PCS 99

Subjective Tests with Animated Talking Faces: What are they good for?

Jörn Ostermann, David Millen, Igor S. Pandzic[1],

AT&T Labs – Research

Room 3-231, 100 Schultz Dr., Red Bank, NJ, 07701

email: {osterman,drm}@research.att.com, Igor.Pandzic@cui.unige.ch

Abstract

Computer simulation of human faces has been an active research area for a long time, resulting in a multitude of facial models and several animation systems. Current interest for this technology is clearly shown by its inclusion in the MPEG-4 standard. However, it is less clear what the actual applications of Facial Animation (FA) will be. We have therefore undertaken experiments on 190 subjects in order to explore the benefits of FA. Part of the experiment was aimed at exploring the objective benefits, i.e. to see if FA can help users to perform certain tasks more accurately or efficiently. The other part of the experiment was aimed at more subjective benefits, like raising the level of appeal to the user, gaining more users' interest, filling in the waiting times for server access so the users do not get bored. At the same time a preliminary study into comparison of different FA techniques was undertaken. We present the experiment design and the results. The results show that FA aids users in understanding spoken text in noisy conditions (error rates drop from 16% to 8%); that it can effectively fill the waiting times and make them more acceptable to the user; and that it makes services more attractive to the users, particularly when they compare directly the same service with or without the FA.

1 Introduction

Computer simulation of human faces has been an active research area for a long time, resulting in a multitude of facial models and several animation systems [Kalra92, Fischl93, Kalra93, Terzopoulos93, Ostermann97, Parke97, Pandzic97, Eisert97, Cosatto98, Kampmann98]. Current interest for this technology is clearly shown by its inclusion in the MPEG-4 standard [Ostermann98, Doenges97].

The advances in animation systems, such as those mentioned above, have prompted interest in the use of animation to enrich the human-computer interface. To understand further the utility and usability of a facial display interface, we have completed several experiments. In the first experiment, we consider the benefits of a facial display as a distinct channel in a multi-modal interface. In this experiment we explore the performance benefits of using Facial Animation in a number intelligibility task. It was expected that facial animation synchronized with speech would result in better performance over speech alone in a noisy ambient environment.

In the second and third experiments, we tested user performance and preferences in a kiosk application across a variety of interface conditions. These experiments were intended to explore the more subjective benefits of facial animation displays, such as increasing the task interest and appeal, and minimizing the negative aspects of system delays.

In all three experiments, several different facial animation techniques were used. The results, therefore, provide a preliminary study of the performance and preferences of different FA techniques.

In the next section we present the experiment design, describing in detail the technical setup, the experimental tasks for different experiments and the subjects. In section 3 we present the detailed results of the experiments followed by a summary of most important results. Finally we give conclusions and discuss issues for further study.

2 Experiment design

Three experiments were undertaken, each examining different aspects of FA.

Experiment 1 was primarily aimed at measurable effects of FA (Figure 1a,c), rather then based on subjects' evaluation of certain criteria (though a questionnaire was also used as a second source of information). The measurement was performed by observing how well the subjects can perform a task with or without Facial Animation (FA), and under different conditions.

The following effects have been explored:

- Effect of FA on speech understanding in optimal acoustic conditions
- Effect of FA on speech understanding in noisy conditions

[1] This work was conducted while Igor Pandzic worked at AT&T Labs-Research during the summer of 1998.

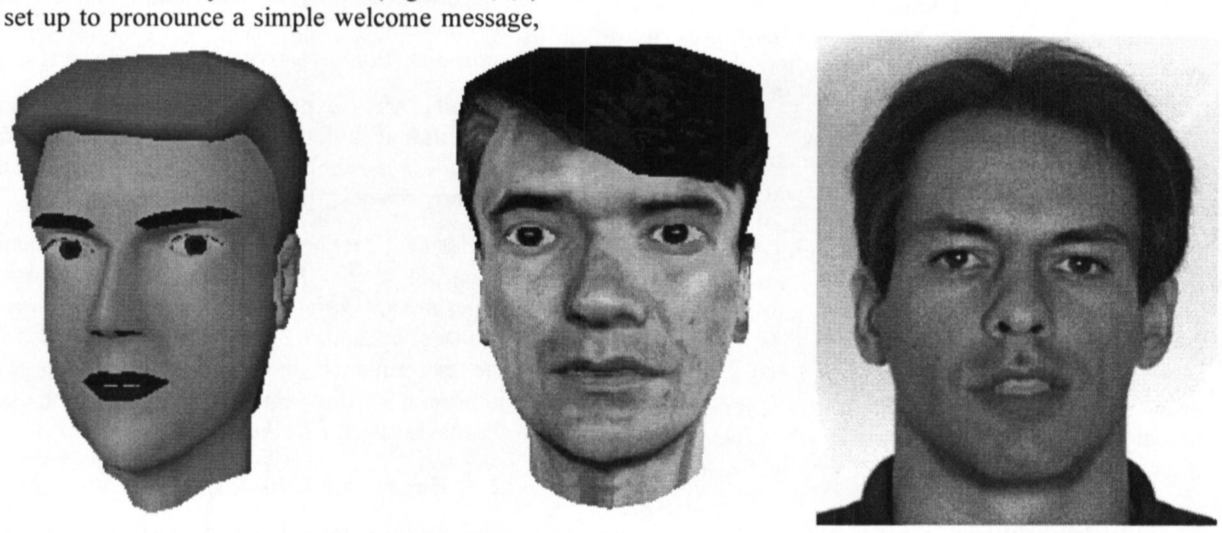
- Effect of changing FA frame rate on speech understanding in both noisy and optimal conditions
- Effect of changing FA techniques on speech understanding in both noisy and optimal conditions

Experiment 2 was aimed at more subjective benefits of FA (Figure 1a): the general appeal to the user, making a service more friendly, filling the waiting times and in general improving the users' satisfaction. For this purpose a simple service with a limited scope was conceived and the subjects were asked to use it and then to answer questions related to their level of satisfaction with the service. The response to the service with and without facial animation and synthetic voice has been compared.

Experiment 3 was a preliminary study into comparison of different methods to generate synthetic faces. Three different synthetic faces (Figure 1a,b,c) were set up to pronounce a simple welcome message, and the subjects were asked to compare and evaluate different faces.

2.1 Technical setup

A Text-To-Speech (TTS) system is coupled and synchronized with a FA system, yielding a Visual TTS (VTTS) system that simulates a talking head pronouncing arbitrary text in real time [Sproat95]. The FA system is based on a 3D polygon mesh face model with defined facial actions allowing the simulation of speech and facial expressions like smiling, being angry etc. The coarticulation model is the one from [Cohen93]. The facial model can be modified [Osterman97]. The actual facial models used in the experiments are shown in Figure 1 a) and b). The cartoon like character (Figure 1a) was used in all experiments, the texture mapped character (Figure 1b) was only used in Experiment 3.



a) b) c)

Figure 1: a) standard face, b) texture mapped face, c) sample-based face.

Additionally, the sample-based FA system has been used in experiments 1 and 3 [Cosatto98]. In this system a set of samples of the mouth area is extracted from a video of a real person talking. The samples are classified according to the mouth shape. This database of mouth shapes is used to generate appropriate mouth movement according to the speech pronounced by the TTS system. Functionally this yields the same system as VTTS, but the visual result is different. The image looks much more realistic, however the lip movement looks less natural. Currently, this method does not run in real time, therefore the utterances needed for the experiment were recorded off line and shown as video clips during the experiment. Figure 1 c) shows the sample-based face.

During the tests, subjects were using headphones for better control of the acoustic environment.

2.2 Experimental task

2.2.1 Experiment 1

The subject's task was to listen to several series of numbers (digits), and type them in. There were five numbers per series. The numbers were pronounced by the TTS system. The subject could type the number in only when each series of five numbers was fully spoken out. The error rate was measured on a digit-by-digit basis. The subjects were given two trial series of five digits each, then ten measurement series of five digits each. Each subject repeated the task in noisy and optimal acoustic conditions. The order of noisy and optimal condition was randomized and different for each subject. In the noisy conditions, the Signal to Noise Ratio was –2 dBA. This is a very difficult hearing condition, corresponding roughly to talking on the phone in a noisy airport while a flight announcement is heard from a nearby loudspeaker. Such difficult conditions were chosen in order to have a significant error rate and to be able to measure improvements when FA is deployed.

The subjects were split into four groups, each group having different visual conditions. The summary of different visual conditions is shown in Table 1.

After each test the subjects were given a questionnaire. For each question an appropriate six-point scale of answers was offered to the subject.

Table 1: Summary of different visual conditions in Experiment 1.

Condition name	Face rendering	Frame rate
No Face	none	-
Low Frame Rate	3D	10 Hz
Standard Face	3D	18 Hz
Sample-Based Face	sample based	30 Hz

2.2.2 Experiment 2

The subjects were asked to use a simple interactive real-time system giving information about theatre shows. The service was conceived in such a way that it can perfectly be used without FA. FA is just a gadget to make the service more engaging and friendly. The service involves waiting time (simulating Internet and server access waiting times) that is filled by FA and/or speech synthesis. The face acts as the representative of the service, welcomes the users and asks questions about what they would like to find out. Expressions (smiles) were used in an effort to make the face more pleasing.

The service starts with a welcome message, 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 prices. When the user has chosen which information he/she wants, there is a waiting time before the information is actually displayed, simulating the waiting times on the Internet. After reading the desired information, the user can choose to get more information about the same show, to get information about another show or to exit.

To insure that subjects spend sufficient time using the system, they were asked to choose a theatre show and find some information about it: the review (was it good or bad?), venue and the ticket price. They were given a data sheet where they had to write down this information. This insured that the subjects went through all features of the system.

The experiment was performed in varying conditions with respect to the presence of the visual and acoustic stimuli (FA and TTS). In addition, one group of users was tested using a text-only version of the interface. The duration of each test was under 5 minutes.

The questionnaire was used to interrogate the subjects on usefulness of the system, usefulness of FA, friendliness of the system, ease of use, perceived sound quality and possible distraction or annoyance by the face animation. For each question an appropriate scale/choice of answers was offered.

