High-realistic and flexible virtual presenters

David Oyarzun, Andoni Mujika, Aitor Álvarez, Aritz Legarretaetxeberria, Aitor Arrieta, María del Puy Carretero

doyarzun@vicomtech.org
Vicomtech Research Centre
P. Mikeletegi, 57
20009 San Sebastián, Spain

Abstract. This paper presents the research steps that have been necessaries for creating a mixed reality prototype called PUPPET. The prototype provides a 3D virtual presenter that is immersed in a real TV scenario and can be driven by an actor in real time. In this way it can interact with real presenters and/or public. The key modules of this prototype improve the state-of-the-art about these systems in four different aspects: real time management of high-realistic 3D characters, animations generated automatically from actor's speech, less equipment needs, and flexibility in the real/virtual integration. The paper describes the architecture and main modules of the prototype.

Keywords: 3D virtual presenters, mixed reality, real time animation

1 Introduction

The television is a world where technologies with some level of maturity are sooner or later applied. And 3D computer graphics are not an exception. In fact, 3D virtual images have been appearing together with real ones during the last years. For example, they are very common in some weather reports.

In 2006, an Australian TV channel (Channel Ten) went a step forward and broadcasted a talk-show called 'David Tench Tonight show', which was conducted by a 3D virtual character. This character performed interviews to real people, being a mixed reality system shown on live [1].

Its conceptual way of working was very simple. A real actor drove the virtual presenter in real time. He's voice caused a synchronized animation of the virtual presenter's lips and he drove the corporal animations by means of a motion capture system. The company behind it was Animal Logic.

Although the show was cancelled some months later, it was initially successful, being one of the 10 most watched programs in Australia[1].

[1] Statistics from eBroadcast:
http://www.ebroadcast.com.au/enews/Third_Time_Lucky_for_Seven_180806.html

Nowadays, the interest on this kind of mixed reality applications for TV continues. For example, companies like Nazooka have created 3D characters that have been broadcasted in different TV programs *via* mixed reality and it focus its business model in these kind of technology [2].
However, the applicability of 3D real-time virtual presenters to the TV environment presents some lacks yet. Although most of these lacks are not appreciated by the audience, they imply costs that could be reduced, and interfaces that are not very comfortable for the actors. Concretely, some of the main lacks are:

- Character flexibility. Companies provide the whole system, including the character modeling. Costs would be considerably reduced if TV producers could (re)use characters not created exclusively for the mixed reality system.
- Actor's comfort. Some of current applications require the actor wear a motion capture system or s/he needs to memorize and launch in real time a lot of facial and corporal animations *via* joysticks or keyboards.
- Equipment needs. Apart from the possible motion capture system need, some applications require a chroma system for creating the mixed reality. It implies space, high cost equipment and time for setting up the TV program.
- Mixed reality flexibility. Real cameras are usually fixed when the 3D virtual presenter appears. Cameramen cannot make zoom or change cameras while virtual character is on-screen. Being able to *play* with camera parameters would increase the mixed reality illusion.

This work presents a research project, called PUPPET whose requisites on the whole improve current state-of-the-art on these applications. The initial prototype developed solves the lacks explained above and so, it provides a low-cost and very flexible solution.

Sections below explain the TV virtual presenter prototype in detail, going into development related research lines in depth. So, section 2 present the state-of-the-art about systems related with this prototype and section 3 explains briefly our system architecture. Sections 4, 5 and 6 explain the main modules that solve the lacks mentioned above: section 4, the animation engine that provides character flexibility; section 5, the audio analyzer that improves the actor's comfort and section 6 the mixing module that reduces the equipment needs and improves the real/virtual flexibility. Finally, section 7 explains the resulting prototype tests and section 8 presents the conclusions and future work.

2 State of the Art

The prototype that is presented in this article involves several research fields like mixed reality, real time animation, speech technologies, etc.
Probably the applications that mix these fields in the most related way to this prototype are the works of Nazooka [2] and the David Tench Tonight show, developed by Animal Logic [1].

