Avatars 4 All – An Avatar Generation Toolchain

Miroslav Sili
Health & Environment Department
AIT Austrian Institute of Technology GmbH
miroslav.sili@ait.ac.at

Elisabeth Broneder
Health & Environment Department
AIT Austrian Institute of Technology GmbH
elisabeth.broneder@ait.ac.at

Martin Morandell
Health & Environment Department
AIT Austrian Institute of Technology GmbH
martin.morandell@ait.ac.at

Christopher Mayer
Health & Environment Department
AIT Austrian Institute of Technology GmbH
christopher.mayer@ait.ac.at

ABSTRACT
The work-in-progress paper presents an application driven approach to create a versatile toolchain to create avatars. It targets the lack of an economic solution that enables the integration of avatars on multiple platforms and multiple devices. The requirements are based on Active and Assisted Living (AAL) domain-specific research and application-experiences of avatar-based user-interaction in the last decade. The created toolchain consists of phases: i) pre-processing, ii) processing and iii) optional post-processing. To achieve a high level of anthropomorphism not only the visual representation and synchronized voice is required but also combined movements, emotions and in particular affective reactions on the user's behavior. The flexible toolchain enables an integration of ongoing and future improvements. The first results are promising and show that an economic but high-quality solution that can be integrated easily in different services can be developed.

CCS Concepts
• Social and professional topics~Assistive technologies • Computing methodologies~Intelligent agents • Software and its engineering~Virtual worlds software

Keywords
Virtual Agents; Affective Interaction; Avatars; Assistive Technologies;

1. INTRODUCTION
In recent years, there has been an increasing interest in avatar-based Human Computer Interaction (HCI) in the healthcare context, in particular concerning older adults. Several studies [1,2,3,4,5] have documented the benefits of avatar-based user interaction.

Ortiz et al. [6] summarize these as follows: "All of them reached the same conclusion, that the presence of an avatar neither has a positive nor a negative effect but it may have a positive effect on the subject's impression, as the user's experience may be perceived as less difficult and more entertaining". Indeed, simplicity and understandability are crucial, especially for the target group of older adults.

Older adults could benefit from avatar-based user interaction but the remaining question is: How can developers/designers utilize avatars in their applications and services? Is there any Software Development Kit (SDK), software library or module, widget or cloud based service, which is able to provide an avatar that can be embedded in their application or service with low development barriers? Although there are several solutions available on the market, it is very often difficult to find the right setting for its own requirements and circumstances. One example for such a solution is the "Virtual Human Toolkit" provided by the University of Southern California [7]. The toolkit offers a large number of modules and features and it is available free of charge for academic research. However, even such comprehensive solutions have certain limitations, e.g., the hardware requirements or the multi-platform applicability, which is still in progress, or mobile device support.

In order to overcome these drawbacks, we decided to follow a similar approach and to design our own avatar generation toolchain which can be used in a more flexible way. Our solution is useable for different applications and/or services, especially for the Active and Assisted Living domain and the target group of older adults.

2. REQUIREMENTS ENGINEERING
Our avatars pursue the goal to be flexible and extremely versatile but at the same time also affordable. In order to achieve this goal we had to consider requirements from several fields during the design and development phase. The following sections provide an overview of the elaborated requirements from three different perspectives.

2.1 The End User Perspective
The avatar appearance is based on end user requirements which have been elaborated in several national and international AAL projects over the last 8 years. The following list summarizes the main characteristics regarding the appearance from the perspective of the end users (older adults) as elaborated in [8]:

PervasiveHealth 2016, May 16-19, Cancun, Mexico

DOI 10.4108/eai.16-5-2016.2264655

- Human figure (in contrast to comic-like avatars)
- Middle aged
- Female sex
- Neutral, casual business clothes (in contrast e.g., to nurse dress)
- Neutral surrounding (in contrast e.g., to a health care facility)
- The right balance between a character that radiates competence and professionality and a character that has a sense of humor and entertainment value skills.

