Realistic Talking Head for Human-Car-Entertainment Services

Kang Liu, M.Sc., Prof. Dr.-Ing. Jörn Ostermann

Institut für Informationsverarbeitung, Leibniz Universität Hannover

Appelstr. 9A, 30167 Hannover

Email: kang, ostermann@tnt.uni-hannover.de

Tel: +49-511-762 5323, Fax: +49-511-762 5333

Abstract

The rapid development of in-car communication systems in recent years has created problematic human interface challenges. Modern human-computer interfaces are integrated into car's communication systems to make its functionality easily accessible. This paper presents a realistic talking head, which enables a user interface with multi-modality. A talking head in human-car interfaces allows a more realistic means of communication as it simulates human-human interaction. This paper focuses on the use of talking heads for human-car-entertainment services. The architecture for integrating a talking head into human computer interfaces is also described.

Keywords:

Talking Head, Human-Computer Interface, TTS (Text-to-Speech), Human-Car Entertainment Service

1. Introduction

The rapid development of in-car communication systems in recent years has created problematic human interface challenges. There is a need to integrate modern human-computer interfaces into car's communication systems to make its functionality easily accessible. Today, human-car communication input is dominated by touch screens and buttons, while the computer output produces text, graphics and synthesized speech. The next step is to add realistic visual information in human-car interfaces to give the look and feel of human-human interaction.

Presently, the in-car communication system is a central control system, which controls the entire electronic equipment, such as air conditioning, audio and video player, telephone, and navigation system. One additional function is Internet access, so that any information is available while travelling. We believe that a natural interface accompanying the in-car communication system will be of major importance given the increasing number of applications available in a car. We have investigated that the human interaction with a computer, by means of a realistic talking head, enables the user to feel more trustworthy towards the information being communicated as the interaction between computer-human is much more similar to human-human interaction [1]. More encouraging, the talking head, combined with a dialogue system, will recognize the words of users due to automatic speech recognition technology. The talking head will have a positive impact on the perception of human-car communication.

Realistic face animation is still challenging, especially when we want to automate it to a large degree. Faces are the focus of attention for any audience, and the slightest deviation from normal faces is immediately noticed, especially for the mouth part. According to the underlying face model, talking faces can be categorized into 3D-model-based animation and image-based rendering of models [1]. Image-based facial animation can achieve more realistic animations, while 3D-based approaches are more flexible to render the talking head in any view and under any lighting condition.

The image-based approaches [2] [3] [4] analyze the recorded image sequences, and animations are synthesized by combining different facial parts extracted from the videos. A 3D model is not necessarily required for animations. In order to synthesize animations with speech, Bregler et al. [5] proposed a system called "video rewrite" which uses a database with video snippets of triphones. A new video is synthesized by selecting and concatenating the most appropriate triphone snippets. Cosatto et al. [2] [3] [4] described another image-based approach with higher realism and flexibility using a database of images labeled with phonemes. Based on Cosatto's approach, we developed a facial animation system presented in [6]. We implemented an image-based facial animation system according to Cosatto. Initially, three areas for improving the facial animation system were identified and implemented. The first improvement is to consider the coarticulatory features to model the lip synchronization, which is optimized by Pareto-optimization algorithm [7]. The second improvement [8] is using Active Appearance Model (AAM) [8] to detect the facial features, which performs more robust than color-based detection [2]. The detected features are also used to compute smooth transitions between different mouth images by morphing techniques. The third contribution for the animation is to use LLE to parameterize the texture of mouth images, which can achieves better results than PCA [9].

The paper is organized as follows. Section 2 describes the image-based facial animation system, including analysis and synthesis. Section 3 presents the applications of the talking heads and some concluding remarks are drawn in Section 4.

