From Talking to Thinking Heads: Report 2008

Burnham, D. [1] *Abrahamyan, A.* [1] *Cavedon, L.* [2] *Davis, C.* [1] *Hodgins, A.* [1] *Kim, J.* [1] *Kroos, C.* [1] *Kuratate, T.* [1] *Lewis, T.* [3] *Luerssen, M.* [3] *Paine, G. Powers, D.* [3] *Riley, M.* [1] *Stelarc*[1], *Stevens, K.* [1]

[1]MARCS Auditory Laboratories, University of Western Sydney, Sydney, Australia
[2]School of Computer Science and IT, RMIT University, Melbourne, Australia
[3]School of Computer Science, Engineering, Mathematics, Flinders University, Adelaide, Australia

{D.Burnham,A.Abrahamyan,Chris.Davis,A.Hodgins,J.Kim,C.Kroos,T.Kuratate,GA.Paine, KJ.Stevens}@uws.edu.au; marciariley@optusnet.com.au; lcavedon@cs.rmit.edu.au; {David.Powers,Trent.Lewis,Martin.Luerssen}@flinders.edu.au; stelarc@va.com.au

Abstract

The Thinking Head project has as it aims to develop (i) a new generation Talking *Thinking* Head that embodies human attributes, and improves human-machine interaction; and (ii) a plug-and-play research platform for users to test software in an interactive real-time environment. Here, project progress is discussed in terms of the four teams: 1. Head Building – (i) Plug-and-Play architecture, (ii) Thinking Media Framework, and (iii) Animation; 2. Human-Head Interaction (HHI) – (i) Wizard of Oz studies, and (ii) joint attention by human and head; 3. Evaluation; and 4. Performance in (i) the Beijing Head and (ii) the Pedestal Head. Directions for future research are outlined as appropriate.

Index Terms: talking heads, human-computer interaction, evaluation, performance

1. Introduction

The Thinking Head project [1] originated from the interaction of performance artist Stelarc [2] and his 'Prosthetic Head', with a number of cognitive, computer, speech and language scientists. Building on the Prosthetic Head, the Thinking Head project seeks to establish:

1. a new generation Talking *Thinking* Head that embodies human attributes and improves human-machine interaction
2. a plug-and-play research platform for users to test software in an interactive real-time environment.

Progress towards these aims thus far are set out below in terms of the four teams: Head Building, Human-Head Interaction, Evaluation, and Performance.

2. Team Advances

2.1 Head Building Team

The Thinking Head is being developed using a combination of existing and new components, e.g., audio-visual speech processing, multimodal interpretation, dialog management, language generation, speech synthesis, and visual rendering. Novel research is focused on such issues as adaptive conversational abilities, emotion recognition and rendering, using visual and non-verbal cues for interaction and dialog management, and audio-visual processing and integration. Two particular aspects of the head building will be discussed here, the plug-and-play architecture and the thinking media framework.

2.1.1 Plug-and-Play Architecture

A specific focus of the project is to develop a software architecture that allows external groups to integrate components, replacing existing Thinking Head capabilities or contributing new ones. As in other dialogue-system platforms [3] we use an event-driven framework, which has a number of desirable properties, such as: naturally modeling the non-linear nature of human interaction; providing the flexibility required for easy integration of components into a distributed architecture; dynamically prioritising software components and event types; and optimizing the system, via inter-component configuration commands for particular interaction states.

The event framework supports components written in multiple languages running on diverse software platforms, e.g., as done in the Open Agent Architecture [4]. The framework also allows multiple versions of similar-type components, with a policy for selecting contributions from components to be specified. For example, the system may contain 2 dialogue managers, with a "dialogue event" being sent to both, with each dialogue manager processing that event and suggesting a response. The selection policy chooses amongst the responses. We are currently using this to select between suggested responses from a sophisticated domain-specific dialogue manager and a wide-coverage, but less sophisticated, one.

2.1.2 Thinking Media Framework

Another challenge, not typically addressed in other frameworks, is handling enormous volumes of continuous video and audio stream data. Some components, in particular the audio- and visual- processing components, are naturally more tightly coupled than others. The multiple associated data streams are captured in parallel and need to be accessed by a diverse set of TH components: events corresponding to parallel-processed streams (e.g., a multimodal communication) need to be forwarded to other TH components, as well as those gleaned from individual streams (e.g., a face-recognition or speech-start event).