2.2.3 Experiment 3

The subjects viewed/listened to short audio-visual sequences showing a face pronouncing a welcome message: "Welcome to AT&T global communication services". The welcome message was pronounced by three different synthetic faces: non-textured 3D model, textured 3D model and sample-based model, as illustrated in Figure 1. No facial expressions (smiles or other) were used on the synthetic face. The three sequences were shown to the user in random order. After the first showing, all three sequences were shown a second time (in the same order as the first time), and this time after each sequence the subject was asked how he/she liked that particular welcome message. An appropriate scale was offered for the answer to this question.

The purpose of this experiment was to compare different approaches to FA and also see to what extent users are sensitive to artifacts in mouth movement that may occur in some of the approaches.

2.3 The subjects

A total of 190 subjects have completed the experiment at Princeton University. Their age was between 20 and 35 years with a median of 25. It can be observed that the subjects were in general young, very frequent computer users and relatively often (39%) not native English speakers.

3 Results

In this section we present and analyze the results of all three experiments. Significance tests [ANOVA and Scheffe Post Hoc tests] were performed for each experiment, for each of the performance variables and attribute ratings. The observed results that are significantly different from chance (i.e., probability values less than .05) will be reported.

3.1 *Experiment 1*

In the following subsections we present the error rate and timing results in all conditions, as well as the subjective responses collected in the questionnaire.

3.1.1 Error rates

Figure 2 shows the error rates in all conditions. The most obvious observation is that there are much more errors in noisy conditions, as expected ($p< .001$).

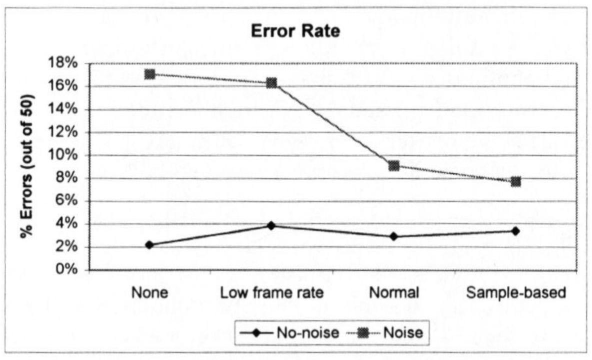


Figure 2: Average error rates; comparison between different conditions with and without noise.

There was an interesting difference in error rates as a function of noise and presentation condition ($p< .001$). In optimal acoustic conditions (no noise) there is no significant difference in error rates between different visual conditions, i.e. all subjects did very few mistakes when no noise was present. However, in noisy conditions significant differences can be observed between different visual conditions. Subjects doing the experiment without the face, or with a low-frame-rate face did much worse then those with the standard or sample-based face, with mean error rates dropping from approximately 16% to approximately 8%.

No significant difference is observed between low-frame-rate face and no face at all. This suggests that 10 Hz is not a high enough frame rate to provide a useful visual speech pronunciation. An increase in frame rate to 18 Hz (normal face) provides obvious improvement in error rates. Although the sample-based face was played at 30 Hz, the face synthesis method is different and this cannot be used for direct comparison in terms of frame rate.

The fact that there is no significant difference between the standard 3D face and sample-based face is somewhat surprising because both the experts and the subjects agree that the sample-based face shows jerkiness in lip movement and lower quality of lip synchronization. Therefore it was expected to show higher error rates than the standard 3D face. One possible explanation of these results is that the sample-based face was played at 30 Hz vs. 18 Hz for the standard 3D face. If we assume that higher frame rate yields better results, this may have compensated for the artifacts in the sample-based face.

3.1.2 Subjective responses

After completing the experiment the subjects were asked to fill in a questionnaire asking some questions about their experience.

As expected, subjects rated several interface attributes lower in the noise condition compared with the "no noise" condition. In particular, subjects rated both the sound quality and the ease of understanding reliably lower in the noise condition. Furthermore, subjects rated the presence of facial animation to be more useful and less distracting in the noise condition.

The "ease of understanding" ratings were similarly high across all presentation conditions (4.6 out of 6). On the other hand, there were several aspects of the task that varied as a function of the presentation condition. As can be seen in Table 2, the 3D face, both in the low-frame-rate and normal conditions, was found somewhat less distracting and more useful than the sample-based face. The normal 3D face was also rated slightly more useful than the low frame rate face.

Table 2: Subject attribute ratings across presentation conditions (1=no face, 2=low frame rate face, 3=normal, 4=sample-based).

	1	2	3	4
Ease of Understanding (6 = easiest)	4.3	4.5	4.6	4.6
Sound Quality Ratings? (6 = best)	3.9	4.2	4.6	4.1
Was the face useful? (6= most useful)	-	3.0	3.2	2.3
Was the face distracting? (6=least distracting)	-	5.0	5.1	4.3

It is interesting to note the discrepancy between the subjects' estimate of the usefulness of the face (Table 2) and its objective usefulness (Figure 2). Although the subjects show better performance results (Figure 2), they do not seem to attribute this improvement to the presence of the face, as the scores on the usefulness of the face are relatively low. Even more surprising is the comparison of usefulness scores of the low frame rate and sample-based faces (Table 2). Sample-based face obtained a substantially lower usefulness score despite the fact that objective results show the exactly opposite effect. This may suggest that the visual cues (lip reading) are used subconsciously.

Further, Table 2 indicates that the subjects found the sound quality of the speech better with the standard 3D face then in other conditions, meaning that the presence of face positively influenced their perception.

3.2 Experiment 2

All the results in experiment 2 are the subjects' responses to a questionnaire. We present and analyze the results related to timing, user satisfaction and some specific questions concerning the synthetic face.

3.2.1 Timing

Table 3 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 audio support and those having audio 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.

Even more noticeable is that subjects preconditioned with the experiment using the Audio+face or Audio show less satisfaction and more annoyance with the waiting time if they use the Text-only service afterwards. This may be explained by the fact that they have already used the service with audio/face offering distraction and notice more the waiting times in the simple text-only service.

Table 3: Subjects attribute ratings across presentation conditions.

	Audio	Aud + face	Text only	P value
Satisfaction with speed (6 = best)	4.4	4.4	3.8	p <.01
Satisfaction with waiting time (6 = best)	4.0	4.2	3.5	p <.001
Overall satisfaction (6 = most satisfied)	4.4	4.6	4.3	p>.1
Was service user friendly? (6 = best)	4.9	5.1	4.7	p <.01
Easy to use?	5.2	5.4	5.3	p>.1
Positive or negative emotions? (6 = most positive)	4.3	4.3	4.2	p>.1
Was the service human-like?(6 =most human-like)	2.7	3.0	2.4	p <.05
Estimate of sound quality (6 =best)	4.5	4.7	(NA)	p>.1

3.2.2 User satisfaction

Table 3 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. Audio and Audio+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, Text Only service is judged to be slightly easier to use than the others; however the ease of use for all conditions is judged so high that differences here are minimal.

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

The last row in Table 3 would show any influence of the presence of the face on the perceived sound quality. There is no such influence.

3.3 Experiment 3

In the third experiment the subjects were asked to compare different face models as illustrated in Figure 1 by looking at a simple welcome message pronounced by each face and answering how they liked each presentation. The mean *appeal* rating for synthetic faces were 5.0 for the standard face (Figure 1a), 2.7 for the texture mapped face (Figure 1b), and 3.3 for the sample-based face (Figure 1c). The standard 3D polygon mesh facial model was clearly preferred by this group of users.

4 Conclusions

Experiments have been undertaken in order to examine the potential usefulness of Facial Animation (FA) combined with Text-To-Speech (TTS) technology for interactive services. Three experiments were run with a total of 190 subjects.

The results can be summarized in the following points:

- In optimal acoustic conditions FA does not help understanding; in case of significant artifacts of the mouth motion, it can slightly worsen the understanding.
- FA at 18 Hz and sample-based FA at 30 Hz help understanding significantly in noisy conditions (error rate drop from 16% to 8%), though they also slightly increase the time the subjects (by less then 10%) spend on the task.
- In general the face is not found distracting.
- In general the face is not perceived as very useful, even when results (e.g. error rate) show that it actually is useful.
- FA can distract users during waiting times and make the waiting time appear shorter. However, TTS (audio) has a very similar effect even without FA.
- People react more positively to a service with FA then without; this difference is even more pronounced when users try the service without FA after trying the one with FA.
- A service with FA is considered more human-like and provokes more positive feelings than the one with TTS (audio) only (both by 1 point on a six-point scale).

It can be concluded that a wider deployment of FA may be worthwhile in interactive services, making them more attractive to users. It will be of interest to learn whether the level of satisfaction with the plain-text service drops significantly after the users have been introduced to the same service with FA enhancement.

Current facial models and animation techniques are still rather crude and technical, which is to a good extent the result of their being created by engineers rather then artists. Experiment 3 has shown that the general appeal level of all used faces is rather low. It is therefore expected that the results in terms of user appeal may improve greatly by making the faces/animations more attractive.

5 References

[Cohen93] M.M.Cohen and D.W.Massaro, "Modeling Coarticulation in Synthetic Visual Speech." In M.Thalmann & D.Thalmann (Eds.) Computer Animation'93. Tokyo: Springer-Verlag.

[Cosatto98] Cosatto E., Graf H.P., "Sample-Based Synthesis of Photo-Realistic Talking Heads", Proc. Computer Animation '98, Philadelphia, USA, pp. 103-110.

[Doenges97] Doenges, P.K., Capin, T.K., Lavagetto, F., Ostermann, J., Pandzic, I.S., Petajan, E.D., "MPEG-4: Audio/Video & Synthetic Graphics/Audio for Mixed Media", *Image Communication Journal, Special Issue on MPEG-4*, Vol.9, No. 4, May 1997.

[Eisert97] P. Eisert, S. Chaudhuri and B. Girod, "Speech Driven Synthesis of Talking Head Sequences," 3D Image Analysis and Synthesis, pp. 51-56, Erlangen, November 1997.

[Fischl93] J. Fischl, B. Miller, J. Robinson, "Parameter tracking in a muscle–based analysis/synthesis coding system", Pic-ture Coding Symposium (PCS '93), Lausanne, Switzerland, No. 2.3, March 1993.