2. Image-based facial animation system

Using image-based rendering face animation requires a face model mainly consisting of a database of phonetically labeled face images. Using image-based rendering, the developed talking head can be driven by text or speech. The facial animation system consists of two parts: analysis and synthesis. The audio-visual analysis part creates the face model of a recorded human subject. The synthesis part generates facial animations by concatenating appropriate mouth images from the database.

2.1 Analysis

The audio-visual analysis of recorded human subjects is depicted in Fig. 1. The analysis results in a database of mouth images and their relevant features suitable for synthesis. In a first step the audio and video of a human subject reading text of a predefined corpus are recorded. The recorded audio and the spoken text are processed by speech recognition, which uses Hidden Markov Models to recognize and temporally align the phonemes to the speech signal. Finally, the timed sequence of phonemes is aligned with the corresponding video. Therefore, for each frame of the recorded video, the corresponding phoneme and phoneme context are known. The phonetic context is required due to the coarticulation that indicates that a particular mouth shape depends not only on its associated phoneme but also on its preceding and succeeding phonemes.

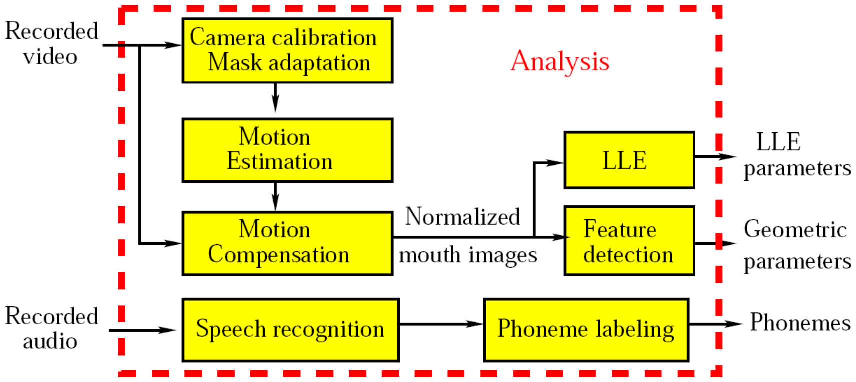


Fig. 1. Audio-visual analysis: the parameters are estimated for building the database.

The camera is calibrated and a face mask is adapted to the first frame using the calibrated camera parameters and some facial feature points [10]. In the next step, motion estimation [11] is carried out to compute the rotation and translation parameters of the head movement in the later frames. These motion parameters are used to compensate for the head motion such that normalized mouth images can be saved in the database. LLE parameters are calculated by LLE (locally linear embedding), which describe the texture of a mouth image. The geometric parameters, such as mouth corner points and lip position, are obtained by feature detection. All the parameters associated with an image are also saved in the database. Therefore, the database is built with a large number of normalized mouth images. Each image is characterized by geometric parameters, texture parameters (LLE parameters), phonetic context, etc.

2.2 Synthesis

The block diagram of the synthesis of a talking head is shown in Fig. 2. The whole synthesis part in Fig. 2 is also defined as a visual text to speech synthesizer (VTTS). First, a segment of text is inputted to a text-to-speech synthesizer (TTS). The TTS provides the audio track as well as the sequence of phonemes and their durations, which are sent to the unit selection engine. Depending on the phoneme information, the unit selection selects mouth images from the database and assembles them in an optimal way to produce the desired animation. The unit selection balances two competing goals: lip synchronization and smoothness of the transition between consequent images. For each goal a cost function is defined, both of them are functions of the mouth image parameters defined above. The cost function for lip synchronization considers the coarticulation effects by matching the distance between the phonetic context of the synthesized phoneme and the phonetic context of the selected mouth image in the database. The cost function for smoothness reduces the visual difference at the transition of images in the final animation. In the next stage, an image-rendering module stitches these mouth images to the background video sequence. Background videos are recorded video sequences of the human subject with typical short head movements. Finally, the talking head is displayed.

