The Virtual Counselor

Automated Character Animation for Ambient Assisted Living

Sascha Fagel, Andreas Hilbert
Research & Development
zoobe message entertainment GmbH
Berlin, Germany
{fagel|hilbert}@zoobe.com

Martin Morandell, Christopher Mayer
Health & Environment
AIT Austrian Institute of Technology GmbH
Vienna, Austria
{martin.morandell|christopher.mayer}@ait.ac.at

Abstract—We present a system for automated animation of text or voice messages suitable for Ambient Assisted Living user interfaces. Input to the system can be text, a pre-recorded speech file, or the speech signal captured directly from the microphone. Speech animation parameters are calculated by a co-articulation model either for the voice audio or – if available – from the phone chain extracted from the Text-To-Speech processing step in case of text input. An animation script that layers body movements and speech animation is generated. This script is rendered and converted into an h.264 video by a computer game engine. The system is developed to be used in care services for elderly users within a European research project.

Keywords: automatic character animation; embodied conversational agent; ambient assisted living; multimodal user interfaces; audiovisual speech synthesis

I. Introduction

Like in many other places in the world the average age of the inhabitants of the European Union is significantly increasing. This poses challenges to the society in various respects one of them being the need for more efficient healthcare for elderly people. Many current and future healthcare services will consist of a combination of human personnel and automated ict-systems, i.e., the service will at least partly be provided by a computer application.

Even though health- and homecare technology is often very complex in design, implementation and maintenance, the user interface towards the end user – in this case elderly people 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 [1]. 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 also for the market-stakeholders.

The AALuis Project (Ambient Assisted Living user interfaces) [2,3] 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, elderly people 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 [4]. Therefore, the concept of an avatar that represents the service to the user as virtual personification was chosen as a core component of the AALuis-project user interface development. 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 [5] as known from natural speech [6].

This paper is structured as follows: The next chapter describes the animation system with its components animation generator, text-to-speech transformation, audio processing, and rendering. Chapter III describes how we embed the generator client and generated animations into user interfaces. We finish the paper with a discussion, and conclusions and future work.

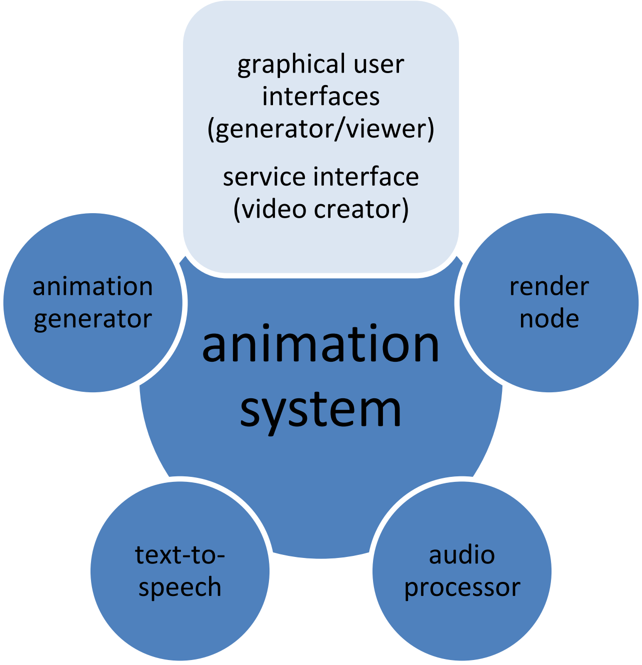
graphical user interfaces (generator/viewer)
service interface (video creator)
animation generator
render node
animation system
text-to-speech
audio processor

Figure 1. System overview.

II. Animation System

The core of the animation system is implemented in a JEE web application on a Red5 Media Server [7]. The web application provides a custom API that is accessible over the Real Time Messaging Protocol (RTMP, [8]). Besides the animation generation (see next section) the web application serves as a dispatcher of some sort – it manages the communication between the nodes that are necessary to fulfill the job. The data is exchanged between the components via TCP/IP. Figure 1 shows an overview of the system.

A 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
- the camera settings
- the audio track
- technical data such as paths and encoder settings

References to all these elements (except speech animation parameters, the audio track and the technical data) 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.

