A 3D Audio-Visual Animated Agent for Expressive Conversational Question Answering

J.-C. Martin [*] , C. Jacquemin [**], L. Pointal [*], B. Katz [*]
[*]*LIMSI-CNRS, BP 133, 91403 Orsay*
[**]*Université Paris Sud, Orsay, France*

Abstract

This paper reports on the ACQA (Animated agent for Conversational Question Answering) project conducted at LIMSI. The aim is to design an expressive animated conversational agent (ACA) for conducting research along two main lines: 1/ perceptual experiments (eg perception of expressivity and 3D movements in both audio and visual channels): 2/ design of human-computer interfaces requiring head models at different resolutions and the integration of the talking head in virtual scenes. The target application of this expressive ACA is a real-time question and answer speech based system developed at LIMSI (RITEL). The architecture of the system is based on distributed modules exchanging messages through a network protocol. The main components of the system are: RITEL a question and answer system searching raw text, which is able to produce a text (the answer) and attitudinal information; this attitudinal information is then processed for delivering expressive tags; the text is converted into phoneme, viseme, and prosodic descriptions. Audio speech is generated by the LIMSI selection-concatenation text-to-speech engine. Visual speech is using MPEG4 keypoint-based animation, and is rendered in real-time by Virtual Choreographer (VirChor), a GPU-based 3D engine. Finally, visual and audio speech is played in a 3D audio and visual scene. The project also puts a lot of effort for realistic visual and audio 3D rendering. A new model of phoneme-dependant human radiation patterns is included in the speech synthesis system, so that the ACA can move in the virtual scene with realistic 3D visual and audio rendering.

Index Terms: *virtual agent, interactive control, expressive speech, real-time facial animation*

1. Introduction

Current applications of Animated Conversational Agents (ACA) are mostly designed for desktop configurations. Mixed and virtual reality applications call for coordinated and interactive spatial 3D rendering in the audio and visual modalities. For example, the rendering of the movements of a virtual character in a virtual scene (locomotion of the character or rotation of its head) and the 3D spatial audio rendering of the synthetic speech during these movements need to be coordinated. Furthermore, the expressiveness of the agent in the two modalities needs to be displayed appropriately for effective affective interactions, and combined with audiovisual speech. This requires experimental investigations on how to control this expressiveness and how it is perceived by users.

Since the 70's, research in audiovisual speech uses model-based approaches and image and video-based approaches [1]. Control models have been defined using visemes, coarticulation models [2], n-phones models grounded on corpora [3], or a combination of rule-based and data-driven articulatory control models [4]. For example, a set of four facial speech parameters have been proposed in [5]: jaw opening, lip rounding, lip closure and lip raising. Expressive qualifiers are proposed by [6] to modulate the expressivity of lip movements during emotional speech. Audiovisual discourse synthesis requires the coordination of several parts of the face including lips, but also brows, gaze and head movement [7]. With respect to facial animation [8], one reference formalism is MPEG-4 that defines a set of Face Animation Parameters (FAPs) deforming a face model in its neutral state [9]. FAPs are defined by motion of feature points, vertices that control the displacement of neighboring vertices through a weighting scheme. An overview of expressive speech synthesis can be found in [10].

Few of these studies and systems enable the coordinated display of 3D speech and facial expressions (directivity measurement of a singer was

computed by Kob and Jers), nor propose real-time interactive means for controlling expressive signals. Moreover, several systems are limited to a single face model where experimental studies and various interactive applications require the use of several models at different resolutions.

Our goal is to enable this coordinated spatial rendering in 3D of the speech and face of a talking head, and to provide means for real-time interactive control of its expressive signals. With respect to visual rendering, we aim at real time 3D rendering of different models at different resolutions that can be easily embedded in different mixed reality applications.

In this paper we describe the components that we have developed in order to meet our research goals and the current state of their integration within an experimental platform for conducting perception studies about audiovisual expressive communication.

2. Platform overview

Our components are integrated within a software platform made of five modules described in Figure 1. They communicate through simple text/xml data exchanged with UDP packets between modules and via files for voice synthesis sound. The application model is expected to produce tagged strings [message 1] for the multimodal module. Text is then sent [message 2] to our TTS which produces [message 3] a sound file and returns a set of lexeme descriptions.

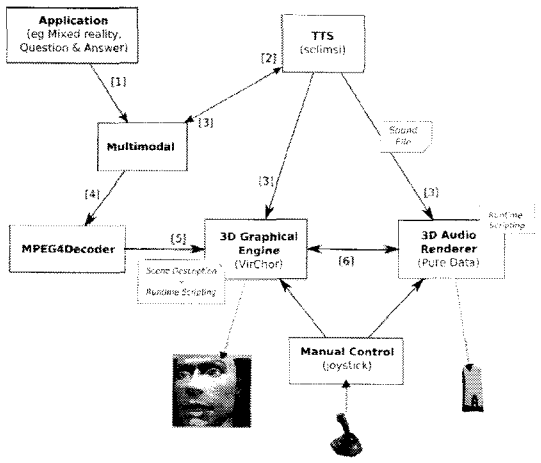

Figure 1: A platform for 3D audiovisual rendering and real-time interactive control of expressive signals.