2.2 The Economical Perspective

The economical perspective fosters reusability of the avatar. Thus, the overall goal is to build a flexible but natural-looking and anthropomorphic solution which can be used in different situations and scenarios without the need to modify or to re-implement the avatar all over again. The following aspects need to be considered:

- Basic gestures and body movements for the speech mode
- Basic gestures and body movements for the presentation mode (e.g., pointing at a presentation wall)
- Viewing direction towards the conversion partner (end user)
- Natural aspects, like head movements, synchronization of the lips or eye blinking
- Variety of moods (e.g., angry, sad, happy)
- Different scenes (e.g., in the office, living room, in front of the presentation wall, in the nature)
- Several camera positions (e.g., long shot, medium shot or close-up)
- Support of multiple languages and voices

2.3 The Technical Perspective

From the technical point of view it is crucial to design a system which can be used fluid and natively on different operating systems and platforms but also in form of pre-rendered video files, embeddable in classical web applications. The following points summarize the requirements from the technical point of view:

- Real-time client-based rendering on various platforms and devices including mobile devices
- Multiple real-time server-based rendering into video files
- Support of various resolutions and aspect ratios
- Real-time rendering of life-sized avatars (e.g., in Ultra HD resolution)
- Possibility to control the avatar via physical events (e.g., gestures or tangible objects) but also via software commands

3. THE AVATAR GENERATION TOOLCHAIN (AGT)

Avatars are in general complex structures consisting of different audio-visual elements. Moreover, they include spatial and temporal aspects and as already mentioned in the previous section, we expect them also to reach a high level of anthropomorphism. Considering these aspects leads to the need for utilizing different techniques and tools. In the following section, we discuss necessary steps and required tools towards a solution, which is able to provide universally applicable avatars.

3.1 Architecture

Figure 1 illustrates the overall architecture of our AGT. It consists of six main components (modeling, animation, audio processing, rendering, video generation, video delivery) classified in three different processing steps (pre-processing, processing and the optional post-processing). The pre-processing part defines the avatar's character and appearance as well as the domain and scenario-specific setting including the avatar's movements. The processing part is responsible for the audio processing and for the run-time animation and rendering. This part is influenced and controlled by internal and external parameters (e.g., the text which has to be spoken, the avatars mood, the intended camera position or the required resolution). The post-processing part is an optional part and just required in situations where a live rendering is either not possible or not applicable (e.g., if the avatar needs to be embedded in existing services or specific Web applications). The performed renderings are in this case stored as separate MPEG4 video files which can be delivered via the internet and remotely played back by any compatible video player.

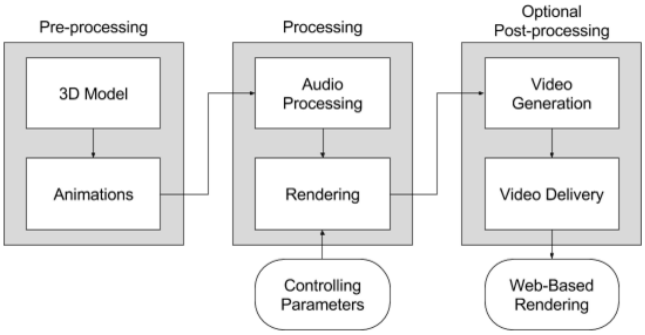

Figure 1: The overall architecture of the AGT conducted of 6 modules (modeling, animation, audio processing, rendering, video generation, video delivery), classified in three processing steps (pre-processing, processing and the optional post-processing).

3.1.1 Avatar's Appearance – The Pre-processing Part

The 3D Model

Our AGT uses the game engine Unity3D [9] and therefore supports conventional 3D model formats such as FBX, Collada or obj. In general, 3D models can be designed by a professional artist or even generated by existing software tools like MakeHuman [10] or Adobe Fuse CC [11]. However, in order to obtain the full feature set of the AGT the provided 3D model needs to fulfill following requirements:

- Definition of joints (rigging) responsible for the control of the body including head and eye rotations used for the animations.
- Blend shapes (blend shapes deform the geometry in order to create a specific look) for visemes, different facial mood expressions and eye blinking

Animations

As already mentioned, naturalness is one of the key factors for avatar-based interaction. Lifelike movements of the character are able to increase the natural perception and therefore increase the overall acceptance. From the technical point of view movements of the character can be defined by artists on a frame-to-frame basis in the 3D modelling tool, programmatically inside the game engine (e.g., Unity3D) or by utilizing motion capturing techniques. Motion capturing techniques allow also nonprofessionals to create natural looking animations. In the AGT solution the low-cost gaming device Microsoft Kinect [12] and the Unity plugin "Kinect with MS-SDK" [13] were used to capture several pre-defined movements (e.g., pointing, hand-supported explaining or waiting).

