Workshop Proceedings of the 11th International Conference on Intelligent Environments
D. Preuveneers (Ed.)

doi:10.3233/978-1-61499-530-2-220

Observing, Coaching and Reflecting: A Multi-modal Natural Language-based Dialogue System in a Learning Context

Joy VAN HELVERT[a,1], Peter VAN ROSMALEN[b], Dirk BÖRNER[b],

Volha PETUKHOVA[c] and Jan ALEXANDERSSON[d]

[a] *University of Essex*
[b] *Open University of the Netherlands*
[c] *Saarland University*
[d] *German Research Centre for Artificial Intelligence GmbH*

Abstract. The Metalogue project aims to develop a multi-modal, multi-party dialogue system with metacognitive abilities that will advance our understanding of natural conversational human-machine interaction and dialogue interfaces. This paper introduces the vision for the system and discusses its application in the context of debate skills training where it has the potential to provide learners with a rich, immersive experience. In particular, it considers a potentially powerful learning analytics tool in the form of a performance reflection dashboard.

Keywords. Natural conversational interaction, mixed-reality, multi-modal dialogue systems, immersive, debate skills, learning analytics, reflection

1. Introduction

As we move towards a world of smart and immersive environments we are seeking new ways of interfacing and engaging with our technologies that more closely reflect natural human interaction. Human to human communication embodies multiple modalities such as speech, gesture, facial expressions, gaze, and body posture, so it follows there is an inherent desire to communicate with our technologies in the same way. This aspiration was illustrated recently in Spike Jonze' film 'Her' [1] in which a writer encounters 'Samantha'; a multi-modal dialogue system capable of understanding, expressing, and responding to emotion to the extent that it was possible for him to fall in love. While at present this level of immersion and engagement remains in the realms of science fiction, one of the aims of Metalogue, a three-year EU project, is to push the existing boundaries further in this direction.

So where to start? Absolutely free natural interaction is clearly not feasible at this point; however, educational dialogues and tutoring interventions offer some valuable

[1] Corresponding Author: Joy van Helvert.

constraints. The application of multi-modal, natural language interfaces in this domain is already being explored, for example Nye et al [2] text and speech, Johnson and Valente [3] discourse with "virtual humans", and Yang et al [4] use of motion sensors; however, increasing computing power offers the possibility of combining a range of traditional and new modalities in a single dialogue system. In this paper we will outline the proposed Metalogue multi-modal dialogue system and a potential application in the domain of debate skills training. It has been shown that digital immersion can enhance learning in three ways: by providing (1) multiple perspectives; (2) situated learning; and (3) transfer [5,6,7]. While the vision for the Metalogue debate skills trainer may not be considered "digitally immersive" in the traditional sense, it can be said to offer an augmented reality that would incorporate these three types of experience.

We start with an outline of the physical system and details of its composition, we then describe its implementation in the form of a debate skills trainer and discuss, in detail, learning analytics and the design for reflective feedback. To close we review Metalogue's immersive potential, and finally layout plans for the implementation and evaluation of the vision.

2. Metalogue: A Multi-Modal, Multi-Party Dialogue System

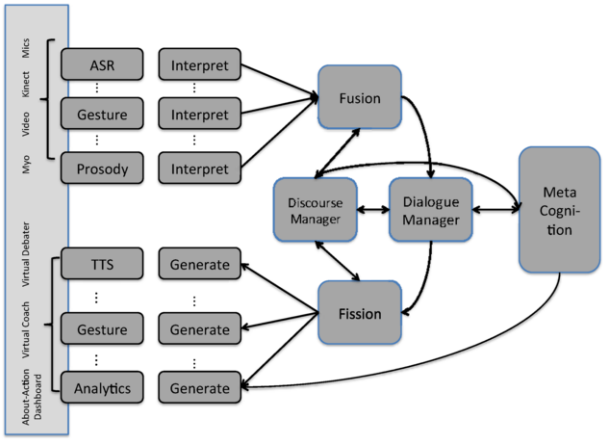

Figure 1. Metalogue components and workflow.