B Text-To-Speech

Data generated by different processing stages of the open-source-system MARY TTS [9] are used for Text-to-Speech-conversion. Alternatively, the high quality commercial CereProc cServer [10] is used. 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:

1. 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.
2. 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.

C Audio Processing

Animation parameters for jaw opening, lip opening and lip spreading are calculated from a language independent, mostly rule-based audio analysis if speech recordings instead of TTS data are used. In a first step the audio signal is normalized, de-clicked, cropped and noise reduced.

Audio analysis for animation parameter extraction is carried out on the speech signal re-sampled to 22050Hz. Frequency-domain representations are gathered from that signal with an FFT and a step-size that equals or is greater than the video frame-rate with a minimum window overlap of 50%.

To determine the appropriate values for *lip spreading, jaw opening* and *lip opening* the following measurements are considered: broad-band energies narrow-band energies, tonality measure (1 - spectral flatness in the range 180Hz–1250Hz) and spectral difference (75Hz–3kHz) between two subsequent analysis frames (i.e., first order temporal deviation of the square-summed spectrum). All energy-based values are logarithmized. *Tongue tip raising* parameters are not calculated from the audio signal.

The narrow band energies cover the characteristic formant features of vowels which are then categorized as /a/-like, /e,i/-like and /o,u/-like sounds by different linear combinations of those measures and mainly help to determine the *lip spreading*. Those are later also weighted by the tonality measure which accounts for the amount of harmonic content in the signal and hence indicates vowels and voiced consonants.

Most influential for the *jaw opening* parameter are broad-band-energy measure, tonality measure and inversely the higher frequency measure accounting for sibilants.

The *lip opening* parameter is calculated solely based broad-band-energy measure and tonality measure.

Measurements are smoothed by a weighted moving average of 100ms.The weights of the measures are determined empirically and the parameters are normalized to range between 0 and 1. Final frame-wise animation parameters are extracted from the analysis parameter sequence by a band limited sinc interpolation.

D 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.

III. Embedding the Animation

The animation system itself provides external interfaces for the generation and the delivery of animated speech messages. The user interface for the generation process either takes text as input, records a voice message from the microphone, or provides a selection of pre-recorded voice messages. This front-end is implemented in Flash ActionScript 3.0 [15] and uses the web browser's Flash plug-in as runtime environment. Figure 2 shows the text-to-avatar Flash interface.

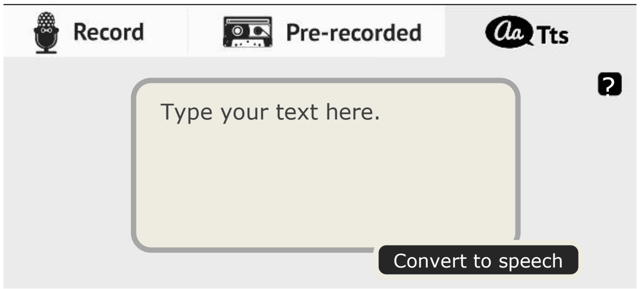

Figure 2. Text-to-avatar user interface.

Additionally to the (graphical) Flash video generation interface, a secure websocket based interface to create videos from text or speech audio is implemented in the OSGi framework [16] to create avatar videos from within any OSGi service. Character and scene settings are then defined by the service and given to the animation system by query parameters together with the text to be spoken or a reference to a (pre-) recorded speech audio.

When the video is rendered in the server back-end and encoded in h.264. This is delivered to the generating Flash application as preview, to a second Flash application (a simplified version of the one for generation), or can be embedded as a <video>-tag in an html5 website. The respective OSGi interface gets – as response to the call to create a video – a url for pseudo-streaming over http(s). Figure 3 shows a screenshot of the first prototype of the Virtual Counselor.

IV. Discussion

There have been several approaches to use avatars in the field Ambient Assisted Living environments for elderly and personal home care assistants. Morandell et al [17] describe some advantages and a comparison of different types of avatars for this application field.

A very promising approach is the use of photo realistic avatars as personal assistants, in particular for the Human–Computer Interaction with elderly people with a diagnose of mild cognitive impairment. The presence of familiar faces can bring benefits such as higher acceptance, attention creation and creation of personal relation. A personalized non-photorealistic 3D avatar may have a comparable positive effect assuming it is clearly recognized as the intended person.