3.1.2 Avatar in Action – The Processing Part

Audio Processing

One crucial part of the AGT is the audio processing. According to the elaborated requirements the audio processing module is responsible for the main communication channel in a direct- as well as in an indirect manner. Direct communication refers to audio output in different languages and sex characteristic voices using an appropriate Text to Speech (TTS)-engine. The indirect communication refers to the synchronization of the audio and its visual representation, e.g., the lip synchronization.

The TTS-engine represents the central part of the audio processing module. Our solution is not limited to a specific speech synthesis. Depending on the target platform we are supporting native implementations (e.g., iOS 7 or Android speech synthesis) or external SDKs or cloud-based services, such as the Cerevoice solution [14]. Speech recognition is currently not supported because of different aspects influencing the accuracy, but future research is intended in order to provide a more practicable and natural solution for the end user.

Our AGT supports lip synchronization in two ways: lip synchronization according to a delivered audio file or according to concrete phonemes. In the easier but more inaccurate audio-file approach the synchronization is realized based on various features extracted from the audio file. The more sophisticated and more realistic approach is based on phonemes and their translation to corresponding visemes. Considering the German language, 15 visemes are enough to visualize the set of possible phonemes [15]. In our current implementation, we grouped these visemes in 7 groups. Consequently, the avatar requires 6 different blend shapes for using lip synchronization based on phonemes.

Rendering and the Influence of Controlling Parameters

Our solution uses the Unity3D game engine for rendering. Although there are multiple alternatives available on the market, there are only a few offering builds for multiple target platforms as well as a professional level-editor (e.g. [9, 16, 17]). The flexible licensing model of the Unitiy3D game engine, which also offers a free of charge version for non-commercial use, attracts a lot of developers also beyond the professional game community. Thus, a lot of Unity plugins have been developed over the last years and a large community has grown, which is very helpful in case of implementation problems or for general support. Moreover, Unity offers great flexibility due to its supported scripting languages.

We use these scripting languages and offer the user the possibility to control different parts of the character by sending different parameters over RabbitMQ [18] or HTTP requests. The user can control the avatar's characteristics by changing his mood or eye opening, which changes the avatar's facial expression over blend shapes. The user can also influence the avatar's behavior by his presence and behavior if a camera is available on the target device. Therefore, we provide, e.g., the possibility that the avatar follows the user with his eyes and head. Thus, the avatar actually looks towards the user while talking. Apart from the avatar's characteristics and behavior, the user can also control the visual appearance of the rendered scene like the background setting, e.g., the avatar is standing in an office or in front of a TV, or the camera position, which changes the perspective the avatar is shown of. Further, additional elements can be faded in like buttons or presentation areas to show a text or video.

3.1.3 Avatar Goes Video – The Optional Post-processing Part

Video Generation

The AGT has been designed in a way to support existing services and applications, where live rendering is either not possible or not applicable, as well. Thus, it is possible to deliver the avatars in form of video sequences, which can easily be embedded, e.g., via the HTML5 video tag. For this purpose, it makes sense to run the AGT on a server machine accepting HTTP requests with a high-end graphic card such as MSI Geforce GTX 980. From the technical point of view videos are generated by capturing rendered frames into single images. We use the ffmpeg tool [19] to build a continuous video out of the captured single images. The resulting video is finally combined with the corresponding audio track into one single MPEG4 file and stored into a specific directory on the file system for the video delivery module.

Video Delivery

The video generation process is decoupled from the video delivery process. The first HTTP client request initializes the video generation process and the second HTTP client request starts the delivery of the generated video. This two-step process allows the concurrent processing e.g., the simultaneous generation and the streaming of videos. For each client request a unique identification number is generated and via an URL immediately returned. The number identifies the request and its corresponding video. The client uses a second HTTP request to obtain the video specified by this identification number. In order to allow multiple clients (e.g., on multiple devices) to obtain the generated video simultaneously, we decided to rely on well-established Web server solutions such as the Apache Web server [20]. Our current implementation is not yet optimized nor tested towards a high throughput. In order to fulfill such requirements it may be necessary to adjust the AGT.