[Kalra92] 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

[Kalra93] Kalra P. "An Interactive Multimodal Facial Animation System", PhD Thesis nr. 1183, EPFL, 1993

[Kampmann98] M. Kampmann, B. Nagel:, "Synthesis of Facial Expressions for Semantic Coding of Videophone Sequences", Computer Graphics International (CGI98), Hannover, Germany, June 1998.

[Osterman97] Jörn Ostermann, Erich Haratsch, "An Animation Definition Interface - Rapid Design of MPEG-4 Compliant Animated Faces and Bodies", The International Workshop on Synthetic-Natural Hybrid Coding and 3D Imaging, September 5 - 9, 1997, Rhodes, Greece.

[Ostermann98] Ostermann J., "Animation of Synthetic Faces in MPEG-4", Proc. Computer Animation '98, Philadelphia, USA, pp. 103-110.

[Pandzic97] Igor S. Pandzic, Tolga K. Capin, Elwin Lee, Nadia Magnenat Thalmann, Daniel Thalmann, "A flexible architecture for Virtual Humans in Networked Collaborative Virtual Environments", Proceedings Eurographics 97, Budapest, Hungary, 1997.

[Parke97] Parke F.I., Waters K., "Computer Facial Animation", A.K. Peters, Wellesley, Massachussets, 1997.

[Sproat95] R.Sproat and J.Olive, "An Approach To Text-to-Speech Synthesis." In W.B.Kleijn & K.K.Paliwal (Eds) Speech Coding and Synthesis, Elsevier Science, 1995.

[Terzopoulos93] 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, pp. 569-579, 1993.

ICSLP 99

INTEGRATION OF TALKING HEADS AND TEXT-TO-SPEECH SYNTHESIZERS FOR VISUAL TTS

Jörn Ostermann[1], Mark Beutnagel[1], Ariel Fischer[2], Yao Wang[3]

[1]AT&T Labs Research, [2]Institute Eurecom/EPFL, [3]Polytechnic University

ABSTRACT

The integration of text-to-speech (TTS) synthesis and animation of synthetic faces allows new applications like visual human computer interfaces using agents or avatars. The TTS informs the talking head when phonemes are spoken. The appropriate mouth shapes are animated and rendered while the TTS produces the sound. We call this integrated system of TTS and animation a Visual TTS (VTTS). This paper describes the architecture on an integrated VTTS synthesizer that allows defining facial expressions as bookmarks in the text that will be animated while the model is talking.

The position of a bookmark in the text defines the start time for the facial expression. The bookmark itself names the expression, its amplitude and the duration during which the amplitude has to be reached by the face. A bookmark to face animation parameter (FAP) converter creates a curve defining the amplitude for the given FAP over time using Hermite functions of 3rd order.

1. INTRODUCTION

With the new generation of Text-to-Speech (TTS) synthesizers creating human like and pleasant voices, the integration of text-to-speech (TTS) synthesis and animation of synthetic faces allows new applications like visual human computer interfaces (HCI), information kiosks or networked applications like virtual sales agents and virtual company representatives. We name such a system Visual TTS (VTTS). In a VTTS system, the speech synthesizer informs the face renderer about the phonemes of the spoken text and the related timing information. A phoneme to viseme converter computes appropriate mouth shapes that generate the impression of a talking head when rendered while the TTS produces the sound. TTS and face animation are areas covered by the upcoming MPEG-4 standard.

The goal of MPEG-4 is to provide a new kind of standardization that responds to the evolution of technology, when it does not always make sense to specify a rigid standard addressing just one application. MPEG-4 will allow the user configuring and building systems for many applications by allowing flexibility in the system configurations, by providing various levels of interactivity with audio-visual content of a scene, and by integrating as many as possible audio visual data types like natural and synthetic audio, video and graphics [1][2][3]. MPEG-4 will become an International Standard in spring 1999, just in time for the new faster and more powerful media processors and in time for using the upcoming narrow- and broadband wired and wireless networks for audio-visual applications like database browsing, information retrieval and interactive communications

As far as synthetic multimedia contents are concerned, MPEG-4 will provide synthetic audio like structured audio and a text-to-speech interface (TTSI). For synthetic visual contents, MPEG-4 allows to build 2D and 3D objects composed of primitives like rectangles, spheres, indexed facesets and arbitrarily shaped 2D objects. The 3D-object description is based on a subset of VRML nodes [4] and extended to enable seamless integration of 2D and 3D objects. Objects can be composed into 2D and 3D scenes using the BInary Format for Scenes (BIFS). BIFS also allows to animate objects and their properties.

Special 3D objects are human faces and bodies. MPEG-4 allows using decoder resident proprietary models as well as to transmit 3D models to the decoder such that the encoder can predict the quality of the presentation at the decoder [3]. The integration of TTS and facial animation is currently limited. Non-speech related animation parameters are transmitted using a synchronous stream. Since the timing of the TTS is unknown to the sender, synchronization of facial expressions with mouth shapes and sound cannot be achieved. Here, we propose an architecture that allows driving a face model including its facial expressions based on MPEG-4 face animation parameters (FAP) from the text input of the TTS using a bookmark mechanism and an interpolation function to derive the amplitude of the facial expressions over time given the bookmarks.

In Sections 2, we explain how MPEG-4 defines the specification of a face model and its animation using FAPs. Section 3 shows the proposed architecture to combine face animation with text-to-speech capabilities. Section 4 describes different interpolation functions for FAP amplitudes.

2. FACE ANIMATION IN MPEG-4

MPEG4 specifies a set of face animation parameters (FAPs), each corresponding to a particular facial action deforming a face model in its neutral state. The FAP value for a particular FAP indicates the magnitude of the corresponding action, e.g., a big versus a small smile. Deforming the face model in its neutral state according to the specified FAP values for the corresponding time instant generates a particular facial action sequence. Then the model is rendered onto the screen.

The head in its neutral state is defined as follows (Figure 1): Gaze is in direction of Z axis; all face muscles are relaxed; eyelids are tangent to the iris; the pupil is one third of IRISD0; lips are in contact; the line of the lips is horizontal and at the same height of lip corners; the mouth is closed and the upper teeth touch the lower ones; the tongue is flat, horizontal with the tip of tongue touching the boundary between upper and lower teeth.

For the renderer to interpret the FAP values using its face model, the renderer has to have predefined model specific animation rules to produce the facial action corresponding to each

FAP. Since the FAPs are required to animate faces of different sizes and proportions, the FAP values are defined in face animation parameter units (FAPU). FAPU are defined as fractions of distances between key facial features (Figure 1). These features like eye separation, eye-nose separation, mouth nose separation, and mouth width, are defined for the face in its neutral state. They allow interpretation of the FAPs on any facial model in a consistent way, producing reasonable results in terms of expression and speech pronunciation.

MPEG-4 defines the animation rule for each FAP by specifying feature points and their direction of movement. The renderer can either use its own animation rules for his proprietary model or download a face model and its FaceDefTables that define the animation rules for the model.

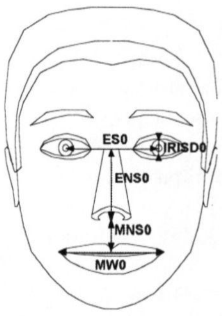


Figure 1: FAPUs [2].

2.1. Face Feature Points

In order to define face animation parameters for arbitrary face models, MPEG-4 specifies 84 feature points located in a face according to Figure 2 in order to provide a reference for defining facial animation parameters. The location of these feature points has to be known for any MPEG-4 compliant face model.

2.2. Face Animation Parameters

The FAPs are based on the study of minimal perceptible actions and are closely related to muscle action [5]. The 68 parameters are categorized into 10 groups related to parts of the face. FAPs represent a complete set of basic facial actions including head motion, tongue, eye, and mouth control. They allow the representation of natural facial expressions. They can also be used to define facial action units [6]. Exaggerated values permit the definition of actions that are normally not possible for humans, but are desirable for cartoon-like characters.

The FAP set contains the two high-level parameters visemes and expressions. A viseme is a visual correlate to a phoneme. Only 14 static visemes that are clearly distinguished are included in the standard set (Table 1). The expression parameter defines 6 high level facial expressions like joy and sadness (Figure 3). In contrast to visemes, facial expressions are animated with a value defining the amplitude of the expression. Two facial expressions can be blended with a weighting factor. Since expressions are high-level animation parameters, they allow animating unknown models with high subjective quality.

2.3. Face Model Specification

MPEG-4 allows the encoder to completely specify the face model the decoder has to animate. This involves defining the static geometry of the face model in its neutral state using a scene graph and defining the animation rules that specify how this model gets deformed by the facial animation parameters [8].

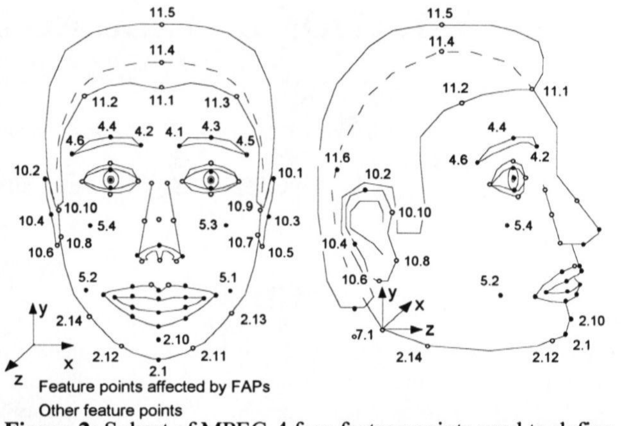


Figure 2: Subset of MPEG-4 face feature points used to define the shape of a proprietary face model. FAPs are defined by motion of feature points [2].

Table 1: Visemes and related phonemes.