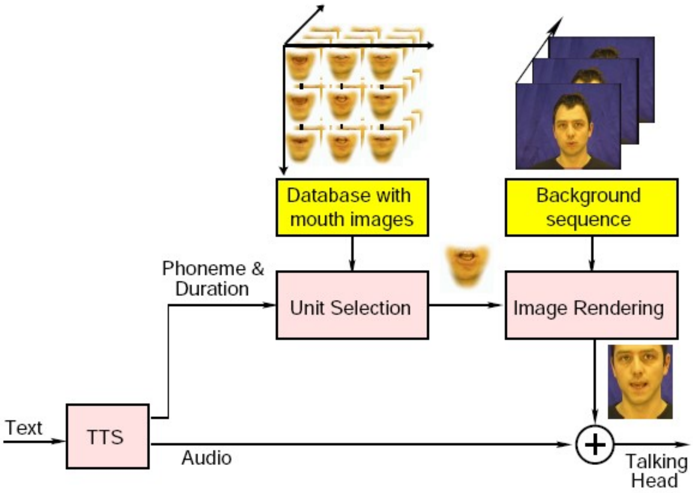


Fig. 2. The system architecture of the image-based talking head synthesizer

Depending on different input modalities, a talking face can be driven by text or speech. The text-driven talking face consists of a TTS and a talking face. The TTS synthesizes the audio with phoneme information from the input text. The speech-driven talking face has original sound, which is processed by speech recognition to determine the phoneme information. In both cases, the phonetic information drives the facial animation. A text-driven talking face is flexible and can be used in many applications, but the quality of speech is not as effective as authentic human speech.

3. Applications with a real time talking head

In this section we present the use of a talking head in Web-based applications as well as in human-car communication. The talking head for interactive services can be implemented using different architectures, depending on the software and hardware requirements of the client as well as the available bandwidth for client-server communication. There are two possible scenarios: the talking head can be rendered either at the server or at the client. Advantages and disadvantages are listed in Table 1.

Since the rendering of a talking head on a server is computationally very expensive, this scenario is only viable for non-interactive applications or applications based on prerendered animations. Therefore, we recommend to render the face at the client. In order to reduce the size of a face model, we compress its database using PCA and H.264 resulting a face model size of 35MB assuming HDTV format. The face model can be rendered in real time using a 1.99 GHz PC.

Table 1. Advantages and disadvantages of the talking head rendered at client or server

	Render at Server	Render at Client
Advantages	- low computational load on the client	- low computational load on the server - only the text and audio data transferred to client, suitable for narrow bandwidth - best video quality - real time
Disadvantages	- very high computational load on the server - streaming audio and video data, requiring broadband connection - low video quality - large latency	- face animation software - require graphics engine

3.1 E-Cogent

We have implemented the E-cogent application (Fig. 3). E-cogent is an electronic convincing agent [12]. E-cogent can help the customers choose a notebook. The talking head will start a conversation with the user as soon as the site is entered. A dialogue system will answer any questions from the user and send the answer to the TTS (Text-to-Speech Synthesizer). The TTS produces the spoken audio track as well as the phonetic information and their duration which is required by the talking head plug-in embedded in the Web site. The talking head plug-in at the client is responsible for generating and rendering speech animations of the virtual customer service agent.

E-Cogent presents two brands of notebooks in the demonstration as shown in Fig. 4. On the left top of the screen our talking head is beginning to initiate a conversation with the user, while on the right side, the notebook and the interface to the dialogue system are displayed. The user can select different notebooks and ask questions. The dialogue system can answer the questions related to the notebooks. The E-cogent provides the user with hints as they surf through the Web page. If there are no related answers in the dialogue database, the questions are forwarding to a customer service centre.

