3D Audiovisual Rendering and Real-Time Interactive Control of Expressivity in a Talking Head

J.-C. Martin, C. d'Alessandro, C. Jacquemin, B. Katz, A. Max,
L. Pointal and A. Rilliard

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
{martin, cda, jacquemin, katz, aurelien.max, laurent.pointal, rilliard}@limsi.fr

Abstract. The integration of virtual agents in real-time interactive virtual applications raises several challenges. The rendering of the movements of the virtual character in the virtual scene (locomotion of the character or rotation of its head) and the binaural rendering in 3D of the synthetic speech during these movements need to be spatially coordinated. Furthermore, the system must enable real-time adaptation of the agent's expressive audiovisual signals to user's on-going actions. In this paper, we describe a platform that we have designed to address these challenges as follows: (1) the modules enabling real time synthesis and spatial rendering of the synthetic speech, (2) the modules enabling 3D real time rendering of facial expressions using a GPU-based 3D graphic engine, and (3) the integration of these modules within an experimental platform using gesture as an input modality. A new model of phoneme-dependent human speech directivity patterns is included in the speech synthesis system, so that the agent can move in the virtual scene with realistic 3D visual and audio rendering. Future applications of this platform include perceptual studies about multimodal perception and interaction, expressive real time question and answer system and interactive arts.

Keywords: 3D animation, voice directivity, real-time and interactivity, expressiveness, experimental studies.

1 Introduction

Current applications of talking heads 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 binaural rendering in 3D 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 / video-based approaches [1]. Control models have been defined using visemes,

co-articulation 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 [11]), 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.

Section 2 describes the modules enabling real time synthesis and head orientation dependant audio rendering of expressive speech. The modules for 3D visual real time rendering of facial expressions are detailed in section 3. An overview of the platform integrating both audio and visual components enabling interactive control for experimental studies is provided in section 4. Although our short term goal is to use this platform for experimental studies, we explain in section 5 how we plan to use it for interactive applications.

2 Real time synthesis and 3D rendering 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 [12] 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.

Special attention is paid to the directional characteristics of the synthetic speech. 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. A broad study of time-varying speech radiation patterns has been conducted [13]. 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 [14]. 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 [15, 16]. 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. Experiments on hand gestures intonation reiteration showed that the performance levels reached by hand-made and vocal reiterated intonation are very comparable [17]. This could suggest that intonation, both on the perceptual and motor production aspects, is processed at a relatively abstract cognitive level, as it seems somehow independent of the modality actually used. Hand-controlled speech is a new tool for studying expressive speech synthesis [18], modeling expressive intonation, designing new interfaces for musical expression [19].

3 A module for 3D real time rendering of facial expressions

The graphic engine used for rendering our animated face is a multi-purpose, robust, and flexible environment for the synthesis and control of audiovisual speech. These

requirements have led us towards MPEG-4, the standardized formalism for face animation [20]. It relies on a predefined set of standard control points, a set of weights that defines the influence of the control points on face vertices, and interpolation mechanisms that are used to combine multiple control point displacements for audiovisual speech [6] and emotion synthesis [21]. We can use a face mesh at any level of detail (LOD) and define feature points through Xface [22]. Weight computation is automatic: weights are proportional to the inverse of a distance from feature points as defined in [23].

We intend to develop faces with lower LODs for applications with critical computing resources such as animations for portable devices. Conversely, for applications with high requirements for realism such as interactive expressive avatars for artistic applications or high resolution rendering, we can design faces with higher LODs. Variation of LOD will not have any impact on the scripting of the animation. In case of lower LOD, unused controlled points will be ignored. Automatic animations of additional control points through interpolations of existing ones will be added in case of higher LODs. Ongoing work on the enhancement of the skin and wrinkle rendering is also used to increase the realism of high definition faces.

With the same purpose of standardization in mind, graphical rendering relies on Virtual Choreographer (http://virchor.sourceforge.net/) (VirChor) a 3D engine based on an extension of X3D (http://www.web3d.org/). VirChor is a generic 3D engine, but has specific data structure and behaviors for face animation. The data structures define the control points and the associated weights on the vertices of a 2D or 3D mesh. The behavior is defined through an event based scripting language.

The steps for the design of a talking head are the following. First, XFaceEd (http://xface.itc.it/) is used to define the MPEG-4 control points and their weights on the vertices of the face mesh. XFaceEd has been enhanced to compute weights automatically through Voronoi distances. It has also been enriched to generate automatically the XML encoding of the face, its control points, and its weighting scheme in VirChor formalism. The automatic weight computation and the automatic code generation associated with the intuitive interface of XfaceEd allow for a fast and efficient transformation of any face mesh into an MPEG-4 ready mesh.

Second, a set of visemes and/or expressions can be defined as displacement vectors of the MPEG-4 control points. For this purpose, we have designed a Python-Tcl/Tk interface for the modification of the control point displacement to generate face animation targets. Through sliders, the control points are translated by sending commands to VirChor via UDP (Figure 1). In the final step, animations are defined as sequences of scheduled targets.

