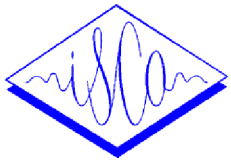
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 – September 3, 2007

Making a Thinking-Talking Head

Chris Davis, Jeesun Kim, Takaaki Kuratate, Johnson Chen, Stelarc, Denis Burnham

MARCS Auditory Laboratories, University of Western Sydney, Australia

Abstract:

This paper describes the *Thinking-Talking Head*; an interdisciplinary project that sits between and draws upon engineering/computer science and behavioural/cognitive science; research and performance; implementation and evaluation. The project involves collaboration between computer scientists, engineers, language technologists and cognitive scientists, and its aim is twofold (a) to create a 3-D computer animation of a human head that will interact in real time with human agents, and (b) to serve as a research platform to drive research in the contributing disciplines, and in talking head research in general. The thinking-talking head will emulate elements of face-to-face conversation through speech (including intonation), gaze and gesture. So it must have an active sensorium that accurately reflects the properties of its immediate environment, and must be able to generate appropriate communicative signals to feedback to the interlocutor. Here we describe the current implementation and outline how we are tackling issues concerning both the outputs (synthetic voice, visual speech, facial expressiveness and naturalness) from and inputs (auditory-visual speech recognition, emotion recognition, auditory-visual speaker localization) to the head. We describe how these head functions will be tuned and evaluated using various paradigms, including an imitation paradigm.

1. Introduction

1.1. Behind Talking Heads

Talking heads implemented by computer graphical methods go back at least to 1971 with Parke's *Initial Parametric Model* [1] that set parameters for eyes, eyelids, mouth; and the later more sophisticated 1974 *Speech Synchronized Animation Parameterized Model* [2]. Such parametric models evolved into Massaro and Cohen's 'Baldi' [3]. A parallel development occurred in talking heads implemented by muscle-based models; Platt and Badler's 1981 *Muscle Based Expression Model* [4] was followed by new muscle models by Waters and others (1987, 1990) [5, 6]. The 1990s saw the introduction of performance-based facial animation with real time speech synchronization (Parke at IBM, Waters at DEC) [7, 8] and the use of data compression methods, e.g., Blanz and Vetter's principle component face model [9].

With sophistication of graphical and mathematical methods came appreciation of the need to make the head appear alive. This has been partially achieved by such techniques as always keeping the eyes in slight motion; enabling simple eye tracking, implementing eye blinks; keeping the head in motion and tying this to acoustic speech parameters; and implementing suitable expressions and coordinating speech parameters and expressions [7, 8]. However, despite this progress in the *look* of talking heads, less progress has been made in the domain of effective *communication* with the head.

1.2. A Communicative Head

Human-human communication is a real-time multimodal dynamic event in which speakers initiate, respond to, and predict actions. Existing machine-human communication systems - from the speech synthesis and automatic speech recognition (ASR) technologies used in commercially deployed systems, through to research-prototype avatars, embodied conversational agents (ECAs), and talking heads - are competent, but to the extent that they attempt to mimic human-human communication, remain limited. The quite obvious difference is that whereas human-human systems involve two or more thinking individuals interacting with each other, machine-human systems involve one less thinker, and lack many of the essential characteristics of human communication.

1.3. Talking-Thinking Head Project Aims

The Thinking Head project is supported by a *Special Initiative* scheme of the Australian Research Council (ARC) and the National Health and Medical Research Council (NH&MRC) [10], and was developed in its initial stages with assistance from the ARC-funded Human Communication Science Network (HCSNet) [11].

The overarching aim of the project is to develop a *Thinking Talking Head*: An advanced embodied conversational agent that combines audio-video processing, speech and speaker recognition, face tracking, utterance interpretation, dialog management, utterance planning, speech synthesis, and animation. There are two parts to this aim: First, we wish to bring together, in one embodiment, a new generation talking *and thinking* head through integration and implementation of software components from various disciplines. Second, we wish to provide a modular plug-and-play research platform for *in situ* evaluation so that researchers can address significant Human Communication Science questions in individual scientific disciplines and technological applications, and more generally in Embodied Conversational Agent (ECA) research.