The Thinking Media Framework (TMF) constitutes a uniform platform for TH components to share these resources. It supports concurrent processing, buffering, and archiving of

streams, establishes inter-stream synchronization and allows for joint control of media sources. By seamlessly tying into higher-level media processing libraries such as OpenCV and CMU Sphinx, TMF keeps out of sight of the developer, while facilitating the independent development of perceptual subsystems for the Thinking Head. The TMF thus provides a mechanism for more tightly integrating some components than others in the overall framework. TMF will be available for multiple operating platforms and is easily extensible to new stream types.

2.1.3 Animation

The current animation component works as a text-to-AV synthesis system: it receives text data intended as speech for the animated face, and generates the speech and corresponding face motion as output. The system consists of four major parts: a TTS module; a phoneme-to-face motion database; a phoneme-to-face animation generator; and a face animation module. The key to this system is the development and use of the newly built phoneme-to-face motion database. This database, as well as the TTS module, is language dependent. Thus far we have developed a successful prototype using Japanese from a Japanese male subject, and are now building an expanded version for an Australian English male voice. We selected the Festival Speech Synthesis System [5] as the English TTS module because of its usability, and for visualization we selected a real-time Talking Head animation system with condensed Principal Component (PC) parameterization based on our previous research. Our animation system supports a large number of realistic face models (currently 200) synthesized from our 3D face static database [6]. Models can also be easily built from 3D reconstruction of photographs, and simplified cartoon-like models can also be accommodated. In the interest of adaptability, the TTS and animation modules can be replaced, although extra processing may be needed to generalise face motion to new visualization parameters. Our current system begins with neutral speech, with planned expansions to include emotional speech and expressions, and non-verbal communication cues.

2.2 Human-Head Interaction (HHI) Team

Two streams are currently being pursued by the HHI Team as set out below.

2.2.1 Wizard of Oz Studies

First, we are laying the foundations for high-level HHI strategies by simulating advanced capabilities of the Thinking Head in *Wizard of Oz* (WoZ) experiments, with an interface that allows the experimenter to drive the Thinking Head's behaviour in a flexible but structured manner. Three video streams, the acoustic signal, and gaze tracking data are recorded to examine participants' responses when they believe they are interacting with an autonomous talking head.

The WoZ setting consists of an interactive customer complaint scenario in which the human participant has been given a specific complaint and goal. This situation tailors the dialogue into a means-ends framework, the success of which can be easily evaluated. The scenario also has been customized to elicit a set of fixed responses from the participant that thus enables the collection of standard data from a range of participants. Furthermore, the collection of human-head dialogue permits the analysis of dialogue structure and the specification of 'potential' out of vocabulary items (for an actual TH).

2.2.2 Joint Attention by Human and Head

Second, we excise aspects of complex HHI, e.g., temporal gaze coordination in joint attention tasks, to study the influence of embodiment variations in communicative behavior. To do this, we combine an open-source game engine, the commercial face animation software 'FaceRobot', and the Thinking Head's own graphics system, in order to control how the Thinking Head is represented - from a 'blob with eyes' to state-of-art detailed rendering of the full face using normal maps.

In order to examine a coordinated exchange between the human and TH, we have developed a set of interactive cooperative tasks, the successful negotiation of which will be used to evaluate the relative effectiveness of Thinking Head expressive characteristics. Task success and on-line performance will be used to also gauge the impact of the pairing-down of specific Thinking Head features (such as structural features - texture maps, and functions such as the accuracy of eye movements).

3.3. Evaluation Team

3.3.1 Introduction

The efficacy of emotional expressions in the current Head [7, see 2.4] has been evaluated with undergraduate participants who were presented with two computer-driven Thinking Head monologues ('Edwin Hubble', 'Machu Picchu').

The aim of the main experiment was to investigate the effect of the Head's facial expressive gestures, such as smiling, winking, rolling eyes, etc, when reciting a text on direct and indirect measures of usability and intelligibility. Rather than explicitly asking participants whether the Head's communication was clearer when expression was present, we used a simple experimental design in which the effect of the manipulation of an independent variable (IV) on two types of dependent variables (DVs) was measured. The IV was the presence vs the absence of emotional expression (counterbalanced across participants and monologues). With respect to the DVs, both indirect and direct measures were employed. The first DV was accuracy on a 6-item comprehension task with the items relating to the 3-minute text that participants watched the Head recite. The assumption in using this indirect DV is that if expression aids communication, intelligibility and/or engagement, then comprehension performance should reflect that enhancement relative to a condition where there is little or no facial expression. The second DV, a more explicit and direct measure, involved participants rating five qualities of the Head: i) Likeable; ii) Engaging; iii) Easy to understand; iv) Life-like; and v) Humorous. At the end of a session, participants assigned ratings across the session to the following statements: i) The Head kept my attention; ii) I would like to interact with The Head again; iii) I enjoyed interacting with The Head; iv) I felt as if The Head was speaking just to me.