#	phoneme	example	#	phoneme	example
1	p, b, m	put, bed, mill	8	n, l	lot, not
2	f, v	far, voice	9	r	red
3	T,D	think, that	10	A:	car
4	t, d	tip, doll	11	e	bed
5	k, g	call, gas	12	I	tip
6	tS,dZ,S	chair, join, she	13	Q	top
7	s, z	sir, zeal	14	U	book

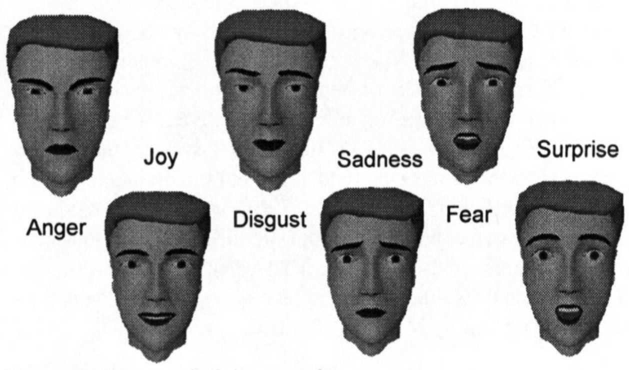


Figure 3: Primary facial expressions.

2.4. Integration with TTS

MPEG-4 acknowledges the importance of text-to-speech (TTS) synthesis for multimedia applications providing an interface to proprietary text-to-speech synthesizer (TTSI). A TTS stream contains text in ASCII and optional prosody in binary form. The decoder decodes the text and prosody information according to the interface defined for the TTS synthesizer. The synthesizer creates speech samples that are handed to the compositor. The compositor presents audio and if required video to the user.

In the current MPEG4 standard, the encoder is expected to send a FAP stream containing FAP number and amplitude for every frame, to enable the receiver to produce desired facial actions (Figure 4). Since the TTS synthesizer can behave like an asynchronous source, synchronization of speech parameters with facial expressions of the FAP stream is usually not given – unless the encoder transmits prosody with timing information for the synthesizer.

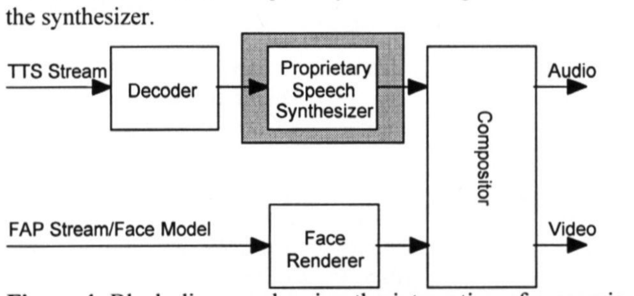


Figure 4: Block diagram showing the integration of a proprietary Text-to-Speech Synthesizer into an MPEG-4 face animation system.

3. ARCHITECTURE FOR VTTS

Figure 5 shows the architecture of the proposed VTTS that allows synchronized presentation of synthetic speech and talking heads. A second output interface is added to the TTS. This interface sends the phonemes of the synthesized speech as well as start time and duration information for each phoneme to a Phoneme/Bookmark-to-FAP-Converter. The converter translates the phonemes and timing information into face animation parameters that the face renderer uses in order to animate the face model [7][9]. In addition to the phonemes, the synthesizer identifies bookmarks in the text that convey non-speech related facial animation parameters to the face renderer. The timing information of the bookmarks is derived from their position in the synthesized speech. Since now the facial animation is driven completely from the text input to the TTS, there is no need to transmit an FAP stream to the decoder. Furthermore, synchronization is achieved since the talking head is driven by the speed of the asynchronous proprietary TTS synthesizer.

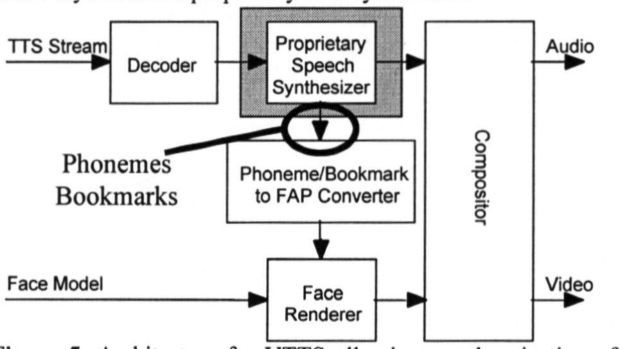


Figure 5: Architecture for VTTS allowing synchronization of facial expressions and speech.

In order to allow for simple bookmarks, each bookmark has to describe for one FAP at a time the transition from the current FAP amplitude to a target FAP amplitude. Simply applying an FAP of constant amplitude and resetting it after a certain amount of time does not allow for realistic face motion. Therefore, we propose that the Bookmark to FAP Converter creates the appropriate transition between current amplitude and the target amplitude. There are 2 ways of designing bookmarks:

1. The position of the bookmark defines the amplitude of the FAP at the time instant of the spoken word. Consequence: In order to generate smooth temporal behavior of the FAP the decoder has to look *ahead* into the TTS stream in order to determine an appropriate behavior. This increases the delay of the decoder.
2. The bookmark defines the *start* point and *duration* of the transition to a new FAP amplitude. Consequence: No additional delay, no look ahead in the bitstream but no precise timing control on when the amplitude will be reached relative to the spoken text.

In our tests we did not find a problem with using option 2 since the transition times for facial expressions is usually less than 1s. As syntax for a bookmark, we use <FAP *n* (*s*) *a* *T*> with FAP number *n*, expression *s* in case *n* equals 2 (Figure 3), the amplitude *a* and the transition time *T* in ms.

4. INTERPOLATION FUNCTIONS

The FAP amplitude a defines the amplitude to be applied at the end of the transition time T. The amplitude a_s of the FAP at the beginning of the transition depends on previous bookmarks and can be equal to:

- 0 if the FAP bookmark is the first one with this FAP n.
- a of the previous FAP bookmark with the same FAP n if a time longer than the previous transition time T has elapsed between these two FAP bookmarks.
- The actual reached amplitude due to the previous FAP definition if a time shorter than the previous transition time T has elapsed between the two FAP bookmarks.

At the end of the transition time T, a is maintained until another FAP bookmark gives a new value to reach. To reset an FAP, a bookmark for FAP n with a=0 is transmitted in the text.

To avoid too many parameters for defining the evolution of the amplitude during the transition time, the function that computes for each frame the amplitude of the FAP to be sent to the face renderer is predefined. Assuming that the transition time T is always 1, we implemented the following functions $f(t)$:

$$f(t) = a_s + (a - a_s)t \tag{1}$$

$$f(t) = a_s + (1 - e^{-t})(a - a_s) \tag{2}$$

$$f(t) = a_s + \left(1 - e^{-\lambda(t-1/2)}\right)^{-1}(a - a_s) \tag{3}$$

$$f(t) = (2t^3 - 3t^2 + 1)a_s + (-2t^3 + 3t^2)a + (t^3 - 2t^2 + t)g_s \tag{4}$$

with time $t \in [0,1]$, the amplitude a_s at the beginning of the FAP at t=0, control parameter λ and the gradient g_s of $f(0)$ which is the FAP amplitude over time at t=0. If the transition time $T \neq 1$, the time axis of the functions (1) to (4) has to be scaled. These functions depend on a_s, g_s, a and T, and thus they are completely determined as soon as the FAP bookmark is known. After extensive subjective evaluations, it turns out that the Hermite function of third order (4) gives the best results, in terms of realistic behavior. Using Splines with more than one Hermite

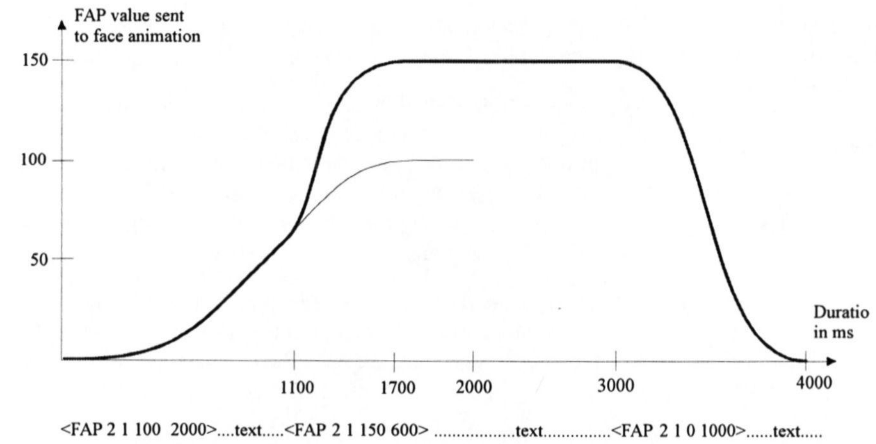


Figure 6: Amplitude of joy (FAP 2 1) as defined by the bookmarks.

segment would increase the flexibility for designing curves but would also require to have some knowledge of bookmarks placed further in the text than the current bookmark which is a significant drawback for a real-time system.

The Hermite function of third order enables one to match the tangent at the beginning of a segment with the tangent at the end of the previous segment, so that a smooth curve can be guarantied. Usually, the computation of the Hermite function requires 4 parameters as input, which are a_s, g_s, a and the gradient of $f(t)$ at t=1. In our implementation we assume a horizontal gradient at the end of the transition time. Figure 6 shows an example of a time curve created with 3 bookmarks for FAP 2 (expression) and expression 1 (joy). As can be seen, the gradient g_s at the beginning of the transition time is 0 for the first and the third bookmark. The gradient g_s for the second bookmark at time t is computed according to

$$g_s(t) = (6t^2 - 6t)(a_s - a) + (3t^2 - 4t + 1)g_s \quad (5)$$

with a_s, g_s ,and a defined by the first bookmark and $g_s(t)$ defining the starting gradient g_s for the second bookmark.

5. Conclusions

MPEG-4 integrates animation of synthetic talking faces into audio-visual multimedia communications. A face model is a representation of the human face that is structured for portraying the visual manifestations of speech and facial expressions adequate to achieve visual speech intelligibility and the recognition of the mood of the speaker. A face model is defined as a static 3D model and related animation rules that define how the model deforms if it is animated with FAPs. The model is defined using a scene graph. Therefore, a customized model with head and shoulders can be defined for games or web-based customer service applications. MPEG-4 defines a complete set of animation parameters tailored towards animation of the human face. Face animation parameters are defined independent of the proportions of the animated face model. Therefore, a face animation parameter stream can be used to animate different models. Successful animations of humans, animals and cartoon characters have been demonstrated.

