Spiliotopoulos
Institute of Computer Science
Foundation for Research and
Technology - Hellas,
Heraklion, Greece
dspiliot@ics.forth.gr

ABSTRACT
In this paper, we present a high-level description of the Metalogue system that develops a multi-modal dialogue system that is able to implement interactive behavior between a virtual agent and a learner outlining the insight to the development of a fully-integrated multimodal interactive system. This system includes several components addressing several research domains: meta-cognitive modeling, skill training, usability testing, prosody analysis, multimodality, dialogue management, speech recognition, gesture recognition and interpretation, and learner feedback. The key issue is the integration of all these components in a single platform, allowing to the users to improve their metacognitive skills. This work reports on the user experience evaluation during the design and development phase of the system that feed back to the design and continuous refinement of the overall approach.

CCS Concepts
• **Human-centered computing~Human computer interaction (HCI)** • **Human-centered computing~Empirical studies in HCI** • *Applied computing~Computer-assisted instruction*

Keywords
multimodal interaction, meta-cognitive skills training, user experience, spoken dialogue.

1. INTRODUCTION
Natural dialogue systems shaped a significant and fast growing market segment, for instance in smart homes [3], intelligent environments [6], education and learning [2, 5], and so on. At the same time, the state-of-the-art dialogue systems do not fully satisfy requirements of natural and rich communication with humans. Multimodal natural language based dialogue is increasingly becoming the most attractive human machine interface, from information offices to smart houses, smart working environments and Internet of Things [8]. Such interfaces offer a mode of interaction that has certain similarities with natural human communication by using a number of input and output modalities which people normally employ in communication, like speech, gesture, facial expressions, voice tone, body language, etc. The multiple effects of multimodality and the sensor data in such environments allow for immersive learning experiences [9].

Our system is a multimodal dialogue system that is able to implement an interactive behavior that seems natural to users and is flexible enough to exploit the full potential of multimodal interaction between a user and a embodied agent (avatar). The objective is to achieve a framework for understanding, controlling and manipulating system and user cognitive processes [1]. The system implements a dialogue manager that incorporates a cognitive model based on meta-cognitive skills training that enables planning and deployment of appropriate dialogue strategies. This is achieved through monitoring both the system and user interactive performance, reasoning about the dialogue progress, calculating the users' knowledge and intentions, and thereby, adapt and regulate the dialogue behavior over the course of the human-system interaction. The meta-cognitive capabilities of the system are based on incorporating transfer of knowledge among skills.

The use case scenarios focus on educational, political, business and coaching situations where negotiation skills play a key role in the decision-making processes [4]. Given the complexity of performing a negotiation or a simple debate, the process of learning how to negotiate has to be incorporated into the system via specific learning patterns [11].

2. MOTIVATION
The prototype system aims to provide the learners with a rich and interactive environment that trains them in order to develop meta-cognitive skills, support motivation, and stimulate creativity and responsibility in the decision making and argumentation process. The system utilizes virtual dialogue agents capable of engaging in natural interaction through gestures, voice, mimicry and body language.

The system was deployed and tested in a specific use-case scenario: in social educational contexts for training young, politically active citizens in the framework of educational activities of the Hellenic Youth Parliament. It is also planned to be adapted for the business education context for training human call center agents to optimally identify customer claims and respond accordingly.

Major effort has been devoted to the system architecture that aimed to integrate multimodal input and output as well as the learning models, driven by multi-perspective dialogue management. In an attempt to involve the end users as early as possible into the design and development process, the system was deployed as an observer in order to investigate the user experience with the multimodal interactive feedback. The user feedback, needs and findings were used as functional requirements for the final prototype.

PETRA '16, June 29-July 01, 2016, Corfu Island, Greece
 ISBN 978-1-4503-4337-4/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2910674.2935860

3. SYSTEM ARCHITECTURE

3.1 Architecture objectives

The overall architecture can be viewed from the perspective of three distinct objectives. The first is a system architecture that supports several input and output modalities, such as spoken natural language, facial expressions, body posture and biosensor data [7], and is designed to be modality-agnostic. The second objective is to provide a toolkit for developers (API) to create interactive learning applications, connect new devices and interpretation functionality for their sensor data. The third objective is that the system architecture is modular, allowing for mix and match of open source components that are suitable for other use cases. For instance, the architecture design aims to be able to support subsets of the available modalities or data inputs, in order to accommodate functional requirements or technical setups.