Metalogue is designed to "hear" and "see" a wide range of human interaction signals, and to "interpret", and "respond" in as natural way as possible, either in a coaching capacity or as a fully-fledged partner in the interaction. Figure 1. depicts the integrated components of the system and its processing workflow. Metalogue gathers three types of sensor specific data [8] that serve as input: (1) speech signals from multiple sources (wearable microphones and headsets for each dialogue participant and an all-around microphone placed between participants); (2) visible movements tracking signals from Microsoft Kinect and Myo sensors capturing body movements and facial expressions; and (3) video signals captured by the camera and recording the whole dialogue including sound.

The speech signals serve as input for two types of further processing: (1) Automatic Speech Recognition (ASR)[2], leading to lexical, syntactic and semantic analysis and ultimately updating the discourse model to answer the question 'what was said?, and (2) prosodic analysis[3] which is concerned with rhythm, stress and intonation of speech and answers the question 'how was it said?' This enables the system to interpret elements such as speech rate (fast, slow, adequate tempo), volume (loud, soft, adequate), emphasis (flat intonation, uneven/unbalanced intonation etc.), and pausing (too long or not enough). Analysis of visible movements gathered by Microsoft Kinect and Myo sensors, enable the system to interpret input related to gaze (re)direction, head movement and head orientation, facial expressions, hand and arm gestures, posture shifts, and body orientation. These outputs are further analyzed to determine factors such as emotional state and, ultimately, argument content, organization and delivery.

At the heart of the system, interpretation and the generation of output depend on advanced linguistic multi-modal, multi-party and multi-perspective discourse models. These combine both social and linguistic signal information and are generated by collecting and annotating a corpus of human-human interaction data. This, in turn, is used to train machine-learning algorithms for the automatic recognition and prediction of a wide range of human interaction phenomena. In addition, the system incorporates metacognitive models that explain metacognition as a set of skills, a cognitive agent that exhibits metacognitive behavior similar to humans, and a learner model able to assess the users' metacognitive skills. Together these aspects of Metalogue enable the system to critically analyse participants' interactions within a certain time frame and generate "events" (i.e. points highlighting areas where performance could be improved and positive interaction behaviors that can be built upon). These are recorded in the form of annotations to the video file.

Output to the user can take three forms; (1) in-performance coaching; (2) post-performance reflective analysis; and (3) an optional debate or negotiation partner in the form of a virtual character that simulates a wide range of both verbal and non-verbal language attributes.

3. Debate Skills Training:

To debate successfully the student must master a range of metacognitive skills [9, 10], such as monitoring and adjusting verbal performance, eye contact, body posture and gestures, also they must know when and how to employ appropriate strategies and arguments to achieve certain goals while at the same time recognising and responding to the oppositions' strategies and arguments. Debate skills training typically involves ad-hoc face-to-face classroom debates combined with more formal organised

[2] The Metalogue ASR system is developed based on the Kaldi (http://kaldi.sourceforge.net/about.html) open-source speech recognition toolkit. The current trained system after augmenting general language models (trained on Wall Street Journal corpus) with a Speaker Adaptive Training using collected Metalogue data lead to a word error rate (WER) of approximately 32%.

[3] OpenSmile audio analysis tool has been used: http://records.sigmm.ndlab.net/2015/01/opensmile-the-munich-open-source-large-scale-multimedia-feature-extractor-a-tutorial-for-version-2-1/

competitions. While the learner gains confidence through these performances, feedback from tutor or panel on such complex combined physical and mental skills played out over a period of time, depends on subjective human judgment and may reflect a summary of the interaction rather than pinpoint specific behaviours that can be improved.