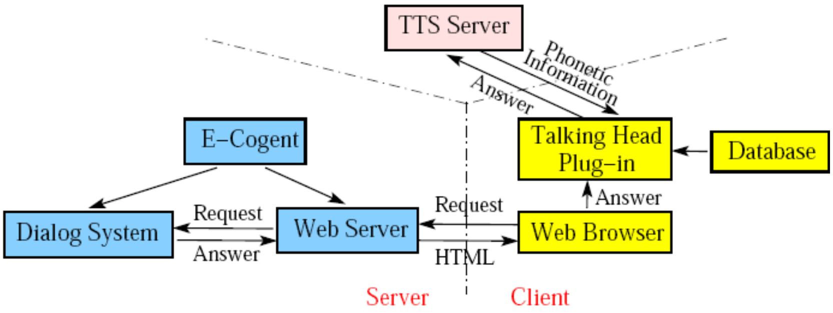
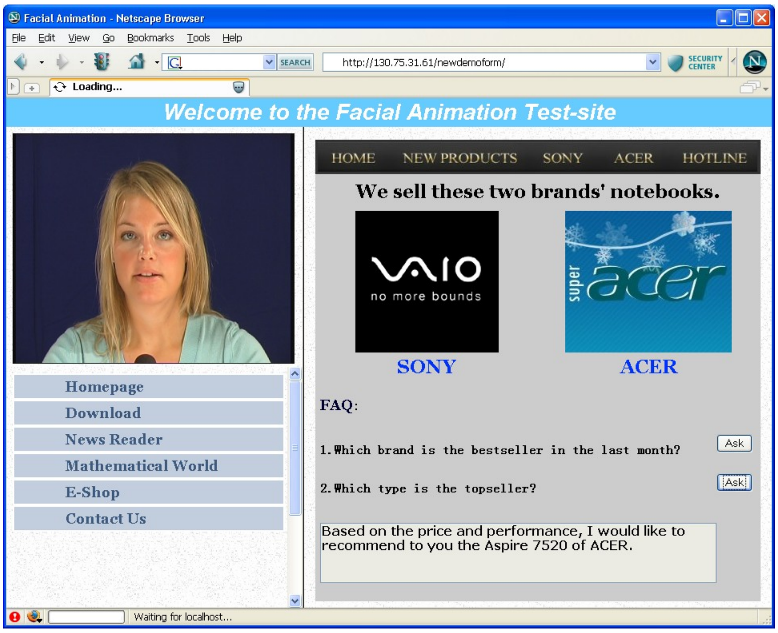


Fig. 3. Architecture of the E-Cogent application.


Facial Animation - Netscape Browser
File Edit View Go Bookmarks Tools Help
SEARCH
http://130.75.31.61/newdemoform/
SECURITY CENTER
Loading...
Welcome to the Facial Animation Test-site
HOME NEW PRODUCTS SONY ACER HOTLINE
Homepage
Download
News Reader
Mathematical World
E-Shop
Contact Us
VAIO
no more bounds
super acer
SONY
ACER
Ask
Ask
Waiting for localhost...

We sell these two brands' notebooks.

FAQ:

1.Which brand is the bestseller in the last month?

2.Which type is the topseller?

Based on the price and performance, I would like to recommend to you the Aspire 7520 of ACER.

Fig. 4. E-cogent presents notebooks with different brands.

3.2 Human-car entertainment service

With the expansion of various mobile devices and entertainment services in cars such as navigation system, in-car video and games, new application areas for virtual humans are opening. For these mobile device and entertainment users, applications with interfaces like text and speech are especially important. Consequently, the animation of virtual characters based on text and speech inputs could enable rich multimedia services in cars. At the same time, talking faces bring personality and human touch into everyday use of mobile devices.

Personalized talking heads can be used as part of multimedia services in cars, such as human-car entertainment services. The interfaces could be an in-car operating console, an in-car conference telephone. The user interfaces with the in-car operating console using either touch screens or buttons or via a microphone and speech recognition. A natural language understanding unit extracts structure and meaning from the raw text. The dialogue manager provides the answers for the talking head based on dialogue data and state information. The dialogue manager also emits an event to drive equipment. Fig. 5 shows the architecture of the human-car communication system.

