

TALKING HEADS AND SYNTHETIC SPEECH: AN ARCHITECTURE FOR SUPPORTING ELECTRONIC COMMERCE

Jörn Ostermann, David Millen

AT&T Labs - Research
100 Schultz Dr., Red Bank, NJ 07701, USA
email: {osterman,drm}@research.att.com

ABSTRACT

Facial animation has been combined with text-to-speech synthesis to create innovative multimodal interfaces. In this paper, we present an architecture for this multimodal interface. A face model is downloaded from a server into a client. The client uses an MPEG-4 compliant speech synthesizer that animates the head. The server sends text and animation data to the client in addition to regular content to be displayed in a web browser. We believe that this architecture can support electronic commerce by providing a more friendly, helpful and intuitive user interface when compared to a regular web browser. In order to substantiate these claims, we undertook experiments to understand user reaction to interactive services designed with synthetic characters. In one experiment, participants played the 'Social Dilemma' game with the computer as a partner. Results indicate that users cooperate more with a computer when an animated face is representing the computer during the game. A simulated commercial application was evaluated, also comparing the performance of facial animation, text-to-speech and text only conditions. According to the results, the use of facial animation in the design of interactive services was favorably rated for most of the attributes in these experiments. Further, the results show that facial animation may effectively fill application-waiting times and make delays more acceptable to the users.

1. INTRODUCTION

Computer simulation of human faces has been an active research area for some time, resulting in the development of a variety of facial models and the development of several animation systems [3][4][7][13][15][21][24]. The advances in animation systems, such as those mentioned above, have prompted interest in the use of animation to enrich the human-computer interface. This prompted ISO to support animation of talking faces and bodies in the MPEG-4 standard [9][10][11][17][19][20]. One important application of animated characters has been to make the human computer interface more compelling and easier to use. For example, animated characters have been used in presentations systems to help attract the user's focus of attention, to guide the user through several steps in a presentation, and to add expressive power by presenting nonverbal conversational and emotional signals [1][22]. Animated guides or assistants have also been used with some success in user help systems [2][6][8] and for user assistance in web navigation [16].

Character animation has also been used in the interface design of communication or collaboration systems. There are several multi-user systems that currently use avatars, which are animated representations of individual users [24][21]. In many cases, the avatar authoring tools and online controls remain cumbersome. The social cues that are needed to mediate social interaction in these new virtual worlds have been slow to develop, and have resulted in frequent communication misunderstandings [9]. Nevertheless, the enormous popularity of Internet chat applications suggests considerable future use of avatars in social communication applications.

In Section 2, we present the client and server architecture for supporting web-based applications like electronic commerce with text-to-speech speech (TTS) synthesis and facial animation (FA). In Section 3, we show the use of facial animation in an information kiosk. We also present the results of subjective tests using this information kiosk. In Section 4, we present the 'Social dilemma' experiment that shows how FA and TTS can influence the users in their interaction with a computer.

2. ARCHITECTURE FOR TTS AND FACIAL ANIMATION FOR WEB-BASED APPLICATIONS

In order to enable web-based FA on a client, the client requires a web browser, a TTS and a FA renderer (Figure 1). Usually, the settings like speech rate of a TTS are determined by the preferences of the user. A server does therefore not know them. In order to enable synchronized speech and facial animation, the TTS must provide phonemes and related timing information to the FA renderer. Using a coarticulation model [4], the renderer can then animate a model downloaded from a server and let it move its mouth synchronously to the speech of the TTS. Driving the face model using the text of the TTS does not allow for animating non-speech related actions like smiles or head nodding. Therefore, the TTS has to handle bookmarks that contain these facial animation parameters (FAP). The bookmarks are placed in the text and their timing is determined by the TTS from the start time of the word following the bookmark [11][18][23].