MPEG-4 defines interfaces to include proprietary TTS synthesizer in an MPEG-4 multimedia application. However, MPEG-4 does not yet provide sufficient means to synchronize a face model with a TTS synthesizer.

Here, we propose an architecture that controls a talking head completely using the text input of the TTS. In addition to the input interface and the sound output interface as defined by MPEG-4 we define an interface that exports the phonemes and their timing from the TTS. In addition, the TTS is extended to be able to recognize bookmarks in the text. These bookmarks are also exported with their timing derived from the words between which it is located. A phoneme/bookmark converter translates the phonemes and bookmarks into appropriate sequences of facial animation parameters that a rendered such that speech and animation are synchronized.

One bookmark defines one facial expression, its amplitude and a transition time after which this amplitude has to be reached. Subjective evaluation showed that a Hermite function of 3rd order allows creating realistic animation of facial expressions.

This proposal for synchronizing TTS with facial expressions is currently considered by MEG-4.

6. References

[1] ISO/IEC JTC1/WG11 N2201, "Text for FCD 14496-1 Systems", *Tokyo meeting, March 1998.*

[2] ISO/IEC JTC1/WG11 N2202, "Text for FCD 14496-2 Visual", *Tokyo meeting, March 1998.*

[3] J. Ostermann, "Animation of synthetic faces in MPEG-4", *Computer Animation 98, pp. 49-55, Philadelphia, June 1998.*

[4] J. Hartman, J. Wernecke, *The VRML handbook*, Addison Wessley, 1996.

[5] Kalra P., Mangili A, Magnenat-Thalmann N, Thalmann D. "Simulation of Facial Muscle Actions Based on Rational Free Form Deformations", *Proc. Eurographics 92, pp. 59-69, 1992.*

[6] P. Ekman, W.V. Friesen, *Manual for the facial action coding system*, Consulting Psychologist Press, Inc. Palo Alto, CA, 1978.

[7] M. M. Cohen and D. W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech," In M. Thalmann & D. Thalmann (Eds.) *Computer Animation "93*, Tokyo: Springer-Verlag.

[8] J. Ostermann, E. Haratsch, "An animation definition interface: Rapid design of MPEG-4 compliant animated faces and bodies", *International Workshop on synthetic - natural hybrid coding and three dimensional imaging, pp. 216-219, Rhodes, Greece, September 5-9, 1997.*

[9] K. Waters, T. Levergood, "An automatic lip-synchronization algorithm for synthetic faces", *Proceedings of the Multimedia Conference, ACM, pages 149-156, San Francisco, California, September 1994.*

Visual Computer

User evaluation: Synthetic talking faces for interactive services

Igor S. Pandzic*, Jörn Ostermann, David Millen

AT&T Labs – Research, Room 3-231, 100 Schultz Dr., Red Bank, NJ, 07701, USA
e-mail: {osterman,drm}@research.att.com, Igor.Pandzic@cui.unige.ch

Computer simulation of human faces has been an active research area for a long time. However, it is less clear what the applications of *facial animation* (FA) will be. We have undertaken experiments on 190 subjects in order to explore the benefits of FA. Part of the experiment was aimed at exploring the objective benefits, i.e., to see if FA can help users to perform certain tasks better. The other part of the experiment was aimed at subjective benefits. At the same time comparison of different FA techniques was undertaken. We present the experiment design and the results. The results show that FA aids users in understanding spoken text in noisy conditions; that it can effectively make waiting times more acceptable to the user; and that it makes services more attractive to the users, particularly when they compare directly the same service with or without the FA.

Key words: Facial animation – Talking head – Animated characters – Nonverbal communication – Subjective tests – User evaluation – Interactive services

* This work was conducted while Igor Pandzic worked at AT&T Labs-Research during the summer of 1998. Currently he works at MIRALab, University of Geneva, 24 rue du General Dufour, 1211 Geneva, Switzerland

1 Introduction

Computer simulation of human faces has been an active research area for a long time, resulting in a multitude of facial models and several animation systems (Kalra et al. 1992; Fischl et al. 1993; Kalra 1993; Terzopoulos and Waters 1993; Ostermann and Haratsch 1997; Parke and Waters 1997; Eisert et al. 1997; Cosatto and Graf 1998; Kampmann and Nagel 1998). Current interest for this technology is clearly shown by its inclusion in the MPEG-4 standard (Ostermann 1998; Doenges et al. 1997; MPEG-N2501; MPEG-M2502; MPEG-N2503).

The advances in animation systems, such as those mentioned above, have prompted interest in the use of animation to enrich the human-computer interface. One important application of animated characters has been to make the 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 (Andre et al. 1998; Rist et al. 1997). Animated guides or assistants have also been used with some success in user help systems (Don et al. 1993; Gibbs and Breiteneder 1993), and for user assistance in web navigation (Milewski and Blonder 1996). Personal character animations have also been inserted, with some success, into documents to provide additional information to readers (Bickmore et al. 1998).

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 (The Palace; Suler 1997; Pandzic et al. 1997). 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 (Damer et al. 1996). Nevertheless, the enormous popularity of Internet chat applications suggests considerable future use of avatars in social communication applications.

The Visual Computer (1999) 15:330–340

The use of *facial animation* (FA) in interface design has been the primary research focus of several studies of multi-modal interfaces. One important area of inquiry has been the nature of the social interaction for applications that use facial animation. In one interview task, researchers found that users revealed more information, spent more time responding, and made fewer mistakes when interacting with an animated facial display compared with a traditional paper and pencil questionnaire (Walker 1994). Furthermore, the increased responsive effects were greater when a *stern* facial animation was used compared to a more neutral face. In a second study, subjective reports about an interview with an animated face revealed that users attributed personality attributes to the face, reported that they were more alert, and presented themselves in better light compared with an interview using only text (Sproull 1996). Finally, users exhibited a degree of cooperation when interacting with animated partners in a social dilemma task. The fact that the animated face was *human* was important, as the same cooperative interaction was not observed using animated faces of dogs (Parise 1996).

To understand further the utility and usability of a facial display interface, we have completed several experiments. In the first experiment, we consider the benefits of a facial display as a distinct channel in a multi-modal interface. In this experiment we explore the performance benefits of using FA in a number intelligibility task. It was expected that FA synchronized with speech would result in better performance over speech alone in a noisy ambient environment.

In the second and third experiments, we tested user performance and preferences in a kiosk application across a variety of interface conditions. These experiments were intended to explore the more subjective benefits of FA displays, such as increasing the task interest and appeal, and minimizing the negative aspects of system delays.

In all three experiments, several different FA techniques were used. The results, therefore, provide a preliminary study of the performance and preferences of different FA techniques.

In the next section we present the experiment design, describing in detail the technical setup, the experimental tasks for different experiments and the subjects. In Sect. 3 we present the detailed results of the experiments followed by a summary of most important results. Finally, we give conclusions and discuss issues for further study.

2 Experiment design

Three experiments were undertaken, each examining different aspects of FA.

Experiment 1 was primarily aimed at measurable effects of FA (Fig. 1a,c), rather then based on subjects' evaluation of certain criteria (though a questionnaire was also used as a second source of information). The measurement was performed by observing how well the subjects can perform a task with or without FA, and under different conditions.

The following effects have been explored:

- Effect of FA on speech understanding in optimal acoustic conditions
- Effect of FA on speech understanding in noisy conditions
- Effect of changing FA frame rate on speech understanding in both noisy and optimal conditions
- Effect of changing FA techniques on speech understanding in both noisy and optimal conditions

Experiment 2 was aimed at more subjective benefits of FA (Fig. 1a): the general appeal to the user, making a service more friendly, filling the waiting times, and in general improving the users' satisfaction. For this purpose a simple service with a limited scope was conceived and the subjects were asked to use it and then to answer questions related to their level of satisfaction with the service. The response to the service with and without facial animation and synthetic voice has been compared.

Experiment 3 was a preliminary study into comparison of different methods to generate synthetic faces. Three different synthetic faces (Fig. 1a–c) were set up to pronounce a simple welcome message, and the subjects were asked to compare and evaluate different faces.

2.1 Technical setup

A Text-To-Speech (TTS) system is coupled and synchronized with an FA system, yielding a Visual TTS (VTTS) system that simulates a talking head pronouncing arbitrary text in real time (Sproat 1995). The FA system is based on a 3D polygon mesh face model with defined facial actions allowing the simulation of speech and facial expressions like smiling, being angry etc. The coarticulation model is the one

Fig. 1. a standard face; **b** texture mapped face; **c** sample-based face

from (Cohen 1993). The facial model can be modified (Osterman 1997). The actual facial models used in the experiments are shown in Fig. 1a,b. The cartoon like character (Fig. 1a) was used in all experiments, the texture mapped character (Fig. 1b) was only used in Experiment 3.

Additionally, the sample-based FA system has been used in experiments 1 and 3 (Cosatto 1998). In this system a set of samples of the mouth area is extracted from a video of a real person talking. The samples are classified according to the mouth shape. This database of mouth shapes is used to generate appropriate mouth movement according to the speech pronounced by the TTS system. Functionally, this yields the same system as VTTS, but the visual result is different. The image looks much more realistic, however the lip movement looks less natural. Currently, this method does not run in real time, therefore the utterances needed for the experiment were recorded offline and shown as video clips during the experiment. Figure 1c shows the sample-based face.

Figure 2 shows the physical setup at the experiment site. Two workstations were used in order to increase the capacity. Subjects were using headphones for better control of the acoustic environment.

2.2 Experimental task

2.2.1 Experiment 1

The subject's task was to listen to several series of numbers (digits), and type them in (Fig. 3). There were five numbers per series. The numbers were pronounced by the TTS system. The subject could type the number in only when each series of five numbers was fully spoken out. The error rate was measured on a digit-by-digit basis. The subjects were given two trial series of five digits each, then ten measurement series of five digits each. Each subject repeated the task in noisy and optimal acoustic conditions. The order of noisy and optimal condition was randomized and different for each subject. In the noisy conditions, the Signal to Noise Ratio was −2 dBA. This is a very difficult hearing condition, corresponding roughly to talking on the phone in a noisy airport while a flight announcement is heard from a nearby loudspeaker. Such difficult conditions were chosen in order to have a significant error rate and to be able to measure improvements when FA is deployed.