3.2 Architecture overview

The system architecture is comprised of five distinct layers (Figure 1). Each layer corresponds to specific main functionalities and includes the respective components of the multi-modal dialogue trainer system:

- The recognition layer contains the components that handle the input to the system. Here, the raw data is first processed and an output for the next layer is generated. The input mainly comes from two types of sources, microphones and Microsoft Kinect. The process splits into two types of raw input. The input from the Kinect devices is analyzed by the motion and gesture recognizers. Raw signals are standardized to specific XML formats for data processing. The microphone input is audio or voice that is analyzed by the speech recognizer (speech to text) and the prosody emotion component. At the same time, the video recording of the conversation, via the Kinect, is stored for later use.

- The interpretation layer utilizes the data from the recognition layer. The data are interpreted as separate streams at this stage. The text from the spoken input is fed to the semantic analysis module. The 3D interpreter analyses the motion recognition data about body and face of the trainee, and the gesture interpreter formalizes the data from the gesture recognizer, classifying the instances of gestures.

- The interpreted events (semantic, face and body postures, gestures) are fused into lists of instances of interaction events as part of the core, dialogue management, layer. The semantic interpretation of the text is used for the dialogue acts, while the rest enable the dialogue management core functions and populate the discourse model of the system. Advanced processing is executed in this layer, more specifically where a synchronized interaction among Fusion, Discourse Model and Dialogue Manager is performed in order to deeply analyze behavior of the participant. The Fusion module acts as a gateway in order to receive standardized data from the previous layers, then it performs two essential operations: temporal synchronization and data smoothing. Subsequently, the Discourse Model monitors the participants' beliefs, desires and intentions. In addition, this module reports on the participant's general performance. Its output is delivered to the Dialogue Manager.

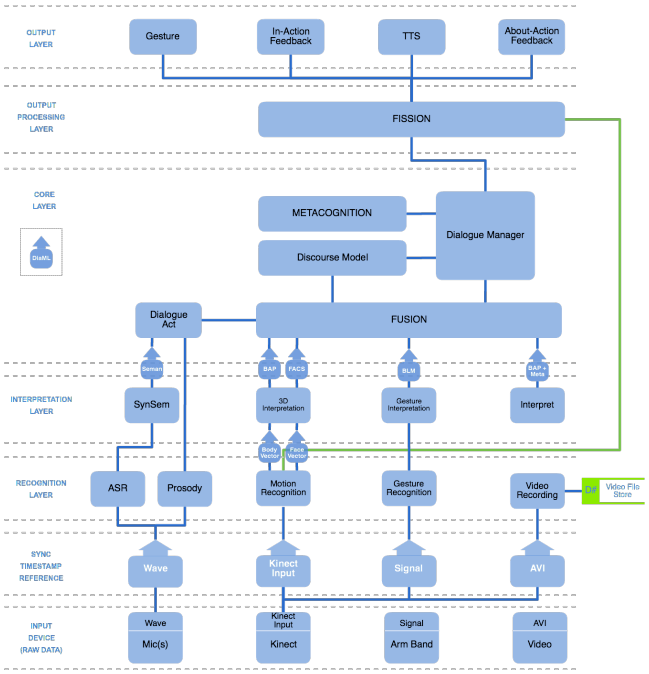

Figure 1. System architecture overview.

- In the output processing layer, a meta-module is built with the Fission component. Here, the Dialogue Manager output is the base for the generation of the system response and the selection of the appropriate means, for instance, as real-time feedback during the interaction (in-action feedback) or as summative revisit of the interaction after it ends (about-action feedback), which are presented to the debate participant.

- The output layer controls the system output modules. It communicates the information and supervises the response to the user according to the Fission calculations. The instructions pass in the form of content and functional parameters.