The study Avatars@Home [18] brought some insights concerning the use of Avatars compared to other output modalities. These findings are incorporated into the present approach that will combine a multimodal user interaction.

From the AALuis approach we expect the following advantages

- Joy of use: The high level of design and animation will lead to joyful use of the application. Combining entertainment, infotainment and edutainment.
- Broader applicability of the given avatars: The designed characters can be used for a broad field of information delivery and services.
- Fewer risks concerning personal relationships: Even though familiar faces could increase personal bindings this could also lead to interpersonal stress when an avatar represents a known person.

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.

V. 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 Adobe Flash or html5.

The communication between the front-end (user interface) and the back-end (animation system) via RTMP was recently extended by secure websocket connectivity to enable a wider variety of front-end solutions. iPhone and Android apps for mobile clients are currently being developed.

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.

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.

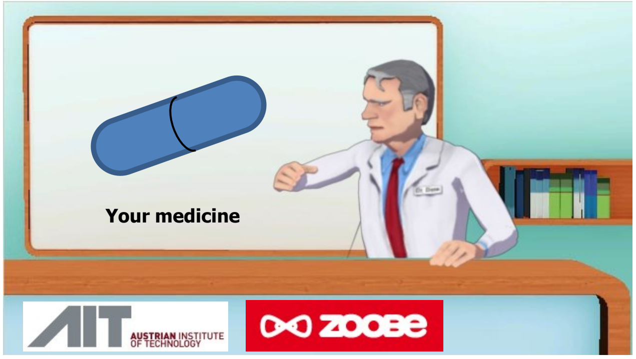

Figure 3. The Virtual Counselor.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] van Berlo, A., “Design guidelines on smart homes”, A COST 219bis Guidebook, 1999.

[2] AALuis project homepage: www.aaluis.eu [online resource, retrieved 18.10.2012]

[3] 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 and Care - AAATE Assistive Technology Research Series 29, 456-463, 2011. DOI: 10.3233/978-1-60750-814-4-456

[4] Foster, M. E., “Enhancing Human-Computer Interaction with Embodied Conversational Agents”, Proc. International Conference on Human-Computer Interaction, Beijing, 2007.

[5] Ouni, S., Cohen, M. M., Ishak, H. and Massaro, D. W., “Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads”, EURASIP Journal on Audio, Speech, and Music Processing, 2007.

[6] Sumby, W.H. and Pollack, I., “Visual Contribution to Speech Intelligibility in Noise”, Journal of the Acoustical Society of America, 26, 212-215, 1954.

[7] Red5 Media Server 1.0, http://www.red5.org [online resource, retrieved 18.10.2012]

[8] RTMP Specification 1.0, Adobe Systems Incorporated, http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/rtmp/pdf/rtmp_specification_1.0.pdf [online resource, retrieved 18.10.2012]

[9] Schröder, M. and Trouvain, J., “The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching”, Intl. Journal of Speech Technology 6, 365-377, 2003.

[10] CereProc, cServer Text-to-Speech Server, http://www.cereproc.com/en/products/server [online resource, retrieved 18.10.2012]

[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

[14] FFmpeg. ffmpeg.org [online resource, retrieved 19.10.2012]

[15] ActionScript 3.0 Reference for the Adobe Flash Platform. http://help.adobe.com/en_US/FlashPlatform/reference/actionscript/3/ [online resource, retrieved 02.01.2013]

[16] The Open Services Gateway initiative framework. http://www.osgi.org [online resource, retrieved 18.10.2012]

[17] Morandell, M., Hochgatterer, A., Fagel, S. and Wassertheurer, S, “Avatars in Assistive Homes for the Elderly A User-Friendly Way of Interaction?”, Lecture Notes in Computer Science 5298, 391-402, 2008. DOI: 10.1007/978-3-540-89350-9_27

[18] Morandell M, Hochgatterer A., Wöckl B., Dittenberger S. and Fagel S., “Avatars@Home InterFACEing the Smart Home for Elderly People”, Lecture Notes in Computer Science 5889, 353-365, 2009. DOI: 10.1007/978-3-642-10308-7_25