From these lexemes the multimodal module requests [message 4] the MPEG4Decoder application to build the corresponding visemes set with timings and send them [message 5] to the VirChor engine. Sound file references are sent to PureData together with 3D coordinates of the virtual character for 3D audio rendering. Once VirChor and PureData have both visual and sound rendering information, they [message 6] start to play, considering possible real-time interaction coming from external events at runtime such as a joystick for interactive control of the agent.

3. Real-time facial animation

Audiovisual speech and interactive expression of emotions require a head animation with the following features:

- it should work real-time with a high frame-rate, because labial movements are fast and must be rendered with accuracy to allow for realism and possibly labial reading,
- it should offer easy keypoint edition and viseme modeling so that visemes can be quickly derived from existing visemes databases for new face models,
- the animation should be flexible enough to allow for interruptions in the case of dialogical interactions, new animation targets should be easily defined on the fly,
- face realism should be good enough to match requirements of interactive applications we aim at.

3.1. MPEG4 Animation

These requirements have oriented the design of ACQA. Since we wanted a flexible model that could be defined at various levels of details and on several face models, we chose to use pure facial synthesis in the line of MPEG4. We do not rely on motion capture techniques that highly depend on the actors used for the video recordings. Standardized syntheses have the advantage to rely on predefined sets of keypoints that can accept the Face Animation Tables (FATs) designed for other face models.

3.2. GPU Programming for Animation and Rendering

The key-point based animation relies on graphic processing unit (GPU) programming: the positions of the key-points are sent as parameters to the vertex shader, a program that computes the location of each vertex of the face mesh. The key-points transformations are used by the vertex shader to compute face animation through a skinning technique.

Each vertex transformation is a weighted sum of the transformations of at most 4 key-points. Since all the computations are performed in the GPU, no geometry is transferred over the graphic bus during animation. The frame rate is 155fps with a dual core Pentium at 3GHz and an NVIDIA GeForce 7800 while it is only 15 fps without GPU rendering.

GPU programming is also used to enhance the visual

quality of the animated face. Animated wrinkles connected to the displacement of some vertices are implemented through bump mapping techniques: depth textures that displace the normals and produce relief effects. Going pale or going red is also rendered in the GPU, through the fragment shader (as for wrinkles), by modifying the color of the skin is some specific areas of the face. More work remains to be done to render the skin translucency through BSSRDF.

Face animation is implemented in *VirChor* through scripting. For the audiovisual speech part, the animation timeline is defined, by the Text-to-Speech synthesizer. It provides the 3D engine with visemes and interpolation durations so that audio speech and lipsync can be synchronized. For the expression of emotions, the MPEG4 decoder similary produces target expressions together with interpolation durations. While lipsync is implemented as a successive list of targets, expression of emotions additionally requires durations that describe more precisely emotion spikes with raise, plateau, and decline. Since these two animation channels (audiovisual speech and emotions) are independent in their timing and animation models, they have been implemented independently and are triggered separately. Because of the limitation of input parameters for vertex shaders, key-point masks have been defined so that only minimal keypoint subsets are used in each channel. Dual-channel animation is controlled through UDP commands as indicated in igure 1 above. This will enable future experimental evaluation of several possibilities for blending the expressions of emotion with the lipsync.

3.3. Scripting and Interruptibility

In addition to efficient real-time dual-channel animation, the 3D engine also allows for interruptions. Through network commands, ongoing animations can be interrupted and redefined towards new targets with new timings. In case of interruption, the state of the ongoing interpolations is saved, and serves as a starting point for the new animation in order to avoid any noticeable discontinuity in the facial animation.

Animation and scripting are integrated in *VirChor*, a generic 3D engine with a formalism close to X3D. Since the engine is a multipurpose 3D renderer, animated faces can be embedded in various applications with variable behaviors and spatial layouts.

The quality of the visual animation has been evaluated through an application for interactive affective communication. The users were asked to reproduce simple or more subtle expressions through a tangible interface that controlled facial animation parameters. The qualitative evaluation shows that users appreciated the interface for its reactivity and its malleability. This shows that through its event-based scripting technique, the animation of the interface can be easily interrupted and redirected towards new targets.

Optimized animations through vertex shaders (Graphic Processing Unit programs) seem appropriate for an accurate synchronization with speech signal.

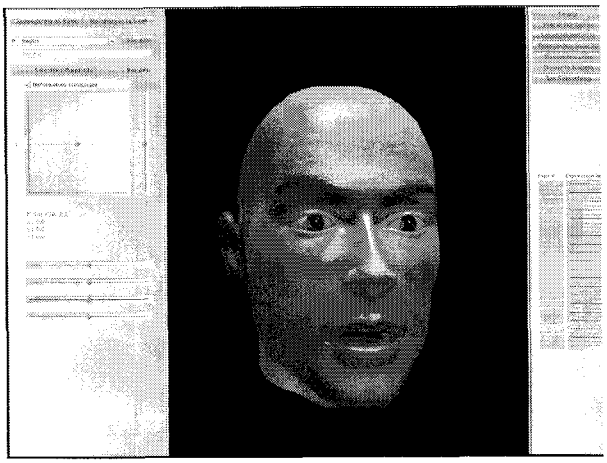

Figure 2. Interface for viseme and expression edition.