1.4. Implementation

Four teams of researchers will work together to achieve the aims. The *Computing Team* will be concerned with adapting Head Zero (see 1.2), and setting up an architecture into which off-the-shelf or newly developed software components may be integrated (see 2.3.4). The *Human-Head Interaction (HHI)* Team will be concerned with leeching out from human-human interactions the essential elements necessary for realistic interactions, and applying these to the Thinking Head. There will be a focus on auditory-visual (AV) speech, and its integration with key aspects of intelligence: prediction, interaction and learning. The *Evaluation Team* will conduct controlled experiments on Head-Human behavioural interactions. Recorded data from these interactions will be analysed and fed back into the n+1 develop-evaluate-update iteration, to drive both component software and overall Head development. The *Performance Team* will be concerned with what Brooks (1997) referred to as "the juice of life" [12]. To overcome the apparent lifelessness of ECAs, we will explore a move to ECA-as-performer, leaning heavily on impetus from performative environments by Thinking Head artist-in-residence, Stelarc, and MARCS Lab and Thinking Head researcher, Garth Paine [13, 14]

2. Building A_Head

2.1. Head Zero: A Solipsistic Thinking Head

Head Zero was developed by Stelarc [13] to explore and manipulate ECAs in performance art. Head Zero (Figure 1) is a simple animated artificial head that speaks to an interlocutor, who uses text input. Head Zero has real time lip-synching, speech synthesis and simple facial expressions (frowns, smiles, although these are not linked to the head's speech). The head also nods, tilts and turns its head, and occasionally changes eye gaze direction (although this is unrelated to the interlocutor).

2.1.1 Components

As shown in Figure 1, Head Zero is comprised of four components: A graphics player (Head player); Program D (Alicebot engine); a Text-to-Speech (TTS) engine, and AIML (Artificial Intelligence Markup Language) database. These components are described below.

The Head Player is the main component of Head Zero. It displays two windows, although they appear to be one (see Figure 1). The top window displays the head and all its animations; and the lower window text window displays users' typed text questions. The player monitors whether the user has entered text into the text window. If not, it repeatedly shows animations such as eye blinking, head rotation, smiling, in 30 seconds intervals. When it detects text and a return key press, it calls *Program D* to form an answer in text format, and then calls the TTS system (IBM ViaVoice) to synthesize the text to a wav file and mouth data file. Once executed, it draws the head animation using OpenGL and simultaneously plays the audio file.

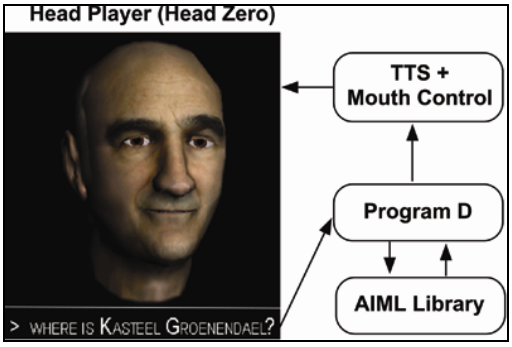


Figure 1. Head Zero. The window above the white line is the head window, below is the text window in which typed interactions with the head are shown (the white border line is not normally shown)

Program D is freely available under the GNU General Public License, and is platform independent (Java 2). It utilizes an AIML database to create responses to questions or statements. When running, Program D preloads all AIML categories and retrieves a response to particular text input by searching all AIML files.

When generating the audio file, Head Player calls a routine to synthesize the text to speech. This routine accepts a text file as input (in this case, answer.txt), and outputs an audio file (answer.wav) and a data file (answer.dat). The audio file reads the answer, and the data file contains information about mouth/jaw and cheek movements. When the Head Player plays the audio file, it uses the data file for generating mouth animation. OpenGL controls the drawing of Head Zero in the Head Player and performs transformations if needed (head rotation, head tilt, eye blinking).

2.2 Head History

Head Zero was first implemented as a *Prosthetic Head* installation in huge displays at 'Transfigure' the Australian Centre for the Moving Image in Melbourne in 2003 [15] and in Glasgow, London, and Toronto. The product was an impressive and commanding interactive agent, but it is clear that this is only a preliminary step in developing a sophisticated thinking head. Below we outline the next few steps we are undertaking to achieve our aims - to build a new generation talking *and thinking* head, and to provide a modular plug-and-play research platform.

2.3 Where to now? Looking A_Head

The look of Head Zero and how it moves is based on composite 2D photographs and parameterized manipulations of a wire-mesh. We plan to make the look of the new head more natural and flexible.

2.3.1 Facial Motion Mapping

The approach we are exploring to create a perceptually accurate talking head animation derives from the work of Kuratate [16]. This is based on a

large scale 3D face database from multiple people consisting of a fixed set of basic speech and non-speech related face postures (see Figure 2 Top). The database is statistically analysed using Principle Component analysis in order to derive similar cross-person components. The 3D face database is also used to estimate face postures as in Figure 3 from single face posture data or reconstructed 3D faces using photographs which is also achieved by a multi-linear technique using the same face database.