Nazooka presents some nice developments, however they present limitations regarding the change of camera parameters. That is, the camera remains fixed while the virtual presenter is visible.
On the other hand, David Tench Tonight was initially a successful TV program both from technological and audience level points of view. However, the actor had to use a motion capture system for reproducing all his/her movements in real time. It implied a complex setup and high hardware costs.
There is not many companies and prototypes for creating the mixed reality on TV, however, mixed reality applications are used in several fields such as marketing [3], leisure [4], medicine [5], learning [6], etc.
In this way, techniques for obtaining realistic mixing between virtual and real world has been widely studied:

- Lighting, for achieving the shadows of virtual objects over the real world and *vice versa*. Methods like *shadow mapping* [7] or *shadow volumes* [8] are used frequently.
- Occlusions, for calculating virtual world elements occluding real ones and *vice versa*. Different models like a 3D representation of the real world [8], stereo vision-based depth maps [9] or multi-camera 3D reconstruction [10] are used. Each of them has their advantages and disadvantages.

Regarding speech driven animations, most of the previous works are related to the lip synchronization and coarticulation. Phonemes analysis has to transform the speech into phonetic sounds [11] and map them to visemes (the visual representation of each phoneme). However, most of them concerns English [12, 13]. On the other hand, some approaches have been done for the generation of non-verbal facial expressions from speech. For example, works in [14, 15] generate head movements from fundamental frequency and real time speech driven facial animation is addressed in [16]. However, obtaining a coherent and realistic animation is a state-of-the-art field of research yet.

3 System Overview

The PUPPET prototype system architecture is designed for achieving independence among concrete input devices and the animation and mixed reality modules. In Fig. 1, the conceptual schema of the architecture is presented.

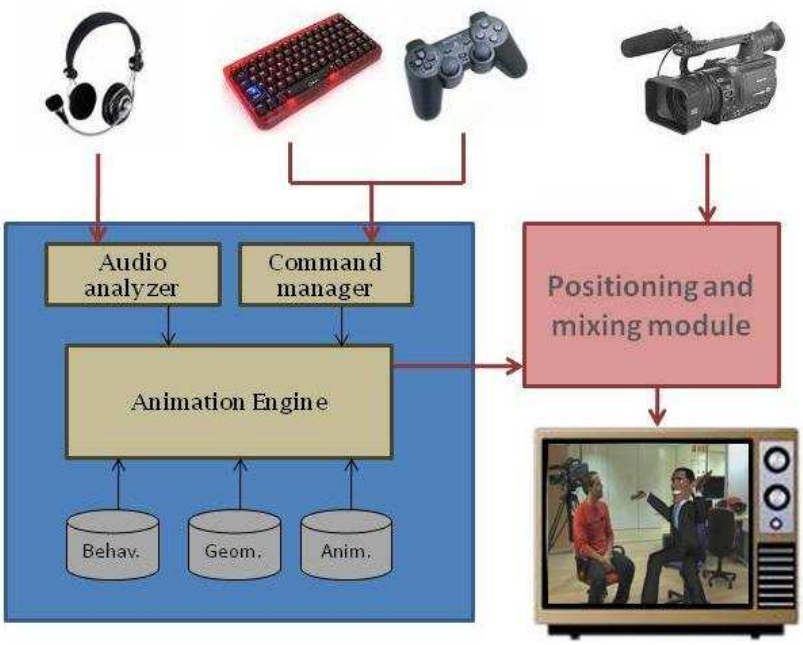

Fig. 1. Simplified architecture schema

Basically, the input devices are on the one hand the microphone, command devices like keyboards, joysticks, data gloves… and, on the other hand, the cameras of the TV studio.

Microphone input, that is, the voice signal, is managed by the *Audio Analyzer* module. Command devices inputs are retrieved by the *Command Manager*. It is an abstraction layer that avoids device-related dependences in the Animation Engine.