3.4. Face Edition

Viseme edition must be made easy so that new faces can be added without important workload. We designed a Java interface that can import existing FATs, modify them through linear or square sliders, and export them back as FATs that will be used as new targets for future animations. In addition, we also extended the Xface toolkit (http://xface.itc.it/). This toolkit is intended for the edition and modification of the weighting scheme of vertices on keypoints. We added the computation of keypoint weights based on Voronoï distances and the import/export to the XML formalism of VirChor.

4. Interactive control of expressive speech

The core component for synthetic speech generation is LIMSI's selection / concatenation text-to-speech synthesis system (SELIMSI). This core system has been augmented with three specific components: a radiation component accounting for relative head orientation and spatial motion of the virtual agent; a gesture control device allowing for direct control and animation of the virtual agent and a phoneme-to-viseme conversion module allowing for speech sounds

and lips movements' synchronization.

The text-to-speech synthesis system is based on optimal selection and concatenation of non-uniform units in a large speech corpus. The system contains two main components: text-to-phoneme conversion using approximately 2000 linguistic rules and non-uniform unit selection and concatenation. The annotated speech corpus (1 hour) contains read text and additional material such as numbers, dates, time. The selection algorithm searches for segments in the corpus according to several criteria and cost functions for ensuring optimal prosodic rendering and segmental continuity. The system receives text as input and outputs the audio speech signal together with a text file describing the phonemic content and the prosody of the utterance.

Realistic rendering of a moving speaking agent in 3D space requires the acoustic signal to be adapted according to the relative position and orientation of the speaker and listener. We conducted a broad study of time-varying speech radiation patterns. For speech synthesis, these results provide phoneme-dependant 3D radiation patterns that are integrated as post-processing of the synthetic speech signals. This enables visual movements of the agent to be accompanied by correlated audio movements.

Real-time interactive control of expressive speech signals is managed in the system by a direct gesture interface. The agent is considered as an "instrument" driven by an operator through a manual interface: as a joystick controls the head position, the expressivity of speech may be controlled using hand gestures via a graphic tablet. The gesture input enables interactive control of expressive parameters of the speech signal like fundamental frequency and voice source parameters (amplitude, spectral tilt, open quotient, noise in the source). Expression is controlled through subtle real-time variations according to the context and situation. In our application, as in real life communication, the vocal expression of strong emotions like anger, fear, or despair are more the exception than the rule. As such, the synthesis system should be able to deal with subtle and continuous expressive variations rather than clear cut emotions. Expressive speech synthesis may be viewed from two sides: on the one hand is the question of expression specification (what is the suited expression in a particular situation?) and on the other hand is the question of expression realization (how is the specified expression actually implemented). Our gesture interface is a research tool for addressing the second problem. Finally, phoneme to viseme conversion is handled by a set of rules, and audio and visual speech streams are synchronized using the prosodic description provided by the text-to-speech module.

5. Conclusion

We described a platform that addresses several challenges of mixed reality interactive applications: 1) coordination of the visual display of a 3D talking head and the corresponding spatial audio rendering in 3D of the synthesic speech, 2) real-time interactive control of expressive signals, and 3) management of different head models at different resolution.

6. References

[1] Bailly, G., Bérar, M., Elisei, F., Odisi, M.: Audiovisual Speech Synthesis. International Journal of Speech Technology. Special Issue on Speech Synthesis: Part II. 6 4 (2003)

[2] Cohen, M. M., Massaro, D. W.: Modeling coarticulation in synthetic visual speech. Models and Techniques in Computer Animation. AAAI/MIT Press (1993)

[3] Ma, J., Cole, R., Pellom, B., Ward, W., Wise, B.: Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of diviseme motion capture data. Computer Animation and Virtual Worlds 15 5 (2004)

[4] Beskow, J. Talking Heads - Models and Applications for Multimodal Speech Synthesis. PhD Thesis. Stockholm. 2003.
http://www.speech.kth.se/~beskow/thesis/index.html

[5] Reveret, L., Essa, I.:Visual Coding and Tracking of Speech Related Facial Motion.Hawai, USA

[6] Bevacqua, E., Pelachaud, C.: Expressive audio-visual speech. Comp. Anim. Virtual Worlds 15 (2004)

[7] DeCarlo, D., Stone, M., Revilla, C., Venditti, J.: Specifying and Animating Facial Signals for Discourse in Embodied Conversational Agents. Computer Animation and Virtual Worlds 15 1 (2004)

[8] Cohen, M., Beskow, J., Massaro, D.: Recent developments in facial animation: an inside view. AVSP'98 (1998)

[9] Ostermann, J.: Animation of synthetic faces in MPEG-4. Computer Animation'98 (1998) Philadelphia, USA 49–51

[10] Schröder, M. Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD Thesis. 2004.