It was predicted that if facial expressive gestures aid intelligibility and communication then comprehension accuracy would be greater in response to the expressive than the inexpressive Head. If it is the case that facial expression increases engagement but greater engagement distracts users

from the text, then we would expect a negative correlation between comprehension accuracy and ratings of engagement (i.e., a tendency for lower comprehension scores to be associated with higher engagement scores, and vice versa).

3.3.2 Method and Results

Forty adult participants completed the main experiment. Each was presented with the two monologues recited by the Head with one version of each text accompanied by facial expressions and the other with little or no facial expression.

There was no evidence of a negative correlation between comprehension and engagement, thus suggesting that greater engagement does not necessarily distract the user from the meaning in the text

With that in mind we move to the comprehension scores, the means and standard errors of which are shown in Table 1.

Table 1. *Mean Comprehension Accuracy (as proportions).*

	Hubble		**Machu**	
	Mean	*SE*	**Mean**	*SE*
Expressive	0.72	0.03	0.52	0.04
Neutral	0.58	0.05	0.54	0.06

There was significantly better comprehension for the Expressive Hubble than the Neutral Hubble text, $t(19)=2.22$, $p=0.04$, but not for the Expressive Machu than the Neutral Machu texts, $t(19)=0.364$, $p=.72$. Thus, there is partial support for the notion that comprehension improves when greater expression is provided. This differential effect is illuminated by the fact that there is better comprehension for the Expressive Hubble than the Expressive Machu texts, $t(19)=4.17$, $p=.001$, but no difference for their neutral counterparts, $t(19)=0.55$, $p=0.59$, thus suggesting that the nature of the mark-up may have affected comprehension.

To investigate text effects more closely, an independent text evaluation task was conducted in which we compared the effect of the two different texts only recited (i.e., no visual information) on comprehension and ratings. There was no significant difference between the read-aloud Hubble and Machu texts on comprehension (means of 0.75 and 0.70, respectively). The only rating that differed across the texts was that the Hubble text was rated as significantly more positive (mean=4.00) than the Machu text (mean=3.40), $t(9)=2.71$, $p=.02$. Again, it appears that that the nature of the mark-up may have affected comprehension.

Analysis of the participants' ratings of the Head showed that the expressive version of the Machu text was perceived to be significantly more humorous than the corresponding Neutral text, $t(19)=3.71$, $p=0.001$. There was also a tendency for the Expressive Machu text to result in the Head being more likeable and engaging than the neutral version of the Machu text (uncorrected $p=0.04$). One possibility is that being overtly expressive, and in particular trying to be funny, or being seen to make fun of the text, detracts from comprehension. To tease these out in future experiments it is proposed to have separate ratings such as "witty" and "silly".

User Tracking: A second batch of experiments was performed with a further thirty-two participants. However, in this case the Head's face tracking was switched on so that the Head continually tracked the user (facing and directing eyes at the participant). Means and standard errors are shown in Table 2.

Table 2. *Mean Comprehension Accuracy (Tracking ON).*

	Hubble		**Machu**	
	Mean	*SE*	**Mean**	*SE*
Expressive	0.58	0.06	0.47	0.05
Neutral	0.60	0.05	0.60	0.05

In this case significant increases in Humorous ratings were seen for the Expressive vs the Neutral rendering of both texts (Hubble, p=0.004; Machu, p=0.072), and for the Machu, but not the Hubble text, there was a slight but non-significant reduction in average comprehension (p=0.086) with Expression vs Neutral, providing some limited support for the possibility of a negative relationship between humour and comprehension.

3.3.3 Conclusions

Together the results from the main experiments and two subsequent experiments show that the differences observed in the main evaluation experiment are influenced by the expressive markup, possibly interacting with text nuance. A possible avenue of exploration in this regard is the relative timing of the text and the emotional mark-up; simultaneous presentation may not in fact be the best way to convey emotion to the human interlocutor. In addition, there is some indication that head-humour may negatively influence the degree of comprehension in the human. These issues beg further investigation by the HHI and the Evaluation teams in studies to manipulate expressive markup and the degree of humour systematically as independent variables in future experiments [8].