The server for this client comprises a web server, a TTS/FA server and a database of face models (Figure 1). They are controlled by the application, which could be implemented as a CGI script. When the application is started due to a client request, it downloads a face model from the model library to the client. We use VRML [12] as a file format for the face models. In

order to increase the coding efficiency, the binary MPEG-4 BIFS format [9] can be used. Afterwards, the application provides the HTML-pages for the server and the text with embedded FAPs to the TTS/FA server. This TTS/FA server can be implemented using MPEG-4 [9][11][10].

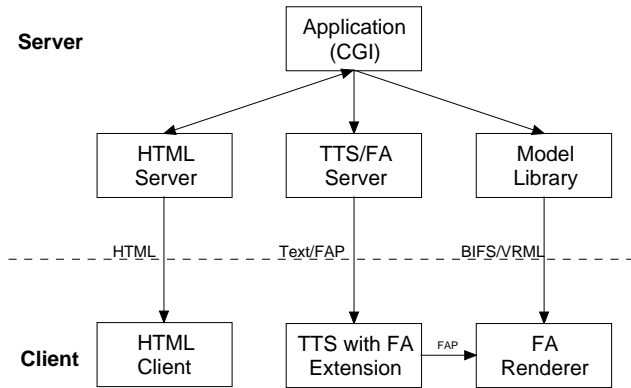


Figure 1: Architecture for using TTS and face animation (FA) in an Internet environment.

3. INFORMATION KIOSK EXPERIMENT

This experiment was designed to assess the user benefits of an animated agent assistant for an information retrieval task. To this end we created an information retrieval service with a limited scope. Users were given a short information retrieval task to provide experience using the service, and then asked to evaluate the service by answering a computer-administered questionnaire.

Participants in the study performed the information search task using one of three interfaces: text only, TTS + text, TTS + FA + text.

3.1 Experimental Procedure

In this experiment 190 students or employees of Princeton University completed the experiment. Participants were paid for their participation. The median age was 25 and the majority of subjects were self-described heavy users of computers at work and at home.

Participants were asked to use a simple interactive system to gather information about theatre shows. The interactive service was designed to be fully functional in a text-only version. The facial animation and text-to-speech utterances were incremental and designed to make the service more engaging and friendly. More important, facial animation and text-to-speech was designed to fill wait times that were designed into the service. The wait times were typical of interactive service delays using a dial-up Internet connection.

The service begins with a welcome message, and then gives the user a choice of Broadway shows. The user chooses a show, and is presented with a choice of available information about the show: review, venue and price. There are built-in delays from the time that the user selected the information that they want to see and the display of that information. Participants can then request additional information about the currently selected show,

look for information about a different show, or exit the information service.

To insure that subjects spend sufficient time using the system, they were asked to choose a theatre show and find some information about it. For example, they were asked to find out whether the review of the show was good or bad, where the show was playing and what was the price of a ticket. The information gathered by the participants was written down on a data collection form.

Upon completion of the information-gathering task, they were asked to fill out a short computer-administered questionnaire.

3.2 Results of Information Kiosk Experiment

Table 1 presents the subjects' answers to the questions relevant to the speed of service and waiting times. Both rows show similar patterns, which can be expected due to the similar nature of the two questions. The subjects having the TTS support and those having TTS and face are both more satisfied with the speed of service and remarked less the waiting times than the subjects using the text-only service. Since the service was in fact exactly the same with respect to speed and waiting times, we can conclude that audio and face distract the users and make the waiting times less noticeable.

Table 1: Average subjective ratings for the Information Kiosk

	TTS	TTS + face	Text only	P value
Satisfaction with speed	4.4	4.4	3.8	p <.01
Satisfaction with waiting time	4.0	4.2	3.5	p <.001
Overall satisfaction	4.4	4.6	4.3	
Was service user friendly?	4.9	5.1	4.7	p <.01
Easy to use?	5.2	5.4	5.3	
Was the service human-like?	2.7	3.0	2.4	p <.05
Estimate of sound quality	4.5	4.7	(NA)	

Table 1 shows the answers to several questions concerning the user satisfaction and different aspects of the quality of service. Looking at the data a general trend may be noticed. TTS and TTS+Face conditions tend to be similar to each other and better than the rest. They are followed by the Text only. There are some exceptions to this trend. Notably, all three conditions are rated similarly on overall ease of use, which likely reflects a ceiling effort of the consistently high ratings.