Implemented to support debate-skills training within a 4C-ID pedagogic framework [11], Metalogue allows the learner to debate with another human being or with a virtual debate partner and be subject to multi-channel digital "observation". It further enhances the learner's debate experience by providing real-time feedback currently envisaged in the form of signs and symbols displayed on a screen [12] indicating simple posture, volume or tone adjustments to improve performance. A wide range of performance analysis is constantly being generated by the system but real-time feedback is carefully balanced to avoid cognitive overload and/or disengagement whilst the learner is performing. The full range however can be accessed via a reflection dashboard that enables both tutor and learner to review and analyse the performance moment-by-moment. Metalogue will support the debate skills training with a consistent feedback loop, i.e. real-time feedback to raise awareness of currently trained aspects/behaviours and about-action feedback to trigger reflection on the previous training sessions and prepare for the following training sessions.

4. Learning Analytics: About-action Feedback

The term Learning Analytics generally refers to the large-scale measurement, collection and analysis of learner data across different systems within learning organisations [13]. Here we use the term to refer to the measurement, collection and analysis of Metalogue generated learner data relating to the individual learner, including comparisons with aggregated learner data from within the system. This type of reflection and analysis can support educational stakeholders in becoming "aware" of their actions and learning processes. Endsley [14,15] described being "aware" as a three level process consisting of the perception of elements in the current situation, the comprehension of the current situations and the projection of a future status. These three steps are seen as a prerequisite for making decisions and effectively performing tasks. Once people are aware of their situation, they can reflect on their actions, choose to adapt their behavior if necessary, and engage in a process of continuous learning [16].

Schön [16] defines reflective practice as the practice by which professionals become aware of their implicit knowledge base and learn from their experience. He uses the terms reflection-in-action (reflection on behaviour as it happens, so as to optimize the immediately following action), and reflection-about-action (reflection after the event, to review, analyse, and evaluate the situation, so as to gain insight for improved practice in future). In the context of the Metalogue project we refer to the system generated coaching feedback delivered to the learner during their debate performance as the in-action feedback, and the post performance reflection and analysis capabilities of the system as the about-action feedback. The in-action capabilities of the system have been outlined above; however, here we discuss our vision for the about-action

reflection dashboard based on two user scenarios: (1) immediate post performance review by tutor and learner together; (2) tutor or learner individually reviewing at leisure.

There are a wide variety of software offering tools available to analyse and/or visualise existing data [17]. However, of particular interest in the context of the Metalogue project is the use of visualisation tools as demonstrated by the Flashmeeting project [18]. Although the application has been developed to support online meetings, the analysis tools provide a useful illustration of how a multimodal system data and analysis could be organized in the form of a reflection dashboard. For example, it is possible to replay the complete interaction, visualising the actual video replay as well as the broadcasting distribution over time (i.e. who spoke when and for how long), and more detailed information such as chat events, specific content annotations, and broadcasting events such as interruptions. It is also possible to view analyses such as broadcast dominance (i.e. the ratio of contribution by the different participants) in the form of a pie chart, and analysis of the interaction content in the form of a key word cloud.

The key criteria for the visualization of Metalogue data are:

- Occurrences of an event (e.g. voice volume (too high, too low etc), confident posture) on a timeline
- Aggregation of a single event (e.g. time used)
- Occurrences of a number of events in relation to each other in time.
- Integrated overview of various events.

In addition, parts of the dashboard may need to be layered, particularly when an integrated event is shown, and the learner or tutor should be able to zoom-in into the underlying aspects for further clarification.

5. The Metalogue About-action Dashboard: An Outline Design

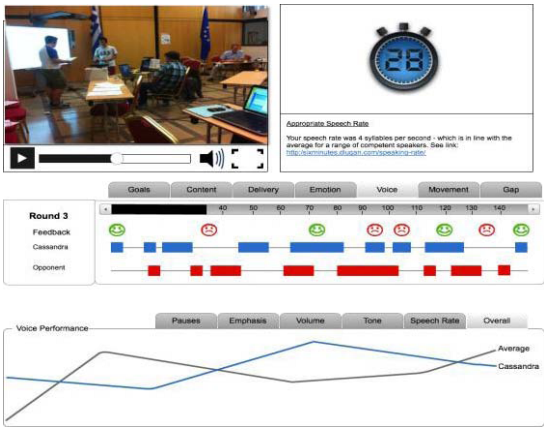

Figure 2. Metalogue About-action Dashboard screen mock-up.