An important note on the above description is that the cognitive, learning and interaction models are closely integrated and direct the operation of a dialogue manager component for the learning environment processing. These components comprise the Metacognition module depicted in Figure 1.

The system architecture is designed to allow conversational interactions in real time and use processing shortcuts where time-critical reactions are desired. For example, if a spoken input is detected, immediate attention feedback via gaze is generated by the virtual characters, bypassing the slower full semantic processing of the input when necessary.

The above feature is very important in creating and preserving the perception of immersion and responsiveness for the user. Furthermore, the system features incremental processing, which results in a more natural and accurate way of achieving high quality interactive behavior. Rather than waiting for a complete utterance from the user (in the form of a sentence), other input (such as an emotion or gesture) is processed in real time, updating system understanding as new information is expressed. The system architecture is designed to be modality agnostic.

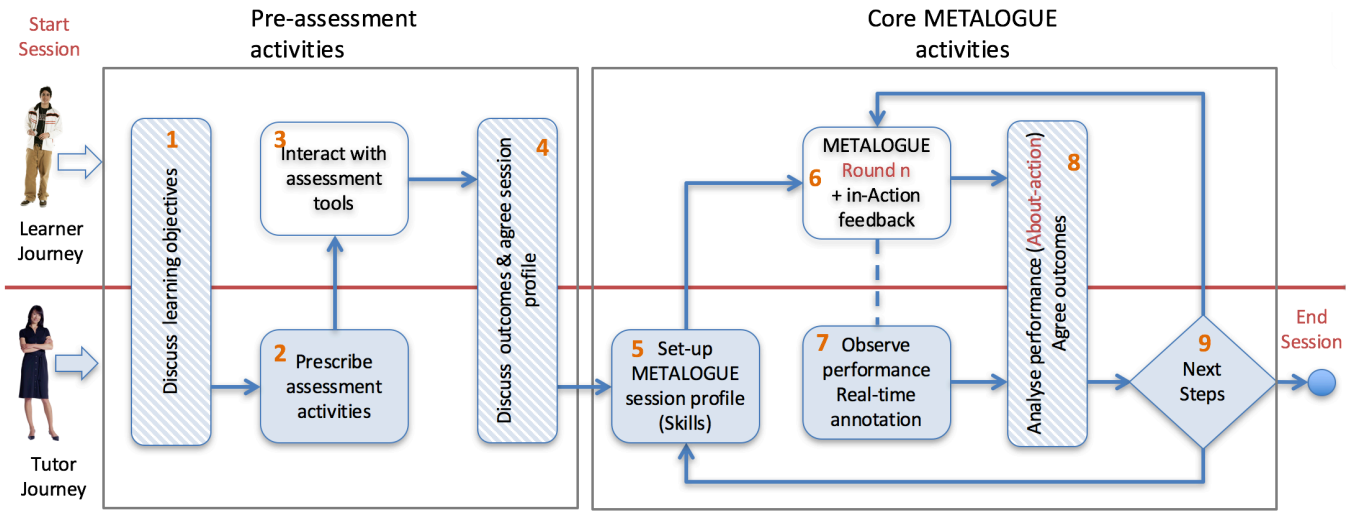

Figure 2: Training session flow.

4. PILOT SETUP

As part of an iterative development, three development phases were planned. At the end of each phase, a user evaluation validates the interaction and user experience, by evaluating the system via specific learning scenarios [10].

This approach concentrates on delivering incrementally improved prototypes for deployment in a consistent environment supporting English language for training metacognitive skills through debate coaching for the Hellenic Youth Parliament participants. Each pilot cycle provides an opportunity to enhance the stability, usability, and applicability of the system instantiation through the learning design validation, testing and evaluation.