Another, weaker trend is for service with the face to be judged somewhat better than the one with TTS; in particular it is judged to be more human-like.

The last row in Table 1 would show any influence of the presence of the face on the perceived sound quality. There is no such influence.

Several questions were asked about the synthetic face itself. The participant responses indicated that the face was reasonable friendly (Mean rating = 4.1, where 6 = most friendly) and not particularly distracting (mean rating = 4.9; where 6 = least distracting). It was found to be only marginally useful (mean =

3.2, where 6 = most useful), which is reasonable considering that the character was deliberately designed and programmed to be more conversational than utilitarian.

4. 'SOCIAL DILEMMA' GAME EXPERIMENT

Using the 'social dilemma' game, we investigated whether user cooperation with a computer is influenced by the presence of text, speech or face animation. Users played the 'social dilemma' game with the computer as partner (Figure 2). The rules of the game are as follows: Partners have to choose project Green or project Blue in a secret ballot. If both partners choose Green, both receive \$6. If both choose Blue, both receive \$3. If one partner chooses Green and one chooses Blue, the first receives \$0 whereas the latter gets \$9. According to these rules, we say that a partner that chooses Green is cooperating because he favors both partners to earn money. Goal of the game is 'Earn as much money as possible'. We do not define whether the partners should maximize their own money or the joint money.

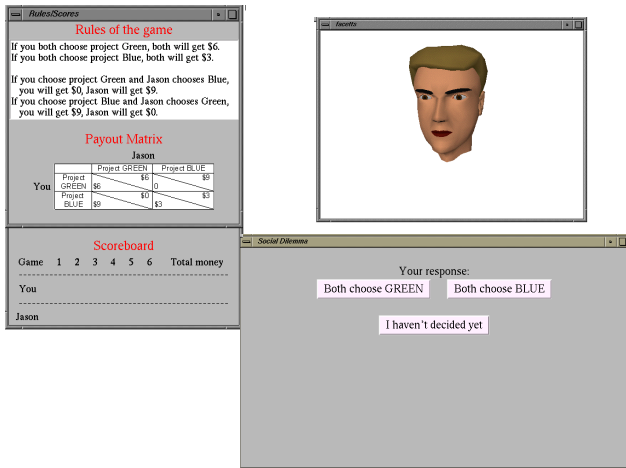


Figure 2: Screen shot of the game. The left window permanently displays the rules of the game and the current score. The top right window shows the face. The bottom right window displays an instant of the dialog box that is used as part of the discussion phase.

The game starts with the computer and the partner introducing each other. For each round of the game, the user and partner discuss on what they will do in this round of the game. After they have communicated their intentions, they secretly select a project using a ballot window. After both partners have selected a project, the scores appear on the scoreboard.

Participants in the study played the game using one of three interfaces: The computer partner was either represented using text only, TTS only, or TTS and FA.

4.1 Experimental Setup

50 students and employees of Princeton University participated in the experiment. They were paid for their participation. The

median age was 25 and the majority of subjects were self-described heavy users of computers at work and at home.

Each participant played 6 rounds of the game. The computer was always initiating the dialog prior to a ballot. Using different wording, the computer suggests cooperating. The human partner indicates his intentions by typing free text and confirming his intention by pressing on of three buttons (similar to Project Green, Project Blue, I don't know, Figure 2). As far as the ballot is concerned, the computer is always cooperating (Project Green), unless the human partner suggested that both choose Blue. However, the human partner did not know the behavior of the computer.

Upon completion of the game, users were asked to fill out a short computer-administered questionnaire.