The subjects were split into four groups, each group having different visual conditions. The summary of different visual conditions is shown in Table 1.

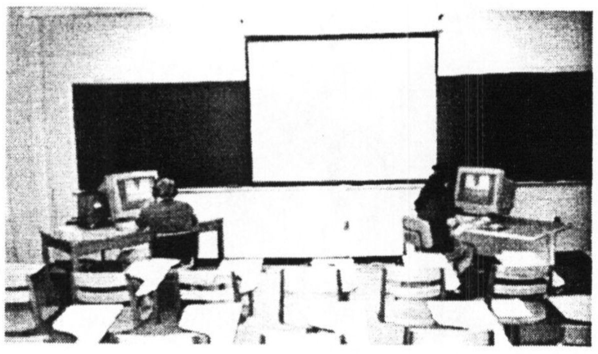


Fig. 2. The physical setup at the experiment site

Table 1. Summary of different visual conditions in experiment 1

Condition name	Face rendering	Frame rate (Hz)
No face	none	–
Low frame rate	3D	10
Standard face	3D	18
Sample-based face	sample based	30

After each test the subjects were given a questionnaire. Following questions were asked:

- Please estimate the time it took you to complete the test.
- Understanding the numbers was: (level of difficulty)
- The sound quality of the speech pronouncing the numbers was: (level of quality)
- Was the video of the face useful?
- Was the video of the face distracting?

For each question an appropriate six-point scale of answers was offered to the subject.

2.2.2 Experiment 2

The subjects were asked to use a simple interactive real-time system giving information about theatre shows. The service was conceived in such a way that it can perfectly be used without FA. FA is just a gadget to make the service more engaging and friendly. The service involves waiting time (simulating Internet and server access waiting times) that is filled by FA and/or speech synthesis. The face acts as the representative of the service, welcomes the users and asks questions about what they would like to find out. Expressions (smiles) were used in an effort to make the face more pleasing.

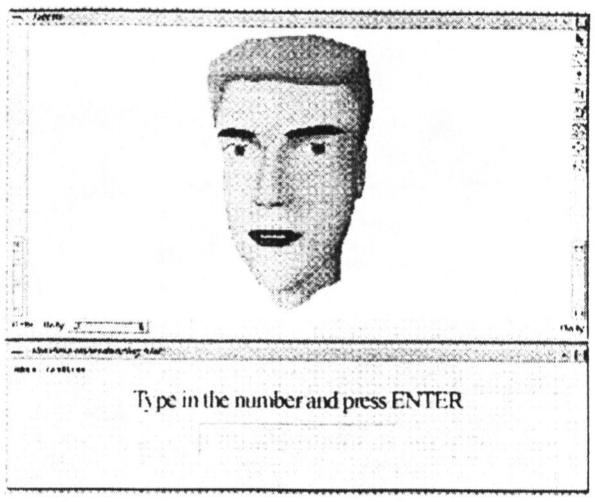


Fig. 3. Screen shot from experiment 1

The service starts with a welcome message, then gives the user a choice of Broadway shows (Fig. 4). The user chooses a show, and is presented with a choice of available information about the show: review, venue and prices. When the user has chosen which information he/she wants, there is a waiting time before the information is actually displayed, simulating the waiting times on the Internet. After reading the desired information, the user can choose to get more information about the same show, to get information about another show or to exit.

To insure that subjects spend sufficient time using the system, they were asked to choose a theatre show and find some information about it: the review (was it good or bad?), venue and the ticket price. They were given a data sheet where they had to write down this information. This insured that the subjects went through all features of the system.

The experiment was performed in varying conditions with respect to the presence of the visual and acoustic stimuli (FA and TTS). In addition, one group of users was tested using a text-only version of the interface. The time spent using the system was measured in order to compare with the subject's estimate of the time spent with the system. The duration of each test was under 5 minutes.

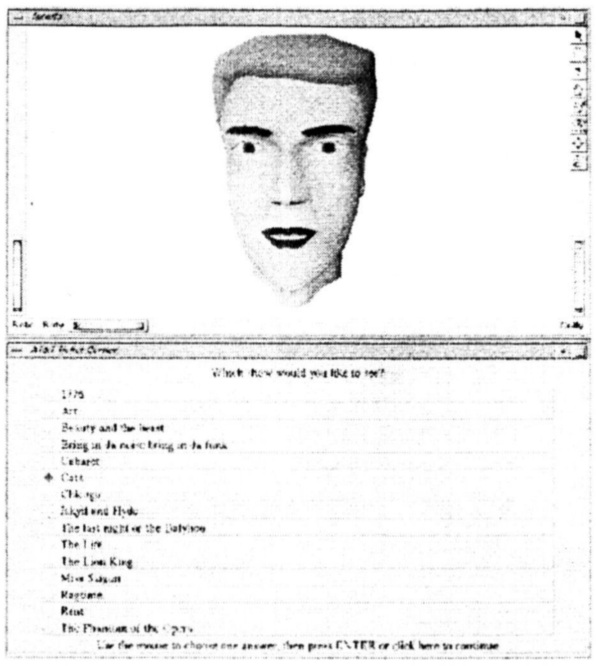


Fig. 4. Screen shot from experiment 2

The questionnaire was used to interrogate the subjects on usefulness of the system, usefulness of FA, friendliness of the system, ease of use, perceived sound quality and possible distraction or annoyance by the face animation.
More specifically, the questionnaire contained the following interrogations:

- Please estimate the time it took you to complete the trial.
- Overall, how satisfied or dissatisfied are you with this service?
- Overall, how satisfied or dissatisfied are you with the speed of this service?
- Is the service easy to use?
- Is the service user-friendly?
- The sound quality of the speech was: (scale)
- How positive or negative are your feelings or emotional reactions when using this service?
- How human-like or computer-like did you find this service?
- Was the video of the face useful?
- Was the video of the face distracting?
- Was the face friendly?
- Did the face look at you?
- Identify the part of the service with the longest waiting time. (One part of the service has a deliberately longer waiting time, we wanted to see if FA can hide it.)

For each question an appropriate scale/choice of answers was offered.

2.2.3 Experiment 3

The subjects viewed/listened to short audio-visual sequences showing a face pronouncing a welcome message: "Welcome to AT&T global communication services". The welcome message was pronounced by three different synthetic faces: non-textured 3D model, textured 3D model, and sample-based model, as illustrated in Fig. 1. No facial expressions (smiles or other) were used on the synthetic face. The three sequences were shown to the user in random order. After the first showing, all three sequences were shown a second time (in the same order as the first time), and this time after each sequence the subject was asked how he/she liked that particular welcome message. An appropriate scale was offered for the answer to this question.
The purpose of this experiment was to compare different approaches to FA and also see to what extent users are sensitive to artifacts in mouth movement that may occur in some of the approaches.

2.3 The subjects

A total of 190 subjects have completed the experiment at Princeton University in June 1998. The subjects were either students or employees of Princeton University. For experiment 1, five subjects were later discarded because they had hearing problems. Further 40 subjects were discarded from experiment 1 due to later-discovered technical problems. Finally, for experiment 1 145 subjects were used, and for the two other experiments all 190 subjects were used. The box-plots[1] in Fig. 5 show the distribution of subjects' age and computer usage habits, extracted by means of

[1] Box-plots are used in this paper. The box represents the middle half of the results, i.e., the lower bound of the box is the 1st quartile and the upper bound is the 3rd quartile. The whiskers above and under the box show the bounds of the data, with any outliers plotted as simple lines outside the bounds

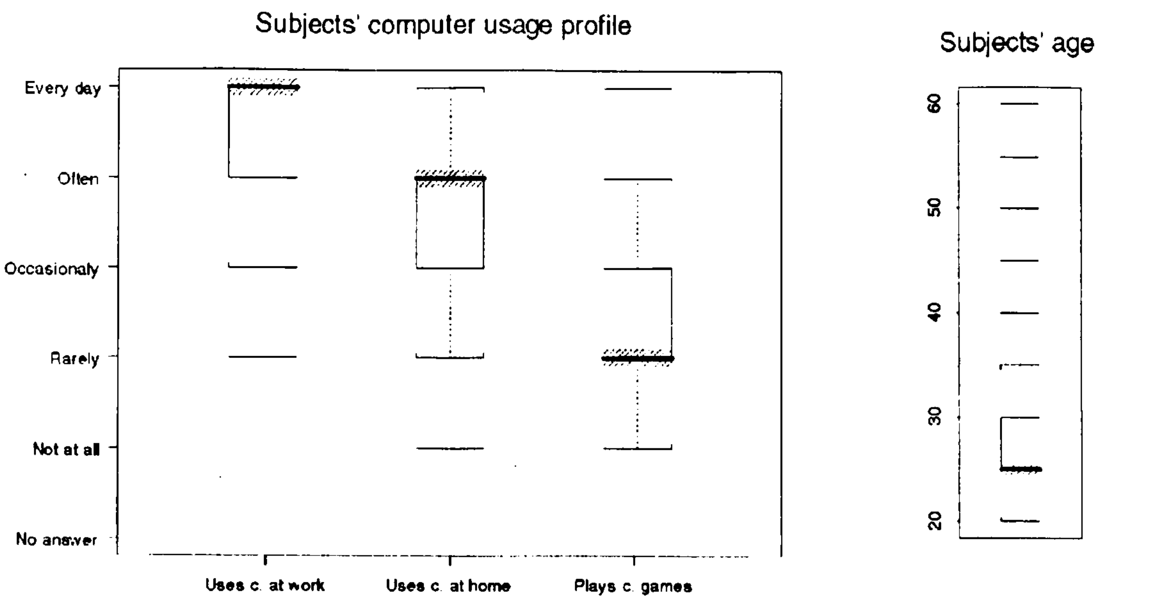


Fig. 5. Subjects' profile charts

a questionnaire. Additionally, the questionnaire has shown that 39% of the subjects were not native English speakers.
It can be observed that the subjects were in general young, very frequent computer users, and relatively often not native English speakers.