Since real time interactivity was also one of our requirements, facial animation in VirChor is implemented in the Graphic Processing Unit (GPU) [24]. The frame rate is approximately 10 times faster than when animation is performed in the CPU: 165 fps instead of 15 for an 8k face and 4k vertex mesh. Through GPU-rendering, both vertex position and normal are computed without the need to transfer the geometry through the graphic bus. At each time step, the interpolation coefficient is computed and sent to the GPU. If a keyframe is reached, the current state of the interpolation is saved as the future source frame in order to ensure a smooth transition between the current interpolation and the subsequent one. The source frame and the target frame are both sent as parameter tables to the vertex shader.

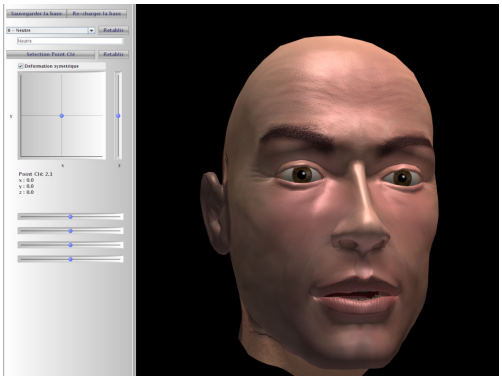

Figure 1. Displaying facial expressions using VirChor real time engine.

4 Platform overview

The components described in the previous sections are integrated within a software platform made of five modules described in Figure 2. 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. 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 produced by the TTS are also sent to the 3D audio PureData module, 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.

5 Conclusions and future directions

We have 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 binaural rendering in 3D of the synthesized speech, 2) real-time interactive control of expressive signals, and 3) management of different head models at different resolution.

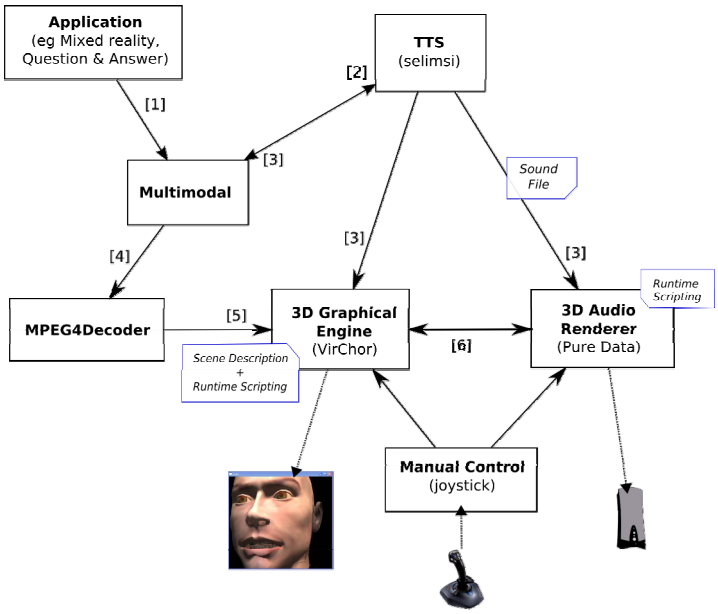

Figure 2. A platform for 3D audiovisual rendering and real-time interactive control of expressive signals in a talking head.

For the definition of synthetic visemes or expressions we rely on video-captured expressions [25]. We have recorded a first video corpus of visemes that suit our goals and our platform. This corpus was used for the definition of a first set of visemes that is used by the talking head. We intend to collect further data to improve the temporal description of the visemes and their co-articulation. We also intend to define expressions of emotion to be displayed with the 3D graphical engine. In order to study expressive speech, we plan to have two separate channels of animations sent to VirChor (one for lip movements, one for the expression of emotion) and explore different directions for their combinations thanks to our experimental platform.

The quality of the visual speech has not yet been evaluated *per se* but the animated face has already been used in an application for interactive affective communication [26]. The qualitative evaluation accompanying this work through a questionnaire on users' satisfaction shows that the users have appreciated the interface for its reactivity, its progressiveness, and its malleability. The smoothness of the animation and its high frame rate make it appropriate for an application on dialogic information querying such as ours. The 3D face can convey subtle and complex expressions, it can be accurately synchronized with speech signal, and its animation can be interrupted and redirected towards new targets.

We intend to use the platform not only for experimental studies but also for the design of interactive applications such as artistic applications or cooperative information retrieval applications. The information retrieval paradigm, exemplified by Internet search engines, which return a list of candidate documents matching a query made up of keywords, is evolving towards so-called Question Answering (QA) systems. It is important that the system indicates its level of confidence into its

answers or justify them with evidence in order to retain the user's trust. Future advances in QA are likely to take the form of *cooperative QA* [27], whereby answers can be augmented with information likely to be of interest to the user (e.g. suggestions, corrections to a previous result) [28]. In an attempt to make such systems more natural to use, speech can be used as the input and output modality [29, 30]. A recent study [31] shows that facial expressions can be better interpreted than linguistic cues for transmitting the certainty of a QA system. It is therefore interesting to consider adding a talking head as a modality to an interactive QA system, and to study the impact on the perceived cooperativity of the system. As an initial application, we intend to use the Ritel system developed at LIMSI [30]. Ritel is an open-domain dialogic QA system for French which uses speech as its input and output modalities, and which thus runs under strict time constraints.