3.2 Affective Avatars

As already mentioned, naturalness is one of the key factors when interacting with avatars. Our perception about what is natural, life-like and anthropomorphic is influenced by various aspects e.g., the natural body movement, the natural rhythm and tone of voice or even the natural sensing and reproduction of moods, emotions and personality. So far, we have implemented several fragments to satisfy these aspects and requirements towards a more affective interaction, but future research and work in this field still needs to be done.

3.2.1 *The Representation of Moods*

Our AGT supports different moods which influence the avatar's facial expression as well as the pitch of the used voice. A change in the pitch of the voice is concretely realized in the used TTS-engine, but not all engines support this feature. The AGT supports (amongst others) following facial expressions: happy, angry and sad. These facial expressions are technically realized by using blend shapes which control the eyebrows, the eyes and the mouth corners.

3.2.2 *Avatar's Sense*

Reflection of emotions and moods support affectivity just partially. The sensing of the situation, environment and the opposite interaction partner is also crucial for an affective interaction. In order to find the right time to interact with the user, person and event recognition are important tasks. In our current implementation, we have focused on vision based approaches to capture certain situations. We use, e.g., conventional cameras for face detection to detect if a person is present before the interaction starts. We also use gesture recognition provided by Microsoft Kinect SDK to react, e.g., avatar reacts on a waving user by waving back. In combination with the speech recognition, tools like openSMILE [21] or EmoVoice [22] could be used to detect the user's emotions and to react accordingly. However, these are just examples how we plan to increase the affectivity of our avatar. As mentioned before, future research and development in these fields needs to be done.

4. RESULTS

Figure 2 illustrates three examples of avatar models in our AGT. The left female avatar (A) was made by a 3D artist whereas the other two avatars (B, C) were made by the above mentioned Adobe Fuse CC tool. In our opinion, this comparison illustrates that for the avatar-based interactions in the AAL domain also affordable automatically generated 3D models can be used. Figure 3 illustrates three out of eight different poses (asking (A), pointing (B) and the idle position (C)) and figure 4 illustrates three out of six facial expressions (angry (A), happy (B) and sad (C)) currently implemented in the AGT.

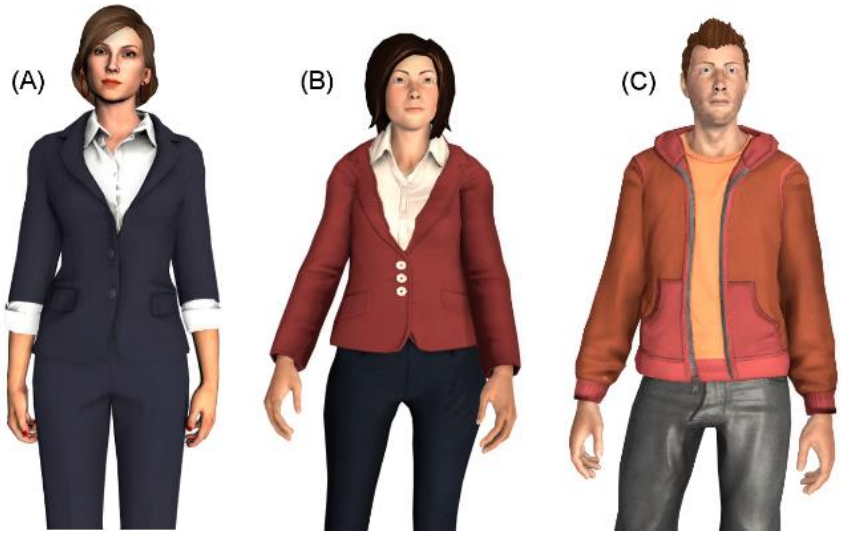

Figure 2: Illustration of three avatars used in the AGT and the comparison of an avatar made by a 3D artist (A) and avatars made by the Adobe Fuse CC tool (B and C).

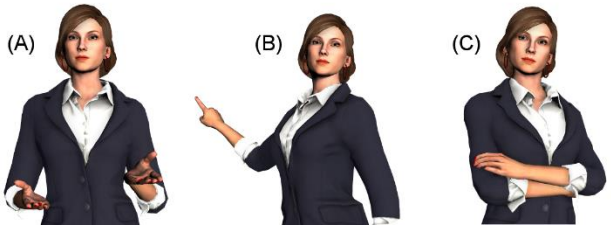

Figure 3: Illustration of three different poses (asking (A), pointing (B) and the idle position (C)) used in the AGT.