The *Animation Engine* creates the 3D virtual scene that is sent to the *Positioning and Mixing Module*. This module creates the visually correct mixing between the virtual scene and the real image, taking into account real camera changes. Sections below detail the technical aspects about the modules that improve the current systems' lacks. They are:

- The *Audio Analyzer*, which provides not only the analysis for synchronizing the real speech with the virtual presenter lips, but facial animations and expressions too.
- The *Animation Engine*, which is able to load characters created by means of commercial tools like Maya or Poser and animate them through standard BVH files.
- The *Positioning and Mixing Module*, which receives the virtual scene and the real camera parameters in real time and creates a coherent real/virtual mixing.

4 Audio Analyzer

The Audio Analyzer provides the synchronization between the actor's voice and the avatar lips as well as some facial expressions and animations.
The analyzer captures the speech signal from the input using a microphone and identifies the appropriate phonemes. As phonemes are recognized, they are mapped to their corresponding visemes. In a parallel process, the speech signal is processed by a pitch and energy tracking algorithm, in order to analyze its behavior and decide non-verbal facial movements. The virtual character is then animated in real time and synchronized with the speaker's voice. Therefore, the speech analyzer developed in this paper consists of four main sub-modules:

- The phoneme recognition system (described in the 4.1 subsection).
- The pitch/energy tracking sub-module (described in the **Error! Reference source not found.** subsection).
- The sub-module that sends the input audio to the recognition system and to the pitch/energy tracking algorithm in real-time. To develop this interface we used the ATK API [17].
- The communication interface between the speech application and the animation platform that was developed with sockets, based on the TCP/IP communication protocol. Through this module we fed the animation module with the recognized unit and facial movements for realistic animation.

Using this module the actor has not to be worried about the facial animation of the virtual presenter. All the aspects including lips synchronization and facial expressions will be automatically and coherently launched by the prototype when s/he speaks.

4.1 Real-Time Phoneme Recognition System

The main goal of this sub-module is to obtain the suitable data to animate the lips of the virtual character in real time. To obtain these data, we trained a triphoneme model using HTK Toolkit [18]. The corpus used for training and testing was Albayzin [19], a phonetic database focused to the development and evaluation of speech recognition and processing systems. It consists of 6800 sentences and 204 speakers. We divided this corpus in two data sets, training (4800 recordings) and test (2000 recordings). All of them are in WAV format (16 kHz/ 16bits/ mono). The feature extraction was performed over 25 ms segments every 10 ms. The parametrization of the speech signal was based on MFCCs, delta and delta-delta coefficients. The Spanish version of SAMPA was used as phoneme set for the recognizer. This set contained 29 phonemes plus the silence and short pause ones. Thiphoneme models were created, which consisted of non-emitting start and end states and three emitting states (except from the short pause model) using Gaussian density functions, whose number of components was increased until no further recognition improvements were observed. The states are connected left-to-right with no skips. The models were trained iteratively using the embedded Baum-Welch re-estimation and the Viterbi alignment, while the resulting was tested using a Viterbi decoder. Algorithm results are resumed in Table 1.

Training		**Testing**	
Correctly Words	*Word Accuracy*	*Correctly Words*	*Word Accuracy*
90.41 %	**84.20 %**	**81.18 %**	**71.23 %**

Table 1. Experimental results (phoneme recognition rate)

4.2 Pitch/Energy Tracking Sub-module

The recognized phonemes are mapped in real-time to their corresponding visemes in order to make the lip-synchronization process. This is the first step for the facial animation, which has been enriched using prosodic information of speech. A statistical model adapted to current speaker is created during the first steps of the recognition, based on the fundamental frequency (pitch) and energy of the speech signal, in addition to some related statistics. According to the values given in real-time by both pitch and energy trackers, some facial animations are shot, mainly related to the head and eyebrows up and down movement, and eyes and mouth more or less expressively movements.

Fig. 2. Several facial expressions automatically generated from the actor's voice

5 Animation Engine

The animation engine has been designed in order to obtain high-quality real time animations and at the same time, be able to load and animate characters not exclusively created for the PUPPET system.
The animation engine is divided on two main modules, the facial animation engine and the body animation engine.