3 Results

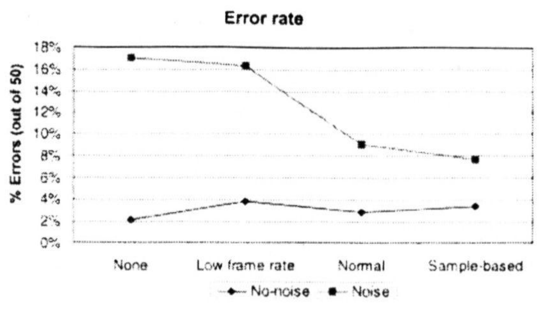
In this section we present and analyze the results of all three experiments. Significance tests [ANOVA and Scheffe Post Hoc tests] were performed for each experiment, for each of the performance variables (i.e., error rates and completion times) and attribute ratings. The observed results that are significantly different from chance (i.e., probability values less than .05) will be reported.

3.1 Experiment 1

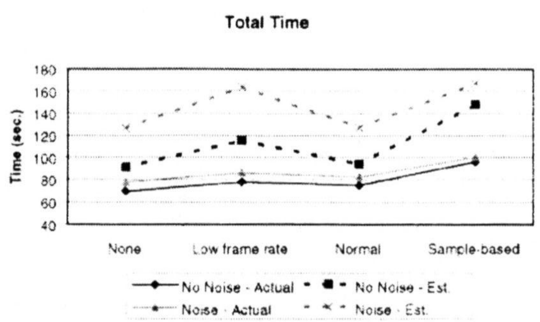
In the following subsections we present the error rate and timing results in all conditions, as well as the subjective responses collected in the questionnaire.

3.1.1 Error rates

Figure 6 shows the error rates in all conditions. The most obvious observation is that there are much more errors in noisy conditions, as expected ($p < .001$).
There was an interesting difference in error rates as a function of noise and presentation condition ($p < .001$). In optimal acoustic conditions (no noise) there is no significant difference in error rates between different visual conditions, i.e., all subjects made very few mistakes when no noise was present. However, in noisy conditions significant differences can be observed between different visual conditions. Subjects doing the experiment without the face, or with a low-frame-rate face did much worse then those with the standard or sample-based face, with mean error rates dropping from approximately 16% to approximately 8%.
No significant difference is observed between low-frame-rate face and no face at all. This suggests that 10 Hz is not a high enough frame rate to provide a useful visual speech pronunciation. An increase in frame rate to 18 Hz (normal face) provides obvious improvement in error rates; however, another study would be needed to determine: a) where between



6



7

Fig. 6. Average error rates; comparison between different conditions with and without noise

Fig. 7. Measured vs. estimated time to complete task in noisy and normal conditions

10 Hz and 18 Hz lies the limit of usefulness; and b) does further increase in frame rate improve understanding even more. Although the sample-based face was played at 30 Hz, the face synthesis method is different and this cannot be used for direct comparison in terms of frame rate.

The fact that there is no significant difference between the standard 3D face and sample-based face is somewhat surprising because both the experts and the subjects agree that the sample-based face shows jerkiness in lip movement and lower quality of lip synchronization. Therefore it was expected to show higher error rates than the standard 3D face. One possible explanation of these results is that the sample-based face was played at 30 Hz vs. 18 Hz for the standard 3D face. If we assume that higher frame rate yields better results, this may have compensated for the artifacts in the sample-based face.

3.1.2 Timing

The time needed to perform each test was measured. After the test the subjects were asked to estimate the time spent to do the test. Measured times, and the difference between the estimated and measured times (estimation error) are shown in graphs in Fig. 7.

The most pronounced effect is the small (approx. 10%) increase in experiment time when a face is presented with respect to doing the test without the face ($p < .001$). This effect is equally pronounced in noisy and optimal acoustic conditions. This may indicate that people get slightly distracted from their main task by the presence of the face. Viewed together with error rate results, it can be remarked that in non-noisy conditions slight increase in time does not bring any improvement in performance. However, in noisy condition the time loss is compensated by a substantial error rate decrease.

Another observation is that the subjects spent slightly more time doing the test in noisy conditions ($p < .001$). That can be explained by more concentration and some hesitations at the moment of entering the numbers.

While the real increase in time when noise is introduced is slight, the increase in subjects' estimated time is more substantial. Figure 7 indicates that subjects tend to overestimate their time in noisy conditions, while being rather accurate in estimation when there is no noise.

3.1.3 Subjective responses

As expected, subjects rated several interface attributes lower in the noise condition compared with the "no noise" condition (See Table 2). In particular, subjects rated both the sound quality and the ease of understanding reliably lower in the noise condition. Furthermore, subjects rated the presence of facial animation to be more useful and less distracting in the noise condition.

The "ease of understanding" ratings were similarly high across all presentation conditions, which can be seen in Table 3. On the other hand, there were several aspects of the task that varied as a function of the presentation condition. As can be seen in Table 3, the 3D face, both in the low-frame-rate and normal conditions, was found somewhat less distracting and more useful than the sample-based face. The normal 3D face was also rated slightly more useful than the low frame rate face.

	No noise	Noise	p value
Ease of understanding (6=easiest)	5.5	3.5	$p < .001$
Sound Quality Ratings? (6=best)	5.0	3.4	$p < .001$
Was the face useful? (6=most useful)	2.4	3.2	$p < .001$
Was the face distracting? (6=least distracting)	4.6	5.0	$p < .05$

Table 2. Average subject attribute ratings for noise and no noise conditions

	No face	Low frame rate	Nor-mal	Sample-based	p value
Ease of understanding (6=easiest)	4.3	4.5	4.6	4.6	$p > .1$
Sound quality ratings? (6=best)	3.9	4.2	4.6	4.1	$p < .01$
Was the face useful? (6=most useful)	(NA)	3.0	3.2	2.3	$p < .001$
Was the face distracting? (6=least distracting)	(NA)	5.0	5.1	4.3	$p < .001$

Table 3. Subject attribute ratings across presentation conditions

It is interesting to note the discrepancy between the subjects' estimate of the usefulness of the face (Table 3) and its objective usefulness (Fig. 6). Although the subjects show better performance results (Fig. 6), they do not seem to attribute this improvement to the presence of the face, as the scores on the usefulness of the face are relatively low. Even more surprising is the comparison of usefulness scores of the low frame rate and sample-based faces (Table 3). Sample-based face obtained a substantially lower usefulness score despite the fact that objective results show the exactly opposite effect. This may suggest that the visual cues (lip reading) are used subconsciously.

Furthermore, Table 3 indicates that the subjects found the sound quality of the speech better with the standard 3D face then in other conditions, meaning that the presence of face positively influenced their perception.

3.2 Experiment 2

Except for the time measurements, all the results in experiment 2 are the subjects' responses to a questionnaire. We present and analyze the results related to timing, user satisfaction and some specific questions concerning the synthetic face.

3.2.1 Timing

Figure 8 shows results of time measurements and subjects' estimates of time. There is no significant difference between the different experimental conditions. The time estimates are not significantly in error. They may be due to the coarseness of the time scale users were offered for the answers (discrete scale with 1 minute intervals).

Table 4 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 audio support, and those having audio 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.

Even more noticeable is that subjects preconditioned with the experiment using the Audio+face or Audio show less satisfaction and more annoyance with the waiting of the Text-only service. This may be explained by the fact that they have already used the service with audio/face offering distraction and notice more the waiting times in the simple text-only service.

3.2.2 User satisfaction

Table 4 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. Audio and Audio+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, Text Only service is judged to be slightly easier to use than the others; however the ease of use for all conditions is judged so high that differences here are minimal.

Table 4. Subjects attribute ratings across presentation conditions

	Audio	Aud+ face	Text only	p value
Satisfaction with speed (6=best)	4.4	4.4	3.8	$p < .01$
Satisfaction with waiting time (6=best)	4.0	4.2	3.5	$p < .001$
Overall satisfaction (6=most satisfied)	4.4	4.6	4.3	$p > .1$
Was service user friendly? (6= best)	4.9	5.1	4.7	$p < .01$
Easy to use?	5.2	5.4	5.3	$p > .1$
Positive or negative emotions? (6=most positive)	4.3	4.3	4.2	$p > .1$
Was the service human-like?(6=most human-like)	2.7	3.0	2.4	$p < .05$
Estimate of sound quality (6=best)	4.5	4.7	(NA)	$p > .1$

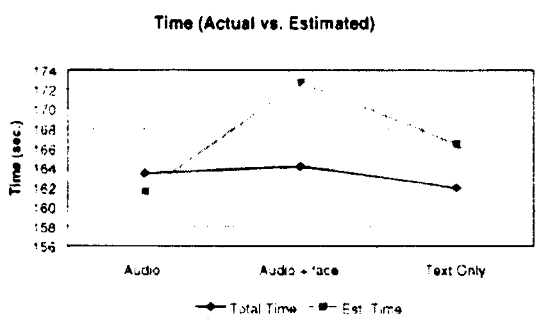


Fig. 8. Measured times and estimated times in Experiment 2

Another, weaker trend is for service with the face to be judged somewhat better than the one with audio; in particular it is judged to be more human-like.
The last row in Table 4 would show any influence of the presence of the face on the perceived sound quality. There is no such influence.

3.2.3 About the face

Several questions were asked about the synthetic face itself. The distribution of answers to these questions is given in Table 5. The figures indicate that the face was found to be fairly friendly and not distracting. It was found marginally useful (in average), which is good considering that it was deliberately designed and programmed in such a way that it is not useful at all.
When asked if the face looked at them 58% of subjects thought that it looked at them, 21% thought it did not, and 21% did not know. It was suspected that the population segment that perceived the face as looking at them might react more favorably to the face in general. However, the statistics did not show any such correlation in the results.

