Avatar User Interfaces in an OSGi-based System for Health Care Services

Sascha Fagel[1], Andreas Hilbert[1], Christopher Mayer[2], Martin Morandell[2], Matthias Gira[2], Martin Petzold[3]

[1]Zoobe message entertainment GmbH, Research & Development, Berlin, Germany
[2]AIT Austrian Institute of Technology GmbH, Health & Environment, Vienna, Austria
[3]ProSyst Software GmbH, Cologne, Germany

{fagel|hilbert}@zoobe.com, {martin.morandell|christopher.mayer|matthias.gira}@ait.ac.at, m.petzold@prosyst.com

Abstract

We present a system to display text information by an animated talking avatar suitable for health care services. Speech animation parameters are calculated by a co-articulation model from the phone chain extracted from the text-to-speech processing step. An animation script that layers body movements and speech animation is generated and then rendered and converted into an h.264 video by a computer game engine. The animation system is attached to AALuis, an OSGi-based system for care services for older adults within a European research project.

Index Terms: character animation; ambient assisted living; audiovisual speech synthesis; health care service; avatar

1. Introduction

The average age of the population of the European Union is significantly increasing [1]. This poses challenges to the society in various respects, one of them being the need for more efficient health care for older adults. Many current and future health care services will be a combination of human service and ICT system, i.e., the service will at least partly be provided by a computer application.

Even though health care technology is often very complex in design, implementation and maintenance, the user interface towards the end user – in this case older adults with all kinds of abilities, preferences and special needs – has to be kept very simple, easy to use and especially enjoyable and attractive. The user interface is the single component in such systems upon which everything else will be judged [2]. Therefore, in particular usability, accessibility, as well as the freedom of choice concerning the interaction with such systems are the crucial points for acceptability, applicability and subsequently the benefit of such systems – for the user him- or herself, for the society and in general all stakeholders.

The AALuis Project (Ambient Assisted Living user interfaces) [3,4] focusses on the aspect of freedom of choice for the preferred ways of user interaction. New approaches such as multi touch technologies and usage of avatars are developed and adopted to the very heterogeneous needs of primary end-users, older adults who can derive a benefit from AAL Systems.

Embodied Conversational Agents (ECAs) that display appropriate non-verbal behavior were shown to enhance user satisfaction and engagement and improve the users' interaction with a computer system [5]. Therefore, the concept of an avatar that represents the service to the user as virtual personification was chosen as a core component of the user interface in the AALuis project. Furthermore, the addition of a visual display to verbal information – i.e., adding a lip-synched animated character to audio speech output – can increase the intelligibility and enhance the robustness of the information transmission [6] as known from natural speech [7].

2. Animation System

The core of the animation system is implemented in a JEE enterprise application on a Red5 Media Server [8]. The application provides a custom API through WebSocket protocol [9]. Besides the animation generation the enterprise application serves acts as a dispatcher – it manages the communication between the nodes that are necessary to fulfill the job. The data is exchanged between these nodes via TCP/IP.

2.1. Animation Generator

The animation generator collects data about the video that is to be created and generates an animation script accordingly. An animation script contains the following elements:

- the character model
- a sequence of animations for the character
- speech animation parameters
- the scenery, the light set, and the camera settings
- the audio track
- technical data such as paths and encoder settings

References to all these elements (except speech animation parameters and the audio track which are dynamically added) and metadata such as the locations of the 3-D source files are stored in a database. The animation generator collects the necessary information and fills an xml data structure. The body animations are taken from a pool of animation cycles that are feasible to accompany speech and concatenated in a random sequence where dedicated animations can be triggered by emphasized parts of the utterance.

2.2. Text-To-Speech

Data generated by different processing stages of the CereProc cServer [10] are used for text-to-speech conversion. Animation parameters for lip-sync speech movements are derived from the duration generation stage which gives as output a chain of phonemes with their respective timings. Animation parameters are calculated for *jaw opening, lip opening, lip spreading,* and *tongue tip raising* which are independently derived by an implementation of the dominance co-articulation principle [11]. Model parameters for ideal articulator positions and their dominances for a given phoneme are available from a study by Fagel and Clemens [12]. The presented method is modified in two details:

2. Instead of generating two target positions per phoneme – resulting in animation parameters not equally distributed over time – the animation parameters are determined based on equidistantly sampled phoneme values with a sample rate equal to the actual video frame rate.
3. The calculation is extended by a hypo/hyper-articulation parameter to generate slower movements with smaller magnitude or faster movements with greater magnitude, respectively. An activation value defined as "low", "medium" or "high" that is assigned to the body animation entry in the database controls this articulation parameter.

2.3. Rendering

The animation script is executed by a modified version of the open source game engine Nebula Device 3.0 [13]. We attached a server function to the engine in order to receive the render jobs and to deliver the results via TCP/IP. Furthermore, the frame sequence is captured from the video memory and passed to ffmpeg [14] for encoding. The final files are written to a network drive shared by the render server and the core animation system.