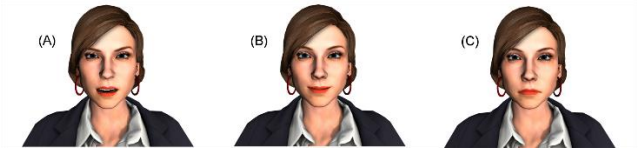

Figure 4: Illustration of three different moods (angry (A), happy (B) and sad (C)) used in the AGT.

5. LIMITATIONS

The aim of this work is to illustrate one possible solution for an avatar generation toolchain. In our literature review we have been focused on feasible and affordable state-of-the-art solutions and approaches. However, the scope of this work is not on an exhausting listing and detailed comparison of these state-of-the-art solutions and approaches. This work provides an overview of a possible setting on an abstract and reasonable level. A known limitation is the missing evaluation in terms of technical verification and validation, and in terms of usability testing involving end users. The description of the AGT on a more technical level and the reporting of evaluation results (i.e., technical and user related) are foreseen in future work.

6. ACKNOWLEDGMENTS

Much of the work reported here was within the international project YouDo which is co-funded by the AAL Joint Programme

(REF. AAL-2012-5-155) and the following National Authorities and R&D programs in Austria, Germany and Switzerland: BMVIT, program benefit, FFG (AT), BMBF (DE) and SERI (CH).

7. REFERENCES

[1] U. Diaz-Orueta, et al. “Role of cognitive and functional performance in the interactions between elderly people with cognitive decline and an avatar on TV.” In Universal Access in the Information Society 13(1), (2014) pp. 89– 97.

[2] M. M. Morandell, et al. “Avatars in assistive homes for the elderly.” In HCI and Usability for Education and Work. Springer Berlin Heidelberg, (2008) pp. 391–402.

[3] M. M. Morandell, et al. “Avatars@home.” In HCI and Usability for e-Inclusion, Springer Berlin Heidelberg, (2009) pp. 353–365.

[4] Y. Sakai, et al. “Listener agent for elderly people with dementia”. In 7th IEEE/ACM International Conference on Human-Robot Interaction (HRI), (2012) pp. 199–200.

[5] H.-H. Huang, et al. “Toward a memory assistant companion for the individuals with mild memory impairment.” In IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC), (2012) pp. 295–299.

[6] A. Ortiz, et al. “Elderly users in ambient intelligence: Does an avatar improve the interaction?” In Universal Access in Ambient Intelligence Environments, Springer Berlin Heidelberg, (2007) pp. 99–114

[7] Virtual Human Toolkit, University of Southern California, https://vhtoolkit.ict.usc.edu/, accesed March 10 2016

[8] M. Sili, et al. “Talking Faces in Lab and Field Trials." In HCI International 2015, Springer International Publishing, (2015), pp. 134-144.

[9] Unity3D https://unity3d.com/, accessed March 10 2016

[10] MakeHuman http://www.makehuman.org/, accessed March 10 2016

[11] Mixamo, Adobe Fuse CC. https://www.mixamo.com/fuse, accessed March 10 2016

[12] Microsoft, XBOX Kinect http://www.xbox.com/de-AT/xbox-one/accessories/kinect-for-xbox-one#fbid=CKcndWw8wBH, accessed March 14 2016

[13] RFilkov, Kinect with MS-SDK http://rfilkov.com/2013/12/16/kinect-with-ms-sdk/, accessed March 14 2016

[14] CereProc, CereVoice https://www.cereproc.com/, accessed March 10 2016

[15] B. Aschenberner et al. “Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis“, (2005)

[16] Unreal Engine https://www.unrealengine.com/, accessed March 10 2016

[17] CryEngine http://cryengine.com/, accessed March 10 2016

[18] Pivotal, RabbitMQ https://www.rabbitmq.com/, accessed March 14 2016

[19] FFmpeg, A complete, cross-platform solution to record, convert and stream audio and video., https://www.ffmpeg.org/, accessed March 14 2016

[20] Apache HTTP Server Project, The Apache Software Foundation https://httpd.apache.org/, accessed March 14 2016

[21] audEERING, openSmile http://www.audeering.com/research/opensmile, accessed March 14 2016

[22] University of Augsburg, EmoVoice https://www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/emovoice/, accessed March 14 2016