Table 5. User satisfaction (Experiment 2)

	Mean
Was the face friendly? (6=most friendly)	4.1
Was the face useful? (6=most useful)	3.2
Was the face distracting? (6=least distracting)	4.9

3.3 Experiment 3

In the third experiment the subjects were asked to compare different face models as illustrated in Fig. 1 by looking at a simple welcome message pronounced by each face and answering how they liked each presentation. The mean *appeal* ratings for synthetic faces were 5.0 for the standard face (Fig. 1a), 2.7 for the texture mapped face (Fig. 1b), and 3.3 for the sample-based face (Fig. 1c). The standard 3D polygon mesh facial model was clearly preferred by this group of users.
In general, it can be said that the subjects were not particularly seduced by synthetic faces. As for comparisons, the normal face (i.e., the 3D model without texture as shown in Fig. 1a) fared much better than the two other face models.

4 Conclusions

Experiments have been undertaken in order to examine the potential usefulness of Facial Animation (FA) combined with Text-To-Speech (TTS) technology for interactive services. Three experiments were run with a total of 190 subjects. The global goals of the experiments were the following:

- Examine FA potential to improve speech intelligibility.
- Examine whether FA can make interactive services more attractive.
- Examine whether FA can alleviate waiting times in services.
- Compare different FA techniques with respect to their appeal to subjects.

The results can be summarized in the following points:

- In optimal acoustic conditions FA does not help understanding; in case of significant artifacts of the mouth motion, it can slightly worsen the understanding.
- FA at 10 Hz does not help understanding in noisy conditions.
- FA at 18 Hz and sample-based FA at 30 Hz help understanding significantly in noisy conditions (error rate drop from 16% to 8%), though they also slightly increase the time the subjects (by less then 10%) spend on the task.
- In general the face is not found distracting.
- In general the face is not perceived as very useful, even when results (e.g., error rate) show that it actually is useful.
- FA can distract users during waiting times and make the waiting time appear shorter. However, TTS (audio) has a very similar effect even without FA.
- People react more positively to a service with FA then without; this difference is even more pronounced when users try the service without FA after trying the one with FA.
- A service with FA is considered more human-like and provokes more positive feelings than the one with TTS (audio) only (both by 1 point on a six-point scale).

It can be concluded that a wider deployment of FA may be worthwhile in interactive services, making them more attractive to users. It will be of interest to learn whether the level of satisfaction with the plain-text service drops significantly after the users have been introduced to the same service with FA enhancement.

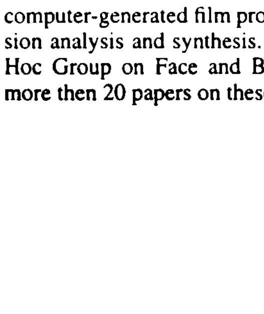
Current facial models and animation techniques are still rather crude and technical, which is to a good extent the result of their being created by engineers rather then artists. Experiment 3 has shown that the general appeal level of all used faces is rather low. It is therefore expected that the results in terms of user appeal may improve greatly by making the faces/animations more attractive.

Acknowledgements. The authors would like to thank Mike Orchard, Yao Wang and Michelle Young of Princeton University for their help in organizing the trials at the University. The authors also wish to thank Eric Cosatto and Hans-Peter Graf of AT&T Labs for the enlightening discussions about the experiment setup and their help with the inclusion of the sample-based face model in the trials. Laurie Garrison of AT&T Labs has provided valuable help in measuring the Signal to Noise Ratio.

References

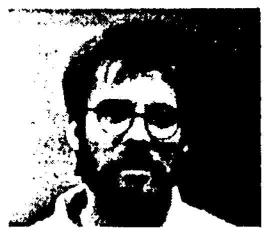
1. Andre E, Rist E, Muller J (1998) Guiding the user through dynamically generated hypermedia presentations with a life-like character. Intelligent User Interfaces. San Francisco, CA, pp 21–28
2. Bickmore T, Cook LK, Churchill EF, Sullivan JW (1998) Animated autonomous personal representatives. Autonomous agents. Minneapolis, MN, pp 8–15
3. Cohen MM, Massaro DW (1993) Modeling coarticulation in synthetic visual speech. In: Thalmann M, Thalmann D (eds) Computer Animation. Springer, Tokyo
4. Cosatto E, Graf H-P, Sample-based synthesis of photo-realistic talking heads. Proc. Computer Animation '98, Philadelphia, PA, pp 103–110
5. Damer B, Kekenes C, Hoffman T (1996) Inhabited digital spaces. Proc. CHI '96, pp 9–10
6. Doenges PK, Capin TK, Lavagetto F, Ostermann J, Pandzic IS, Petajan ED (1997) MPEG-4: Audio/Video & Synthetic Graphics/Audio for Mixed Media. Image Commun J, Special Issue on MPEG-4, 9(4):433–463
7. Don A, Oren T, Laurel B (1993) Guides 3.0 In: CHI-93, Video Preceedings, pp 447–448
8. Eisert P, Chaudhuri S, Girod B (1997) Speech driven synthesis of talking head sequences, 3D Image Analysis and Synthesis. Erlangen, pp 51–56
9. Fischl J, Miller B, Robinson J (1993) Parameter tracking in a muscle-based analysis/synthesis coding system. Pic-ture Coding Symposium (PCS '93), Lausanne, Switzerland, 2.3
10. Gibbs S, Breiteneder C (1993) Video widgets and video actors. UIST '93 pp 179–185
11. Kalra P, Mangili A, Magnenat-Thalmann N, Thalmann D (1992) Simulation of facial muscle actions based on rational free form deformation. Proc. Eurographics 92, pp 65–69
12. Kalra P (1993) An interactive multimodal facial animation system. Ph.D. Thesis nr. 1183, EPFL
13. Kampmann M, Nagel B (1998) Synthesis of facial expressions for semantic coding of videophone sequences. Computer Graphics Int (CGI98)
14. Milewski AE, Blonder GE (1998) System and method for providing structured tours of hypertext files. US Patent # 5760771. June 2, 1998
15. Text for FDIS 14496-1 Systems. ISO/IEC JTC1/SC29/WG11 N2503, MPEG98/October 1998
16. Text for FDIS 14496-2 Video. ISO/IEC JTC1/SC29/WG11 N2503, MPEG98/October 1998
17. Text for FDIS 14496-3 Audio. ISO/IEC JTC1/SC29/WG11 N2503, MPEG98/October 1998

18. Ostermann J, Haratsch E (1997) An animation definition interface - Rapid design of MPEG-4 compliant animated faces and bodies. The international workshop on synthetic-natural hybrid coding and 3D imaging, September 5–9, Rhodes, Greece
19. Ostermann J (1998) Animation of synthetic faces in MPEG-4. Proc. Computer Animation, Philadelphia, PA, pp 103–110
20. The Palace. http://www.thepalace.com
21. Pandzic IS, Capin TK, Lee E, Magnenat-Thalmann N, Thalmann D (1997) A flexible architecture for virtual humans in networked collaborative virtual environments. Proc. Eurographics, Budapest Hungary
22. Parke FI, Waters K (1997) Computer facial animation. Wellesley, MA, A.K. Peters
23. Parise S, Kiesler S, Sproull L, Waters K. My partner is a real dog: cooperation with social agents. Proc. CSCW '96. Cambridge, MA, pp 399–408
24. Rist T, Andre E, Muller J (1997) Adding animated presentation agents to the interface. Intelligent User Interfaces. Orlando, FL, pp 79–86
25. Sproat R, Olive J (1995) An approach to text-to-speech synthesis. In: Kleijn WB, Paliwal KK (eds) Speech coding and synthesis. Elsevier Science
26. Sproull L, Subramani M, Kiesler S, Walker JH, Waters K (1996) When the interface is a face. Human-Computer Interaction 11:97–124
27. Suler JR (1997) From ASCII to holodecks: Psychology of an online multimedia community. Presenation at the Convention of the American Psychological Association, Chicago
28. Terzopoulos D, Waters K (1993) Analysis and synthesis of facial image sequences using physical and anatomical models. IEEE Trans Pattern Anal Machine Intelligence, 15(6):569–579
29. Walker JH, Sproull L, Subramani R (1994) Using a human face in an interface. Proceedings of CHI '94. Boston, MA, pp 85–91



IGOR PANDZIC is a senior research assistant at MIRALab, University of Geneva. He graduated from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 1993. In 1994 he obtained a Masters degree in Computer Graphics from the Swiss Federal Institute of Technology in Lausanne, in 1995, a European Diploma of Superior Studies in Information Systems from the University of Geneva and in 1998. He obtained his PhD in Information Systems from the University of Geneva. His research interests include collaborative virtual environments, parallel computing, computer-generated film production and real-time facial expression analysis and synthesis. He contributes to the MPEG-4 Ad Hoc Group on Face and Body Animation. He has published more then 20 papers on these topics.

DAVID MILLEN'S research interests are in the areas of computer-supported collaborative work (CSCW) and human-computer interaction. He is interested in understanding how individuals and work groups use the Internet and other emerging communication technologies, and how these new technologies change work activities, organizational roles, and patterns of communication. His recent work includes the study of web-based email, media spaces, online communities, and teleworkers. His past work experience includes the user interface design of personal communications (PDA) products, business telephone systems development. He received a Ph.D. in Cognitive Psychology from Rutgers University.



JÖRN OSTERMANN, member IEEE, studied Electrical Engineering and Communications Engineering at the University of Hannover and Imperial College London, respectively. He received Dipl.-Ing. and Dr.-Ing. from the University of Hannover in 1988 and 1994, respectively. From 1988 till 1994, he worked as a Research Assistant at the Institut für Theoretische Nachrichtentechnik conducting research in low bit-rate and object-based analysis-synthesis video coding. In 1994 and 1995 he worked in the Visual Communications Research Department at AT&T Bell Labs. He has been working with Image Processing and Technology Research within AT&T Labs - Research since 1996.
From 1993 to 1994, he chaired the European COST 211 sim group coordinating research in low bitrate video coding. Within MPEG-4, he organized the evaluation of video tools to start defining the standard. Currently, he chairs the Adhoc Group on Coding of Arbitrarily-shaped Objects in MPEG-4 Video.
His current research interests are video coding, computer vision, 3D modeling, face animation, coarticulation of acoustic and visual speech, computer-human interfaces, and speech synthesis.