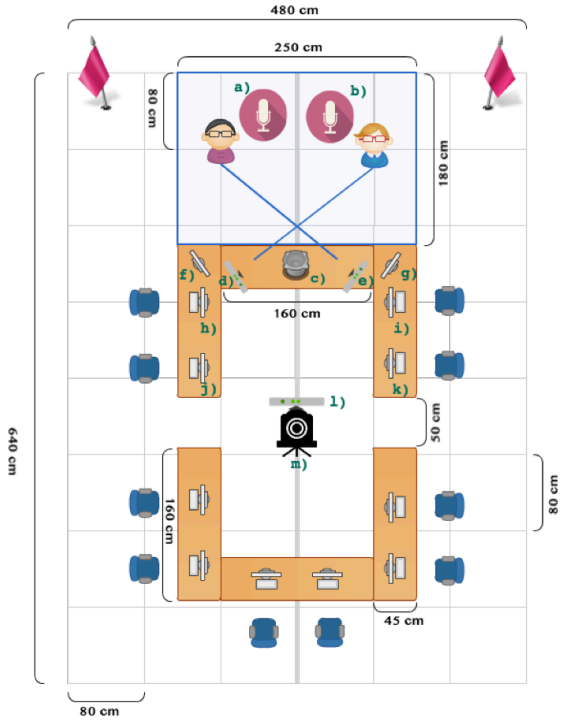

Figure 3. Pilot deployment setup for the Parliament sessions.

At the end of phase one, the system took the role of an observer between two human participants debating over an intensive political issue. Ten participants were asked to participate and debate in pairs. They were invited based on their experience gained from their participation to the annual Hellenic Youth Parliament sessions. Additionally, three tutors from the Hellenic Parliament Foundation were encouraged to observe the debate and provide suggestions for the system about-action component, the means to collaborative summative review of the sessions that the tutors and the trainees may use to reflect upon.

For this reason, a tutor and learner journey overview [12] has been elaborated (Figure 2). The diagram is organized in two horizontal swim lanes; above the central line are the activities (boxes) performed by the learner and below are those performed by the tutor. Both parties are involved in the pre-assessment activities. The tutor and the learner discuss the objectives of the session and, according to those, the tutor prescribes the assessment activities. The aim of those activities are to determine the learner level and construct or update their profile. After the learner finishes with the assessment activities, their profile is updated and the tutor may propose a training session with the system. This is done by selecting the skills that the learner will train and the system consequently selecting the appropriate training scenario.

For this pilot, the system was not used for the pre-assessment activities, in order to control the variables and concentrate on the system feedback to the learners. The users prepared themselves using the same debate scenarios. The scenarios included several aspects for both sides of the debate, and the participants' choice of case (for example, pro smoking banning – against smoking banning) was randomized.

The aim of the evaluation at the end of phase one was to provide several types of feedback to the users via the real-time in-action module and monitor the engagement and general user acceptance of the approach implemented for the initial prototype system. The setup layout is shown in Figure 3. The participants are standing face to face in front of microphones. One Kinect device is dedicated for each participant, monitoring full body and facial movement, and gestures. The system, via the devices, monitors the debate while providing personalized feedback for each participant, though normal displays.

5. EVALUATION

The sessions lasted between 10-15 minutes each. All participants, tutors and students, were debriefed and were asked to fill in a questionnaire on the general user experience. Further discussion followed collecting and analyzing newly found user needs, preferences, system shortcomings and expectations.

According to the users' perception the system should have a unique identity as a training/tutoring system that the users may interact with and train specific skills like civic skills for negotiation. The tutors asked for specific lists of skills that the system may be used to train and voices thoughts on scalability and capability to provide personalized sessions based on learner progress.

Civic skills training was proposed by the tutors as a challenging approach, since they need continuous coaching and refinement. Moreover, based on the system multimodal interaction capabilities, several choices for multimodal system output were indicated, such as real-time coaching tips and learner body posture mirroring via augmented reality visuals for body posture correction suggestions.

Several recommendations were presented for the about-action feedback environment. The evaluators proposed that it should be self-contained, providing easy to access (general consensus was for an iPad app) overview of the progress per skill and type of input, body posture and movement, voice sentiment, etc.

Portability of the system was a serious issue raised by the Parliament counsellors, since the final system should be easily deployable to both educational training establishments and private learning spaces.

6. CONCLUSION

The system architecture presented in this paper was the result of several incremental design and development iterations. The modularity and complexity of the integration were showcased. The system was used an observer in order for students to evaluate the general approach towards multimodal output.