An example of the type of about-action analysis the Metalogue system can provide the tutor and learner is shown as a screen mock-up in Figure 2. There are a wide range of options for selecting and viewing the whole or segments of the learner's performance during a debate round. Video material appears in the top left window with standard video controls immediately beneath. Below (central) is the timeline adapted from the Flashmeeting dashboard as discussed above. This shows the utterances of both the learner and her opponent, plus the Metalogue feedback events against the timeline of the video. Clicking on an utterance block or an event symbol will display the corresponding video segment in the top left window. Similarly, clicking a particular point on the timeline will locate that point in the video.

The tabs along the top of the timeline window allow the user to view different types of event symbol; for example, Figure 2 (central) shows the voice tab has been selected. Accordingly, the symbols along the timeline represent all the Metalogue feedback events relating to voice performance. As the video plays, the top right window displays in detail any feedback events located on the timeline, this includes feedback given in-action, i.e. during performance, and all other events detected by Metalogue during performance but not displayed at the time to avoid overloading the learner. In Figure 2 the feedback is shown as a stopwatch symbol, providing positive feedback to the learner that her speech rate at this point was at an ideal level. It also provides clarification of what the symbol represents and provides links for further exploration.

The lower window is intended to provide various kinds of analysis depending on the tab selected on the timeline window above. In this case, with the voice tab selected (Figure 2, central), the available analysis options for voice are displayed along the top of the lower analysis window i.e. pause, emphasis, volume etc. The 'Overall' tab is shown as selected (lower-right) and the window displays an analysis of the learner's voice performance for the round against the average performance of other learners training with the same game parameters.

6. An Immersive Experience?

Debate, whether human-human or human–virtual human, is by its very nature immersive; however, returning to the "digital immersion" criteria [5,6,7] mentioned in the introduction to this paper (i.e. situated learning, multiple perspectives, and transfer), we will review the proposed Metalogue functionality.

With regard to situated learning, the debate trainer is being designed to support the 4C-ID pedagogic framework [11] which mandates attention to authentic whole tasks based on real life, organized in classes with variation and increasing complexity. The leaner experience is also dynamic and engaging, with the system taking the role of observer/audience and coach. In addition, it offers the option of a virtual debate partner able to employ natural language interaction and different styles of delivery (e.g. aggressive, conciliatory etc.).

Both in-action and about-action feedback offer the learner multiple alternative perspectives on their performance enabling them to become aware of certain behaviours and make choices about how to respond when necessary. For example, in-action: resetting body posture that has become inappropriate or gaze that has become averted; about-action: recognizing and understanding debate strategies and how to employ them.

Finally, in terms of transferability, Metalogue is a mixed-reality system therefore the context of the learner is never entirely removed from the real world. Also the simulations involve realistic debating scenarios that allow the tutor to determine the topic and set the extent of the challenge, thus learning outcomes are envisaged to be highly transferrable.

7. Vision Into Reality

The Metalogue project will be realised in three incremental prototype development cycles culminating in the full functionality outlined in this paper. The first prototype has now been integrated and is currently able to hear and observe learner interactions and provides real-time feedback on posture, hand/arm gestures, and voice volume. This is the first step proving the basic Metalogue concepts from a technical perspective. Each pilot will be comprehensively evaluated with "real learners" in the form of debate students from the Hellenic Youth Parliament. It will involve the technical evaluation of the multi-modal dialogue system (employing user-based and non user-based techniques), user evaluation and satisfaction measurements, and learning effectiveness measurements. Alongside the application of Metalogue as a debate skills training system, the project will pursue its potential application in a call-center environment supporting agent training, and furthermore, investigate its portability into different languages.