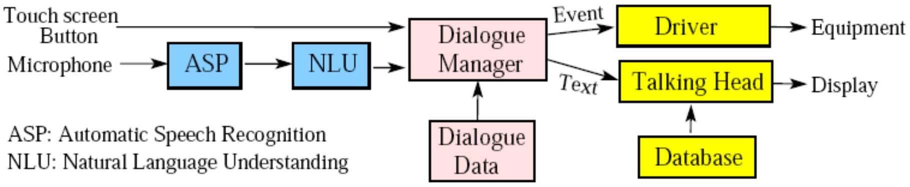


Fig. 5. Architecture of the human-car communication system.

Another example is a text messaging service, which converts SMS (Short Messaging Service) message into a short video clips [13]. It means that a user is able to see the personalized talking face looking like the caller on the mobile screen. Furthermore, personalized talking faces might be used in the teleconference or other multimedia service instead of a real human. Fig. 6 shows different display interfaces for talking heads. The user can get a face in a car for travel information, entertainment and road side information assistant.

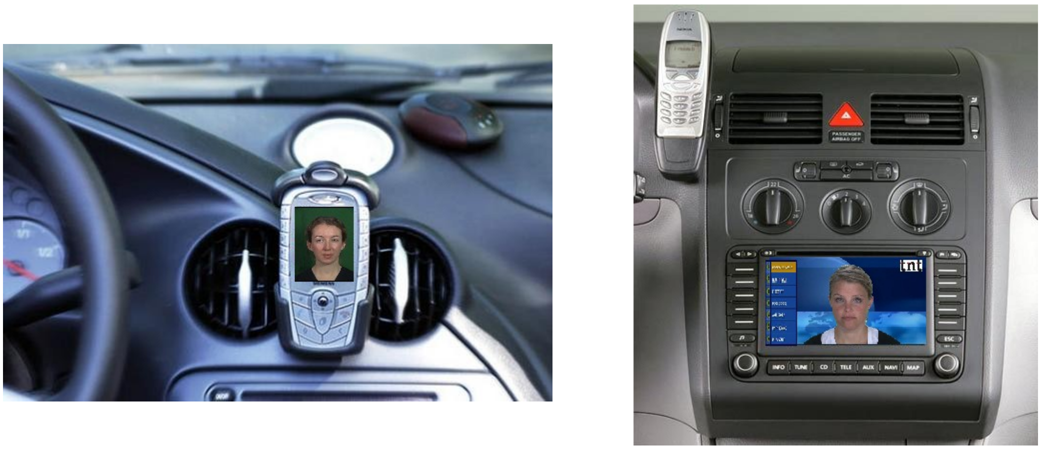


Fig. 6. Talking head display interfaces. (left) SMS conversion system on a cell phone. (right) in-car display units for information, entertainment, and system control.

4. Conclusions

Facial animation has been combined with text-to-speech synthesis to create innovative multimodal interfaces, such as web-based in-car entertainment services. Using image-based rendering, facial animations look more realistic than those facial animations generated by using 3D models. Image-based facial animation system consists of two parts. One is the audiovisual analysis system for recorded human subjects creating a face model, composed of a personalized 3D mask as well as a large database of mouth images and their related information. The other is the synthesis system for facial animations, which concatenates appropriate mouth images from the database.

Based on the developed facial animation system, web-based applications are introduced. Combined with a dialogue system, the talking head interface will give the look and perception of human-human interaction.

An interactive demonstration of the applications, incorporating talking heads, is available at http://www.tnt.uni-hannover.de/project/facialanimation/demo.

5. References

[1] J. Ostermann and A. Weissenfeld, "Talking faces – technologies and applications," Proceedings of 11th International Workshop on Systems, Signal and Image Processing, IWSSIP, 2004.

[2] E. Cosatto, J. Ostermann, H. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," Proceedings of the IEEE , vol. 91, no. 9, pp. 1406-1429, September, 2003.