Several shortcomings were identified that need to be addressed for the final system. At the end of the road, the system will be required to take the role of a participant and support natural free interaction with a human learner, with the system as a tutor. Human tutors observe learners and prompt specific aspects that need to be addressed. On the other hand, the system analyses many data streams, each providing an aspect of interaction (emotion, body language, and so on). However, this information needs to be fused into a single choice of system response, just as a human tutor would do.

The system may also be used as a collaborative environment for tutors and learners, instead of a self-sufficient tutor. Such approach requires the logging of data for several learners, enabling the human tutors and learners to revisit sessions and improve on specific skills.

Further work involves the implementation of the tutor and learner recommendations to result in an integrated environment for skill selection and training. Additional experimentation on the cognitive load from the system real-time feedback to the learner is warrantied in order to model the frequency and type of the visual feedback, as this is generated from the interpreted events.

7. ACKNOWLEDGMENTS

The work described in this paper has been partially funded by the European Union FP7 ICT program Metalogue, under grant agreement 611073. The authors would like to thank the Hellenic Parliament Foundation tutors and students that participated in the pilot experiments for their time and devotion to interact with the system under development and provide crucial feedback.

8. REFERENCES

1. Alexandersson, J., Girenko, A., Spiliotopoulos, D., Petukhova, V., Klakow, D., Koryzis, D., Taatgen, N., Specht, M., Campbell, N., Aretoulaki, M., Stricker, A., and Gardner, M. 2014. Metalogue: A Multiperspective Multimodal Dialogue System with Metacognitive Abilities for Highly Adaptive and Flexible Dialogue Management, *10th Int. Conf. Intelligent Environments (IE'14)*, 2-4 July 2014, Shanghai, China.
2. Cavazza, M., de la Camara, R. S., and Turunen, M. 2010. How was your day? A Companion ECA. *Proc. 9th Int. Conf. on Autonomous Agents and Multiagent Systems*, pp. 1629–1630.
3. De Silva, L. C., Morikawa, C., and Petra, I. M. 2012. State of the art of smart homes. *Engineering Applications of Artificial Intelligence*, 25(7), 1313–1321.
4. Gonzalez, C., and Lebiere, C. 2005. Instance-based cognitive models of decision making. In D. Zizzo and A. Courakis, editors, *Transfer of knowledge in economic decision making*. Macmillan.
5. Griol, D., Callejas, Z., López-Cózar, R., and Riccardi, G. 2014. A domain-independent statistical methodology for dialog management in spoken dialog systems. *Computer Speech and Language*, 28(3), 743–768.
6. Heinroth, T., and Minker, W. 2013. Introducing spoken dialogue systems into Intelligent Environments. New York: Springer.
7. Herzog, G., and Reithinger, N. 2006. The SmartKom Architecture: A Framework for Multimodal Dialogue Systems. *SmartKom: Foundations of Dialogue Systems*, 55–70. doi:10.1007/3-540-36678-4.
8. Minker, W., Haiber, U., Heisterkamp, P., and Scheible, S. 2004. The SENECA spoken language dialogue system. *Speech Communication*, 43(1–2), 89–102.
9. Schneider, J., Boerner, D., Van Rosmalen, P., and Specht, M. 2015. Augmenting the senses: a review on sensor-based learning support. *Sensors*, 15(2), 4097-4133.
10. Spiliotopoulos, D., Petukhova, V., Koryzis, D., and Aretoulaki, M. 2014. Pilot Scenario Design for Evaluating a Metacognitive Skills Learning Dialogue System, *Proc. Int. Conf. in Human-Computer Interaction*, Los Angeles, USA, CCIS 435: 162-166, Springer-Verlag Berlin Heidelberg.
11. Stevens CA, Taatgen NA, and Cnossen F. 2016. Instance-Based Models of Metacognition in the Prisoner's Dilemma. *Topics in Cognitive Science*. 8(1): 322-334
12. Van Rosmalen, P., Boerner, D., Schneider, J., Petukhova, V., and Van Helvert, J. 2015. Feedback design in multimodal dialogue systems. *Proc. 7th Int. Conf. on Computer Supported Education*, Volume 2, pp. 209–217, Lisbon, Portugal.