2.4 Performance Team

There are two areas of importance to report with respect to the performance area, one retrospective, and one prospective.

2.4.1 The Beijing Head

In Beijing, from June 10 to July 3, 2008 the Prosthetic Head Zero+ and the Walking Head [9], a Prosthetic Head mounted on a walking frame, accompanied Stelarc to participate in the New Media Arts Exhibition as part of 'Synthetic Times: Media Art China 2008', under Artistic Director, Zhang Ga in China, and Kim Machan, the Director of MAAP (Multimedia Art Asia Pacific) at the National Art Museum of China (NAMOC). The exhibition, occupying about 7,500m^2 showcased both established and emerging artists from around 27 countries, over 50 seminal and current media art works will be on view along with performances, workshops and symposia.

For this exhibition, the latest version of the Thinking Head, Prosthetic Head Zero+, was taught a wide range of facts about Beijing, China, and the Olympic Games. The Prosthetic Head and the Walking Head were installed in the first room of the exhibition after the entrance lobby. The Prosthetic Head was projected at over 4 m in size, so the scale was impressive. During the opening and the first few days of the exhibition there was, every now and then, good interaction with the Head. Interaction was successful when the person was a good typist and when their English was adequate, so every now and again a group of people were obviously enjoying the Head's responses, they were being engaged by the head.

2.4.2 The Pedestal Head

Building on the Walking Head [9] we are developing a Pedestal Head in which the TH will be displayed on a monitor affixed to a robot arm allowing motion in all 3 axes. Combined with the development of an auditory localization system to control gross movements of the monitor-mounted Thinking Head in response to the physical location of sound, e.g., the interlocutors' speech, and gaze tracking for fine-grained responses to the interlocutors' face, the Pedestal Head should provide a compelling and engaging presence.

Conclusions and an Invitation

The Thinking Head project is progressing via the simultaneous work of four research and development teams. Integration of the output of these teams will result in iterative improvements in the Thinking Head. We also seek and encourage the incorporation of ideas and components from other researchers from across disciplines and continents. With the plug-and-play architecture, this is now becoming possible in the Head Building area, and we also encourage this in the HHI, Evaluation, and Performance areas.

Acknowledgements

We greatly appreciate steering committee (R.Dale, & M. Wagner plus authors Burnham, Cavedon, Hodgins, Powers) input; and Chinese informants' (Shari Li, Hongjin He & Fang Qian) input to the Beijing Head.

References

[1] Burnham, D., Dale, R., Stevens, K., Powers, D., Davis, C., Buchholz, J., Kuratate, K., Kim, J., Paine, G., Kitamura, C., Wagner, M., Möller, S., Black, A., Schultz, T., & Bothe, H. From Talking Heads to Thinking Heads: A Research Platform for Human Communication Science. (Australian Research Council/National Health & Medical Research Council Special Initiatives Grant, TS0669874), 2006-2011; http://thinkinghead.edu.au/

[2] http://www.stelarc.va.com.au/

[3] Herzog G, Reithinger N (2006) The SmartKom architecture: A framework for multimodal dialogue systems. *SmartKom: Foundations of Multimodal Dialog Systems*, Berlin: Springer.

[4] Cheyer, A., Martin, D. (2001) The Open Agent Architecture. *Journal* of Autonomous Agents and Multi-Agent Systems, *4* (1), 143-148

[5] Black, A. & Taylor. P. (1997) The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR-83, Human Communication Research Centre, Univ. of Edinburgh.

[6] Kuratate, T. (2005) Statistical analysis and synthesis of 3D faces for auditory-visual speech animation, Proceedings of . AVSP'05, 131-136.

[7] Synthetic Times: Media Art China 2008, National Art Museum of China (NAMOC), June 10 to July 3, 2008.

[8] See also *Fagel, S., Kuehnel, C., Weiss, B., Wechsung, I., Möller, S.* A Comparison of German Talking Heads in a Smart Home Environment. Paper at AVSP 2008, in which a German-speaking version of the Thinking Head in a smart home environment with a natural sounding speech synthesis system (MARY) was preferred by users over two other systems.

[9] http://www.stelarc.va.com.au/walkinghead/index.html