Ongoing work on interactive control of expressive facial animation not only includes a joystick, but also an anthropomorphic tangible interface that dynamically controls the animation of the 3D face. The evaluation of this experiment showed that the tactile control of a face can be experienced as a unique affective communication medium. Since VirChor is a generic 3D engine for Mixed Reality, animated faces at different resolution can be made available for a large scale of applications without efforts. They are 3D component that can be embedded into 2D/3D scenes and interact with other scene parts.

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.
11. Kob, M., Jers, H.: Directivity measurement of a singer. Journal of the Acoustical Society of America 105 2 (1999)

12. Prudon, R., d'Alessandro, C.: A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation. 4th ISCA/IEEE International Workshop on Speech Synthesis (2001)
13. Katz, B., Prezat, F., d'Alessandro, C.: Human voice phoneme directivity pattern measurements. 4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan (2006) Honolulu, Hawaï 3359
14. D'Alessandro, C., D'Alessandro, N., Le Beux, S., Simko, J., Cetin, F., Pirker, H.: The speech conductor : gestural control of speech synthesis. eNTERFACE 2005. The SIMILAR NoE Summer Workshop on Multimodal Interfaces (2005) Mons, Belgium 52-61
15. Campbell, N.: Speech & Expression; the value of a longitudinal corpus. LREC 2004 (2004) 183-186
16. Martin, J.-C., Abrilian, S., Devillers, L.: Annotating Multimodal Behaviors Occurring during Non Basic Emotions. 1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005) (2005) Beijing, China 550-557
17. d'Alessandro, C., Rilliard, A., Le Beux, S.: Computerized chironomy: evaluation of hand-controlled Intonation reiteration. Interspeech 2007 (2007) Antwerp, Belgium
18. Le Beux, S., Rilliard, A., d'Alessandro, C.: Calliphony: A real-time intonation controller for expressive speech synthesis. 6th ISCA Workshop on Speech Synthesis (SSW-6) (2007) Bonn, Germany
19. d'Alessandro, N., Doval, B., d'Alessandro, C., Le Beux, S., Woodruff, P., Fabre, Y., Dutoit, T.: RAMCESS: Realtime and Accurate Musical Control of Expression in Singing Synthesis Journal on Multimodal User Interfaces 1 1 (2007)
20. Pandzic, I. S., Forchheimer, R.: MPEG-4 Facial Animation. The Standard, Implementation and Applications. John Wiley & Sons, LTD (2002)
21. Tsapatsoulis, N., Raouzaiou, A., Kollias, S., Cowie, R., Douglas-Cowie, E.: Emotion Recognition and Synthesis based on MPEG-4 FAPs. MPEG-4 Facial Animation. John Wiley & Sons (2002)
22. Balci, K.: Xface: MPEG-4 based open source toolkit for 3D Facial Animation. Working conference on Advanced visual interfaces (2004) New York, NY, USA 399-402
23. Kshirsagar, S., Garchery, S., Magnenat-Thalmann, N.: Feature Point Based Mesh Deformation Applied to MPEG-4 Facial Animation. IFIP Tc5/Wg5.10 Deform'2000 Workshop and Avatars'2000 Workshop on Deformable Avatars (2000) 24-34
24. Beeson, C.: Animation in the "Dawn" demo. GPU Gems, Programming Techniques, Tips, and Tricks for Real-Time Graphics. Wiley, Chichester, UK (2004)
25. Fagel, S.: Video-realistic Synthetic Speech with a Parametric Visual Speech Synthesizer. International Conference on Spoken Language Processing (INTERSPEECH/ICSLP 2004) (2004)
26. Jacquemin, C.: Pogany: A tangible cephalomorphic interface for expressive facial animation. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII 2007) (2007) Lisbon, Portugal
27. Benamara, F. WebCoop: un système de Questions-Réponses coopératif sur le Web. PhD Thesis. 2004.
28. Bosma, W.: Extending answers using discourse structure. Proceedings of RANLP workshop on Crossing Barriers in Text Summarization Research (2005) Borovets, Bulgaria
29. Boves, L., den Os, E.: Interactivity and multimodality in the IMIX demonstrator. IEEE conference on Multimedia and Expo (ICME 2005) (2005)
30. Rosset, S., Galibert, O., Illouz, G., Max, A.: Integrating spoken dialog and question answering: the Ritel project. Interspeech'06 (2006) Pittsburgh, USA
31. Marsi, E., van Rooden, F.:Expressing uncertainty with a Talking Head in a Multimodal Question-Answering System.Aberdeen, UK