[3] E. Cosatto and H.P. Graf, "Sample-based synthesis of photo-realistic talking heads", Proceedings of IEEE Computer Animation, pp. 103-110, 1998.

[4] E. Cosatto and H.P. Graf, "Photo-realistic talking heads from image samples", IEEE Trans. Multimedia, vol. 2, no. 3, pp. 152-163, 2000.

[5] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio", Proc. ACM SIGGRAPH'97, in Computer Graphics Proceedings, 1997.

[6] A. Weissenfeld, K. Liu, S. Klomp, and J. Ostermann, "Personalized unit selection for an image-based facial animation system," 7th InternationalWorkshop onMultimedia Signal Processing , Shanghai, China, 2005.

[7] K. Liu, J. Ostermann. "Realistic Facial Animation System for Interactive Services," Interspeech 2008 in the special session: LIPS 2008: Visual Speech Synthesis Challenge. Brisbane, September, 2008.

[8] K. Liu, A. Weissenfeld, J. Ostermann, X. Luo. "Robust AAM Building for morphing in an Image-based Facial Animation System," ICME 2008, Hannover, June, 2008.

[9] K. Liu and A. Weissenfeld and J. Ostermann, “Parameterization of mouth images by LLE and PCA for imagebased facial animation”, Proc. ICASSP 06, Toulouse, France, May 2006.

[10] A. Weissenfeld, N. Stefanoski, Q. Shen and J. Ostermann, “Adaptation of a Generic Face Model to a 3D Scan”, Workshop On Immersive Communication And Broadcast Systems, Berlin, 2005.

[11] A. Weissenfeld, O. Urfalioglu, K. Liu and J. Ostermann, “Robust Rigid Motion Estimation based on Differential Evolution”, Proc. ICME 06, Canada, 2006.

[12] J. Ostermann. “E-COGENT: An electronic convincing agent,” in MPEG-4 Facial Animation: The Standard, Implementation and Applications, I.S. Pandzic and R.Forchheimer, Eds. Chichester, U.K.: Wiley, pp 253-264, 2002.

[13] Jürgen Rurainsky, Peter Eisert. “Text2Video: A SMS to MMS conversion,” ITG Dortmunder Fernsehseminar, pp.163-168, Dortmund, 2005.

6. Biography:

Kang Liu was born in 1977. He studied Mechanical and Electrical Engineering at the Institute of Mechatronic Control Engineering, Zhejiang University, P.R. China. He recieved his Bachelor and Master Degree from Zhejiang University in 2001 and 2004, respectively. Since March, 2004 he has been working toward the PhD degree at the Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany. Currently, he is an active researcher in the 3DTV and facial animation projects. He has published several research papers. His current research interests are image processing, video coding, facial animation, and computer-human interfaces.

Jörn Ostermann 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 on video coding. He was a member of Image Processing and Technology Research within AT&T Labs - Research from 1996 to 2003. Since 2003 he is Full Professor and Head of the Institut für Informationsverarbeitung at the Leibniz Universität Hannover, Germany. In 2007, he became head of the Laboratory for Information Technology.

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. He chaired the Adhoc Group on Coding of Arbitrarily-shaped Objects in MPEG-4 Video. Since 2008, he is the Chair of the Requirements Group of MPEG (ISO/IEC JTC1 SC29 WG11). Jörn was a scholar of the German National Foundation. In 1998, he received the AT&T Standards Recognition Award and the ISO award. He is a Fellow of the IEEE and member of the IEEE Technical Committee on Multimedia Signal Processing and past chair of the IEEE CAS Visual Signal Processing and Communications (VSPC) Technical Committee. Jörn served as a Distinguished Lecturer of the IEEE CAS Society. He published more than 100 research papers and book chapters. He is co-author of a graduate level text book on video communications. He holds more than 30 patents.

His current research interests are video coding and streaming, 3D modelling, face animation, and computer-human interfaces.