4.2 Experimental Results

We define the cooperation rate of a partner as the percentage of choosing Green in his actual decision. Figure 3 shows the average cooperation rate of the human partner as a function of the round of the game. As can be seen, the cooperation rate is highest when the computer is represented using TTS and FA. A TTS representation yields a higher cooperation rate than text only. The decrease of cooperation for the last round is expected as described in the literature [26]. Users know that a lack of cooperation in the last round of the game will not result in future consequences. We believe that we will be able to exploit this higher cooperation rate when TTS and FA are used in ecommerce applications where we want to guide the user through a product selection and buying process.

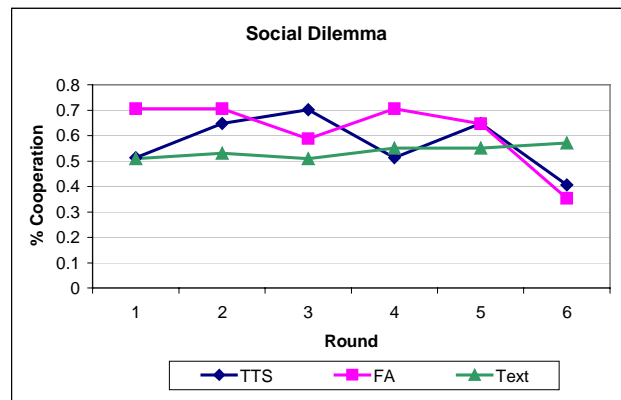


Figure 3: Cooperation rate of a human user as a function of the round of the game. Different modalities representing the computer achieve different cooperation rates.

The evaluation of the questionnaires showed a weak trend that users prefer the interface with FA and TTS. There is one exception in that the intelligence of the computer partner was judged higher if it was represented with text only. Despite of this, cooperation is higher for the FA and TTS interface (67%) than for the text interface (52%) At this point, we are not sure on what caused this discrepancy.

5. SUMMARY

In this paper, we present an architecture that integrates TTS and face animation (FA) into Internet applications. A server sends a face model to a client. The client animates this face model and synthesizes speech according to the text and embedded bookmarks that it receives from the server. The bookmarks enable the animation of non speech-related actions like head nodding and smile.

In subjective tests, we evaluated this technology simulating an information kiosk and a game geared towards measuring cooperation comparing the performance of facial animation, text-to-speech and text only conditions. According to the results, the use of facial animation in the design of interactive services was favorably rated for most of the attributes in these experiments. Further, the results show that facial animation may effectively fill application-waiting times and make delays more acceptable to the users. An important result for Ecommerce is that users cooperate more with a computer if it is represented with FA and TTS instead of TTS only or text only. We measured an increase of the cooperation rate from 50% for text to 70% for TTS and FA in a first human computer interaction. The average cooperation rate over 5 consecutive interactions increased from 52% to 67%. Based on these results we expect that face talking heads will increase the performance of web stores and web-based customer service.