3. System Integration

The animation system itself provides a secure WebSocket endpoint to create avatar videos from text. This endpoint is integrated into an OSGi framework [15] running the AALuis layer. This allows the creation of avatar videos from within any OSGi component. OSGi allows a plug-and-play integration of (web) service into a software system. It is a common platform for home gateways and other embedded devices.

Character and scene settings are defined by the health care service and given to the animation system by query parameters along with the text to be spoken. The video is rendered on the server and encoded in h.264. The request returns an URL to the generated video file. This URL is embedded in a '<video>-tag' in a HTML5 user interface. The video file is streamed from the server and played when the user interface is displayed. The video file is cached locally and played from the file system if the same utterance is requested with the same parameters again.

4. Discussion

There have been several approaches to use avatars in Ambient Assisted Living environments for older adults and personal home care assistants. The study Avatars@Home [16] brought some insights concerning the use of Avatars compared to other output modalities. These findings lead to the presented approach that comprises multimodal user interaction. From the AALuis approach we expect the following advantages

- Joy of use: high level of design combining entertainment, infotainment and edutainment
- Broader applicability of the given avatars: designed for a broad field of information delivery and services
- Fewer risks concerning personal relationships: virtual person representing the service without the risk of interpersonal stress due to the use of familiar faces

Interviews within the AALuis project on the possible usage of (personalized) avatars brought positive answers, in particular for applications such as tutoring. AALuis lab and field trials will bring deeper insights on acceptability, likeability and usability of avatars within Ambient Assisted Living environments.

5. Conclusions and Future Work

We presented a system that generates videos of animated characters from speech or text. Although other applications are obvious, the animated character shown here is especially designed to serve as a virtual personification of an Ambient Assisted Living service in any user interface based on html5.

To ensure maximum platform independence of the user interface the animated voice message is delivered by displaying the accordingly generated video file. In the next version the avatar will be displayed on the screen with idle movements that fill the gaps between the informational outputs in order to represent the provided service persistently to the user. Methods for client-sided rendering of the avatar are currently under investigation. A customized version of the OGRE game engine [17] is under development to replace the currently used Nebula Device in order to increase the rendering quality. Several male and female avatars will be modeled and animated in order to best represent the service and to best fit the user needs. The design of the avatars will be guided by a target group questionnaire.

6. Acknowledgements

The presented work was supported by the European Commission under the AAL Joint Programme, the German Federal Ministry of Education and Research, funding reference No. 16SV5573K, the Austrian BMVIT and the programme "benefit".

7. References

[1] Giuseppe C., Declan C., "Can Europe Afford to Grow Old?", International Monetary Fund Finance & Development 43:3, 2006.
[2] van Berlo, A., "Design Guidelines on Smart Homes", A COST 219bis Guidebook, 1999.
[3] AALuis project homepage: www.aaluis.eu [retrieved 26.04.2013]
[4] Mayer, C., Morandell, M., Hanke, S., Bobeth, J., Bosch, T., Fagel, S., Groot, M., Hackbarth, K., Marschitz, W., Schüler, C., Tuinenbreijer, K., "Ambient Assisted Living User Interfaces", Everyday Technology for Independence & Care 29, 456-463, 2011.
[5] Foster, M. E., "Enhancing Human-Computer Interaction with Embodied Conversational Agents", Proc. HCII, 2007.
[6] Ouni, S., Cohen, M. M., Ishak, H., Massaro, D. W., "Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads", EURASIP JASMP, 2007.
[7] Sumby, W.H., Pollack, I., "Visual Contribution to Speech Intelligibility in Noise", JASA 26, 212-215, 1954.
[8] Red5 Media Server 1.0, http://www.red5.org [retrieved 26.04.2013]
[9] IETF: The WebSocket Protocol. RFC6455, 2011. http://tools.ietf.org/html/rfc6455 [retrieved 26.04.2013]
[10] CereProc, cServer Text-to-Speech Server, http://www.cereproc.com/en/products/server [retrieved 26.04.2013]
[11] Löfqvist, A. "Speech as audible gestures", In W. J. Hardcastle and A.Marchal (Eds.): Speech Production and Speech Modeling, NATO ASI Series, 55, Kluwer, Dordrecht, 289–322, 1990.
[12] Fagel, S. and Clemens, C., "An Articulation Model for Audiovisual Speech Synthesis - Determination, Adjustment, Evaluation", Speech Communication 44, 141-154, 2004.
[13] The Nebula Device. http://sourceforge.net/projects/nebuladevice [retrieved 26.04.2013]
[14] FFmpeg. ffmpeg.org [retrieved 26.04.2013]
[15] The Open Services Gateway initiative framework. http://www.osgi.org [retrieved 26.04.2013]
[16] Morandell M, Hochgatterer A., Wöckl B., Dittenberger S., Fagel S., "Avatars@Home InterFACEing the Smart Home for Elderly People", Lecture Notes in Computer Science 5889, 353-365, 2009.
[17] Torus Knot Software Ltd: OGRE (Object-Oriented Graphics Rendering Engine). http://www.ogre3d.org/ [retrieved 26.04.2013]