- The facial animation engine uses advanced morphing techniques [20] for generating a high quality animation in real time. This is a technique quite extended and it creates the resulting animation by means of the linear interpolation among a set of predefined key faces. The animation engine includes an optimization for avoiding tests among vertices that are equal and it improves the computational cost in this way.
- The body animation engine implements a set of techniques whose objective is to achieve realistic movements with a low computational cost. For obtaining realistic movements the engine supports the loading of animations created by professional animators. They are loaded in the system by means of BVH files [21], a semi-standard format to which almost all commercial modeling

applications are able to export. And, moreover, it includes a set of optimizations that achieve their execution in real time in a standard desktop computer.

The animation is based on smooth skinning techniques. That is, the vertices of the geometry (or geometries) that conforms the virtual presenter are affected by a virtual skeleton. Transformations over this skeleton influence each vertex taking into account weights assigned to this vertex. This weights provides a way for avoiding *cracks* in the geometry and achieving smooth deformations.
The conceptual equation for the animation is:

$$v_r = v + w_i * M_r * v_i$$

Where v_r is the resulting vertex, v is the vertex with the previous transformations in the hierarchy, w_i is the asigned weight, M_r is the rotation matrix corresponding to current node and v_i is the vertex in its initial position.

For having no dependences regarding specific modeling formats a separation has been established between the geometrical and the smooth skinning information.

- Geometrical information. The 3D character can be loaded in any common geometrical format (3ds, obj, vrml, etc.)
- Smooth skinning information. A new file format, called SHF (Simple Hierarchical Format), has been designed for storing the skeleton and weighting information (Fig. 3). A plugin for Maya [22] to SHF has been developed. It allows designers to obtain this information from any Maya modeled character.

Fig. 3. SHF format file description

The animation engine relates both files in execution time and applies the BVH and morphing animations over them. In this way smooth deformations and high-realistic animations are obtained in real time over any virtual character designed with a standard modeling tool.

6 Positioning and Mixing Module

The Positioning and Mixing Module is designed for creating the mixed reality in a coherent way, without need of physical chroma systems or similar. It works in the opposite way than chroma systems. The virtual presenter background is one uniform color and the real scene replaces directly that color.

Moreover, since our application will be used in television, it would be useful to allow the cameras to translate and zoom. Then, the cameras will be able to follow either the real presenters or the virtual characters and get a more detailed view of them, without losing synchronization between real world and virtual worlds.

The camera is motorized and can be handled remotely. With a remote control three parameters of the camera can be changed: pan, rotation with respect to the vertical axis; tilt, rotation that makes the camera look up and down and zoom.

The robot that moves the camera is connected with the computer through a serial port and transmits the values of the parameters to the computer in real-time. The animation engine receives those values and with simple linear transformations parameters' values in degrees are calculated and transferred to the virtual camera.

In conclusion, the real camera is controlled remotely, but the virtual objects change their position in the screen coherently because of the information traffic between the real and the virtual camera. Fig. 4 shows some screenshots changing the camera parameters

Fig. 4. Playing with the real camera parameters: changes in translation and zoom (chroma system is not necessary; it is just for having a clean background. Virtual character's chair is real, non virtual)

7 PUPPET Prototype Tests

Modules described before conforms the PUPPET prototype. It has been tested by professional actors and staff from a Basque TV production company. They all agreed that the system is easy to use and avoids limitations and lacks found in the state-of-the-art.

The system has been tested in a standard desktop PC and using virtual characters from different sources. Concretely, along this paper, Fig. 2 shows some screenshots detailing the speech-based facial animation. The virtual character has been obtained from the Poser commercial tool [23].
Fig 4. showed changes in the parameters of the real camera, concretely translations and zooms, and the coherence between the virtual and real images. In this case, the virtual presenter, that is a caricature of Barak Obama, had been designed by a professional modeling company.