The vision elaborated above is a challenging, multi-disciplinary endeavor and a work in progress. However, the outcomes have the potential to advance our understanding of debate training and learning analytics on one level, and conversational human-machine interaction, dialogue interfaces on another – perhaps moving us one small step towards the visionary capabilities embodied in Jonze' Her [1].

Acknowledgements

The authors would like to thank all Metalogue staff who contributed in word and writing and in many discussions. The underlying research project is partly funded by the Metalogue project; a Seventh Framework Programme collaboration funded by the European Commission, grant agreement number: 611073 (http://www.metalogue.eu).

References

[1] S. Jonze (Dir.), *Her,* Elite Films, Zurich, 2014

[2] B.D. Nye, A.C. Graesser, X. Hu, Auto Tutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence Education* **24** (2014), 427-469.
[3] W.L. Johnson, A. Valente, Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures, *AI Magazine,* **30 (2)** (2009) 72-83
[4] H. Yang, H. Zhang, W. Xu, P. Zhang, L. Xu, The Application of KINECT Motion Sensing Technology in Game-Oriented Study. *International Journal Of Emerging Technologies In Learning (IJET),* **9(2),** (2014), 59-63. doi:http://dx.doi.org/10.3991/ijet.v9i2.3282
[5] C. Dede, Interfaces for Engagement and Learning. *Science – New Series,* **323** (2009), 66-69
[6] W. Sadowski, K.M. Stanney, Presence in Virtual Environments. In K.S. Hale and K.M. Stanney (Eds.) *Handbook of Virtual Environments*, Erlbaum, Mahwah, NJ, (2002) 791-806
[7] J. Lessiter, J. Freeman, E. Keogh, J. Davidoff, A Cross-media Presence Questionnaire: The ITC Sense of Presence Inventory. *Presence Teleoperators and Virtual Environments,* **10 (3)**, (2001) 282-297
[8] J. Schneider, D. Börner, P. van Rosmalen, M. Specht, Augmenting the senses: A review on sensor-based learning support, *Sensors* **15(2)** (2015) 4097-4133; doi:10.3390/s150204097
[9] B. Lybbert. "What Should be the Goals of High School Debate?" *Paper presented at the National Forensic League Conference on the State of Debate, Kansas City, MO,* (1985) 9p
[10] N.R. Trumposky, "The Debate Debate", *The Clearing House* **78** (2004), 42-55
[11] J.J.G. van Merriënboer, L. Kester, Whole task models in education. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer and M. P. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology*. New York: Routledge, Taylor & Francis Group (2008) 441-456
[12] J. Schneider, D. Börner, P. van Rosmalen, M. Specht, Presentation Trainer: A Toolkit for Learning Non-verbal Public Speaking Skills. In C. Rensing, S. De Freitas, T. Ley, P.J. Munoz-Merino (Eds.) *Proceedings of the 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Graz, Austria, September 16-19, 2014. Open Learning and Teaching in Educational Communities, Lecture Notes in Computer Science*, **8719,** Springer International Publishing (2014) 522-525..
[13] G. Siemens. Learning Analytics: Envisioning a Research Discipline and a Domain of Practice, *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*, (2012) 4-8
[14] M.R. Endsley, D.J. Garland, Theoretical Underpinnings of Situation Awareness: A Critical Review. In M.R. Endsley & D.J. Garland (Eds) *Situation Awareness Analysis and Measurement,* Mahwah, NJ: Lawrence Erlbaum Associates (2000) 3-28
[15] M.R. Endsley, Measurement of Situation Awareness in Dynamic Systems, *Human Factors* **37** (1995) 65-84
[16] D.A. Schön, *The Reflective Practitioner: How Professionals Think in Action*, Basic Books, (1983) 374p
[17] W. Kraan, D. Sherlock, Infrastructure and Tools for Analytics, *CETIS Analytics Series* **1 (11)** (2013) 22p
[18] P.J. Scott, E. Tomadaki, K.A. Quick, The Shape of Live Online Meetings, *International Journal of Technology, Knowledge and Society* **3** (2007) 1-16