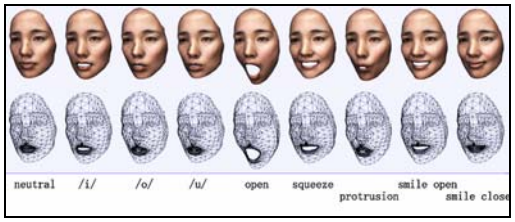


Figure 2. (a) Nine face postures included in the 3D face database (top), (b) Example of mesh adaptation (bottom)

These components can be used to transfer face motion from one person to another. A mesh structure can be fitted to Static 3D postures (see Figure 2 Bottom) to control and render face models and person-specific deformation parameters derived using straightforward multi-linear analysis techniques. Based upon these parameters faces can be synthesized using linear estimation.

Following this, time-varying data obtained for one face (using a motion capture device such as OPTOTRAK - see Figure 3) can be used to control the deformation parameters of other faces.

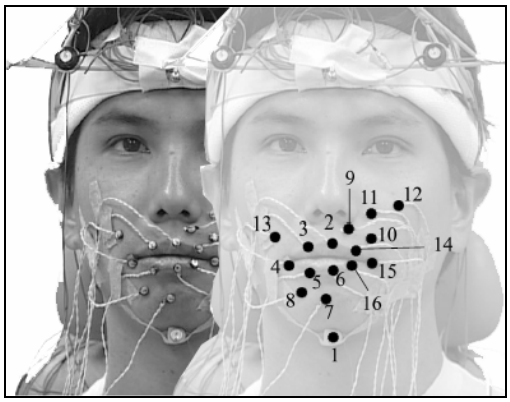


Figure 3. Typical OPTOTRAK face marker configuration

The whole system can then be used by building a phoneme-labeled face motion (PLFM) database to replace the current Head Payer module seamlessly. The face motion of this PLFM database is represented by each diphone (or triphone) to a series of linear combination values of principal components obtained by PCA of motion capture data. The input text data to the TTS system are also used to synthesize time-varying face motion data using the PLFM database according to the phoneme structure of the input text. Then the facial motion mapping will drive any face models in the 3D face database, or pre-estimate any face from single postures or photographs, using the synthesized face motion.

2.3.2 Inputs to the head: Giving the head sense

Head Zero is a platform that illustrates the key notion that people are prepared to attribute limited agency to a talking head even though such a system is very basic [17]. To facilitate this illusion of agency, we will integrate existing speech recognition, speech synthesis, and dialog management systems to allow the Head to converse in a more natural fashion. What is more, the head needs to track lips, faces, heads etc. In order to do this a rudimentary auditory localisation system [18] and an AV ASR and speaker recognition system [19] will be incorporated. In addition, head- and face-motion tracking will be implemented along with basic eye-gaze direction to provide a rough index of the interlocutor's mood and allocation of attention (e.g., see [20]).

2.3.3. The imitation scenario: A perceptual test

Here we outline the first major goal in the development of Head 1.0. In this scenario the head takes auditory and visual inputs from the interlocutor and attempts to imitate both what the person said and the way they said it. This task will require ASR as well as head, face and lip tracking that feed into head motion [21]. With respect to the latter, we will begin with imitating simple exaggerated rigid head motion and then attempt to model finer-grain non-rigid motion. As part of this motion modeling we will explore methods for developing a physical model for the head (see [22]). One benefit of this imitation scenario is that it provides immediate feedback about success. Of course there is a lot more to do: For example, the coordination of face and voice, database construction, dialogue management, etc. Progress will be made through the implementation of a hierarchy of perceptual scenarios (such as the above) that progressively capture aspects of face-to-face interaction.

2.3.4. Plug and Play architecture

The Thinking Head project is a complex software engineering project that invites and will benefit from multiple inputs from both researchers in the project and interested researchers across the wide domain of talking head and associated research. To this end we will employ a plug-and-play architecture that will allow mutual use and portability of across other platforms and systems. We will invite collaboration by researchers plugging in their favorite talking head component and testing it out in a working platform.

One possibility for this plug-and-play architecture is *Artisynth*, a Java-based API for model creation, with GUI support for interactive editing, simulation, and observation. This is an open source platform for collaborative research and development, which provides an interactive simulation environment, developed by Sidney Fels and his Artisynth team [23]. The beauty of this system is that while it

incorporates a particular articulatory synthesis model, it allows the flexible addition or accretion of components in an accessible architecture.

3. Conclusion

What we have attempted to do in this paper is to sketch out the very initial stages of a data driven and perceptually focused approach to developing a talking head whose paralinguistic skills are sufficiently well developed to produce the illusion that the head is "thinking".