6. REFERENCES

- [1] Andre, E., Rist, T. & Muller, J. Guiding the User Through Dynamically Generated Hypermedia Presentations with a Life-Like Character. *Intelligent User Interfaces '98*. (San Francisco, Ca, 1998), 21-28.
- [2] Arafa, Y., Charlton, P., Fehin, P. & Mamdani, A. Personal Service Assistants with Personality. *Proceedings of HCI International '99 - Volume 2*. (Munich, Germany, August 22-26, 1999), Lawrence Erlbaum, 147-151.
- [3] Cosatto E., Graf H.P., Sample-Based Synthesis of Photo-Realistic Talking Heads. *Proc. Computer Animation '98*. (Philadelphia, USA), 103-110.
- [4] Cohen, M. M. and Massaro, D. W. Modeling Coarticulation in Synthetic Visual Speech. In M. Thalmann & D. Thalmann (Eds.) *Computer Animation '93*, Tokyo: Springer-Verlag.
- [5] Damer, B., Kekenes, C., & Hoffman, T. Inhabited Digital Spaces. *Proceedings of CHI '96*, ACM Press, 9-10.
- [6] Don, A., Oren, T., & Laurel, B. Guides 3.0. In *CHI-93, Video Proceedings*, ACM Press, pp. 447-448.
- [7] Eisert, P., Chaudhuri, S., and Girod, B. Speech Driven Synthesis of Talking Head Sequences. *3D Image Analysis and Synthesis*, Erlangen, November 1997, pp. 51-56.
- [8] Gibbs, S, Breiteneder, C. Video Widgets and Video Actors. *UIST '93*. pp. 179-185.
- [9] ISO/IEC IS 14496-1: MPEG-4 Systems, 1999.
- [10] ISO/IEC IS 14496-2: MPEG-4 Visual, 1999.
- [11] ISO/IEC IS 14496-3: MPEG-4 Audio, 1999.
- [12] ISO/IEC 14772-1: 1997, Information Technology – Computer graphics and image processing – The Virtual Reality Modeling Language – Part 1: Functional specification and UTF-8 encoding.
- [13] Kalra P., Mangili A., Magnenat-Thalmann N., Thalmann D., Simulation of Facial Muscle Actions based on Rational Free Form Deformation”, *Proceedings Eurographics 92*, pp. 65-69
- [14] Kalra P. An Interactive Multimodal Facial Animation System, PhD Thesis nr. 1183, EPFL, 1993
- [15] Kampmann, M., Nagel, B. Synthesis of Facial Expressions for Semantic Coding of Videophone Sequences. *Computer Graphics International (CGI98)*, Hannover, Germany, June 1998.
- [16] Milewski, A. E. & Blonder, G. E. System and Method for Providing Structured Tours of Hypertext Files. US Patent # 5760771. June 2, 1998.
- [17] Ostermann J., Animation of Synthetic Faces in MPEG-4”, *Proc. Computer Animation '98*, (Philadelphia, USA, 1998),. 103-110.
- [18] Ostermann, J., Beutnagel, M., Fischer, A., Wang, Y. Integration of talking heads and text-to-speech synthesizers for visual TTS. *ICSLP 99*, Australia, December 99.
- [19] Tekalp, A. M., Ostermann, J., “Face and 2-D mesh animation in MPEG-4,” *Signal Processing: Image Communication*, vol. 15 (4-5), 2000, pp. 387-421.
- [20] van Beek, P., Petajan, E. and Ostermann, J., “MPEG-4 synthetic video” in *Multimedia Systems, Standards and Networks*, Puri, A. and Chen, T. (Ed.), Marcel Dekker, New York, 2000.
- [21] Pandzic, Igor S., Capin, T. K., Lee, E., Magnenat-Thalmann N., Thalmann, D. "A flexible architecture for Virtual Humans in Networked Collaborative Virtual Environments", *Proceedings Eurographics 97*, (Budapest, Hungary, 1997).
- [22] Rist, T., Andre, E., & Muller, J. Adding Animated Presentation Agents to the Interface. *Intelligent User Interfaces '97*, (Orlando, Florida), pp.79-86.
- [23] Sproat, R. and Olive, J. An Approach To Text-to-speech Synthesis. In W.B. Kleijn & K.K. Paliwal (Eds.) *Speech Coding and Synthesis*, Elsevier Science, 1995.
- [24] Suler, J.R. (1997). From ASCII to Holodecks: Psychology of an Online Multimedia Community. Presentation at the Convention of the American Psychological Association, Chicago.
- [25] Terzopoulos D., Waters K., Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. *IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15/6*, 1993, pp. 569-579.
- [26] Van Mulken, S., Andre, E., & Muller, J. An Empirical Study on the Trustworthiness of Life-Like Interface Agents. *Proceedings of HCI International '99 - Volume 2*. (Munich, Germany, August 22-26, 1999), Lawrence Erlbaum, 153156.