8 Conclusions and Future Work

This paper presents a prototype that provides a 3D virtual presenter that is immersed in a real TV scenario. It can be driven by an actor in real time and interact with real presenters and/or public.
The prototype solves some lacks existing in state-of-the-art similar developments. Concretely:

- Character flexibility. There is no need to model animations or virtual characters specifically for their use in the mixed reality platform. The platform supports standard file formats for animating the character and a new file format that supports the smooth skinning data store has been designed.
- Actor's comfort. The platform has been designed for avoiding needs about motion capture systems. It can be used just with a microphone and usual devices like keyboards or joysticks. Speech signal is used to automate not only the lip animation but some facial animations too.
- Equipment needs. There is no need to use chroma systems or similar. The computer creates the real/virtual mix directly.
- Mixed reality flexibility. Almost all current platforms that do not use chroma systems need to fix the camera, without moving. The platform of this work allows the cameraman to change the camera parameters (zoom, movements…) in real time.

Next steps are to include lighting and occlusion techniques that improve the realism and possibilities of the virtual presenter.

References

1. Animal Logic web page. http://www.animallogic.com

2. Nazooka web page. http://www.nazooka.com/site/
3. Metaio Augmented Solutions. www.metaio.com
4. Oda, O., Lister, L. J., White, S., Feiner, S.: Developing an augmented reality racing game. Proceedings of the 2nd international conference on INtelligent TEchnologies for interactive enterTAINment (2008)
5. Carlin, A.S., Hoffman, H. G., Weghorst, S.: Virtual reality and tactile augmentation in the treatment of spider phobia: a case report. Behaviour research and therapy (1997)
6. Tan, K.T.W., Lewis, E. M., Avis, N. J., Withers., P. J. : Using augmented reality to promote an understanding of materials science to school children. International Conference on Computer Graphics and Interactive Techniques (2008)
7. McCool., M.D.: Shadow volume reconstruction from depth maps. . ACM Transactions on Graphics (TOG) **19** (2000) 1-26
8. Fuhrmann, A., Hesina, G., Faure, F., Gervautz, M. : Occlusion in collaborative augmented environments. Computers and Graphics **23** (1999)
9. Fortin, P., Herbert, P.,: Handling occlusions in realtime augmented reality: Dealing with movable real and virtual objects. In Proceedings of the Canadian Conf. on Computer andRobot Vision, Vol. 54 (2006)
10. Matusik, W., Buehler, C., McMillan, L.: Polyhedral visual hulls for real-time rendering. in Proc. 12th Eurographics Workshop on Rendering EGWR'01, London (2001)
11. Lehr, M., Arruti, A., Ortiz, A., Oyarzun, D., Obach, M.: Speech Driven Facial Animation using HMMs in Basque. Proceedings of 9th International Conference on Text, Speech and Dialogue - TSD 2006, Brno, Czech Republic (2006)
12. Goldenthal, W., Waters, K., Van Thong, J.M., Glickman, O.: Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe. Eurospeech, Rhodes, Greece (1997)
13. Massaro, D., Beskow, S., Cohen, M., Fry, C., Rodriguez, T.: Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. AVSP, Santa Cruz, California (1999)
14. Deng, Z., Busso, C., Narayanan, S., Neumann, U. : Audio-based Head Motion Synthesis for Avatar-based Telepresence Systems. ACM SIGMM Workshop on Effective Telepresence (ETP) (2004)
15. Chuang, E., Bregler, C.: Mood swings: expressive speech animation. ACM Transactions on Graphics (TOG) (2005)
16. Malcangi, M., de Tintis, R.: Audio Based Real-Time Speech Animation of Embodied Conversational Agents. Lecture Notes in Computer Science (2004)
17. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book
18. Young, S.: The ATK Real-Time API for HTK
19. Casacuberta, F., Garcia, R., Llisterri, J., Nadeu, C., Pardo, J.M., Rubio, A.: Development of Spanish Corpora for Speech Research (ALBAYZIN). Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy (1991)
20. Alexa, M., Behr, J., Müller, W.: The morph node. Proceedings of the fifth symposium on Virtual reality modeling language (Web3D-VRML), Monterey, California, United States (2000) 29-34
21. Meredith, M., Maddock S.: Motion Capture File Formats Explained. Department of Computer Science, University of Sheffield (2001)
22. Maya Home Page.
http://usa.autodesk.com/adsk/servlet/pc/index?siteID=123112&id=13577897
23. Poser Home Page. http://my.smithmicro.com/win/poser/