The Thinking-Talking Head platform will facilitate research and applications auditory-visual speech processing in various ways: it will advance simulated talking head development, advance the development of components that comprise such heads and stimulate further human-human AV speech processing research in order to feed the Head. Further, the platform will provide a talking head that can be used as a stimulus (1) in investigations of auditory-visual speech perception, (2) in practical applications such as children (with or without learning difficulties, aphasia, forms of dyslexia) learning their first language; child and adult L2 learners; people with limited interactional abilities, e.g., the elderly (pending a suitable interface), autistic children; people with sensory, e.g., hearing impairments; and acting training / imitation / accent modification via comparison of student input to stored expert goals; and (3) to facilitate the development of new media, and film [24] human-machine interfaces, and game animation, and provide the grist for industry linkages.

4. References

[1] Parke, F.I (1972). Computer Generated Animation of Faces. *Proc. ACM annual conference*.

[2] Parke, F. I. (1974). *A Parametric Model for Human Faces*, Ph.D. Thesis, University of Utah, Salt Lake City, Utah, UTEC-CSc-75-047.

[3] M.M. Cohen, D.W. Massaro.(1993). Modeling coarticulation in synthetic visual speech. In N.M Thalman & D. Thalman (Eds.), *Models and Techniques in Computer Animation*, 139–156, Tokyo: Springer-Verlag.

[4] Platt, S.M.& Badler, N. I. (1981) Animating Facial Expressions. *Comp Graphics*, 15, 245-252.

[5] Waters, K. (1987). A muscle model for animating three-dimensional facial expression. *IEEE Computer Graphics*, 21, 17 - 24.

[6] Waters, D. and Terzopoulos, K. (1990). Physically-Based Facial Modeling, Analysis, and Animation. *Journal of Visualization and Computer Animation*, 1, 73–80.

[7] Parke, F. I. (1991) Control Parameterization for facial animation, in N. M. Thalmann and D. Thalmann (Eds.) Computer Animation '91 Tokyo: Springer-Verlag.

[8] Waters, K. (1990) Modeling 3D facial expressions. *SIGGRAPH Facial Animation Course Notes*, 109-129.

[9] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187-194.

[10] 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. (2006-2011). *From Talking Heads to Thinking Heads: A Research Platform for Human Communication Science*, http://thinkinghead.uws.edu.au/index.html (ARC/NH&MRC Special Initiatives, TS0669874).

[11] Dale, R. Burnham, D., Stevens, C., et al. (2005-2009). *Enabling Human Communication*, http://www.hcsnet.edu.au/ (ARC Research Networks, RN0460284).

[12] Brooks, R.A. (1997) From earwigs to humans. *Robotics & Autonomous Systems, 20,* 291-304.

[13] Stelarc: www.stelarc.va.com.au

[14] Garth Paine: www.activatedspace.com/

[15] www.stelarc.va.com.au/prosthetichead/index.html & www.acmi.net.au/transfigure/flash.htm

[16] Kuratate, T. (2004) Estimation of 3D face expressions from front and side view photographs using a 3d face database (in Japanese). *Proceedings of Visual Computing / Graphics & CAD Joint Symposium*, 61–66.

[17] Nass, C., Y. Moon, B. J. Fogg, B. Reeves, & D. C. Dryer. 1995. Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43, 223–239.

[18] Kwok, R., Buchholz, J. M., Fang, G., and Gal, J. (2005). Sound source localization: Microphone array design and evolutionary estimation, *Proceedings of ICIT2005*, Hong Kong.

[19] Goecke, R. (2005) 3D lip tracking and co-inertia analysis for improved robustness of audio-video automatic speech recognition. In Vatikiotis-Bateson, E., Burnham, D. & Fels, S. (Eds.), *Proceedings, Auditory-Visual Speech Processing International Conference*, Adelaide, Causal, 109-114.

[21] Hayashi, K. Onishi, Y. Itoh, K. Miwa, K. & Takanishi, A (2006). Development and Evaluation of Face Robot to Express Various Face Shape. Proc. IEEE International Conference on Robotics and Automation

[22] Lee, C.H., Wetzel, J., Selker, T. (2006). .Enhancing Interface Design Using Attentive Interaction Design Toolkit. Paper in Educators program, SIGGRAPH.

[21] Lee, S.-H.& Terzopoulos, D. (2006). Heads up! Biomechanical modeling and neuromuscular control of the neck in ACM Transactions on Graphics 25, 1188-1198.

[22]Fels, S., Llloyd, J.E. vanden Doel, K., Vogt, F., Stavness, I, Vatikiotis-Bateson, E. Developing physically-based, dynamic vocal tract models using Artisynth. In H. C. Yehia, D. Demolin, Laboissiere, R. *Proceedings of ISSP 2006, 7th International Seminar on Speech Production*, 419-426.

[23] Katzman, A. (2005) Mimesis and praxis in the art of traditional facial animation. *ATR Symp. Cross-modal Proc. Faces & Voices*, 50–51.