A Web-Oriented Java3D Talking Head

O. Gambino¹, A. Augello¹, A. Caronia¹, G. Pilato², R. Pirrone¹, and S. Gaglio¹,²

¹ Università degli Studi di Palermo, Department of Computer Science, Palermo, Italy
{gambino, augello}@csai.unipa.it, alessandrocaronia@inwind.it, pirrone@unipa.it

² Consiglio Nazionale delle Ricerche (CNR), Palermo, Italy
{pilato, gaglio}@pa.icar.cnr.it

Abstract. Facial animation denotes all those systems performing speech synchronization with an animated face model. These kinds of systems are named Talking Heads or Talking Faces. At the same time simple dialogue systems called chatbots have been developed. Chatbots are software agents able to interact with users through pattern-matching based rules. In this paper a Talking Head oriented to the creation of a Chatbot is presented. An answer is generated in form of text triggered by an input query. The answer is converted into a facial animation using a 3D face model whose lips movements are synchronized with the sound produced by a speech synthesis module. Our Talking Head exploits the naturalness of the facial animation and provides a real-time interactive interface to the user. Besides, it is specifically suited for being used on the web. This leads to a set of requirements to be satisfied, like: simple installation, visual quality, fast download, and interactivity in real time. The web infrastructure has been realized using the Client-Server model. The Chatbot, the Natural Language Processing and the Digital Signal Processing services are delegated to the server. The client is involved in animation and synchronization. This way, the server can handle multiple requests from clients. The conversation module has been implemented using the A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) technology. The output of the chatbot is given input to the Natural Language Processing (Comedia Speech), incorporating a text analyzer, a letter-to-sound module and a module for the generation of prosody. The client, through the synchronization module, computes the time of real duration of the animation and the duration of each phoneme and consequently of each viseme. The morphing module performs the animation of the facial model and the voice reproduction. As a result, the user will see the answer to question both in textual form and in the form of visual animation.

1 Introduction

A talking head is a 3D animated model aimed at simulating a human head. The model is capable to emulate the articulation of word pronunciations, by synchronizing an audio flow with labial movements and reproducing emotional expressions of the face. The application fields of this technology include telecommunication [1, 2],
teaching [3, 4], speaking rehabilitation [5] and all systems requiring a friendly interaction with a human user. The problems related to a talking head realization can be summed up in three main tasks: viseme-phoneme association, animation and synchronization. The phoneme is the smallest unit of distinguishable sound: the concatenation of phonemes forms words and phrases.

Conversely, a “viseme” identifies the equivalent contractions of face muscles that produce the phoneme sound.

More phonemes can be expressed by the same viseme: the viseme-phoneme association task consists of detecting this connection. Furthermore, the visemes must be linked together in order to simulate the articulation of entire words pronunciation movements. The subsequent task is to smooth the lips movements in order to produce a fluent video stream. The most used techniques to accomplish this task are Keyframe Interpolation [6], Muscle-Based Facial Animation [7] and Direct Parametrization [8].

The last task, consisting in the correct synchronization between sound and animation flows, is obtained adapting the duration of the viseme for the whole duration of the current phoneme. This produces a pleasant audiovisual effect. This step can be realized using a hybrid architecture based on Hidden Markov Models and Artificial Neural Network (HNN/ANN) [9].

This method cannot be applied to realize a real-time synchronization since it requires heavy computational resources. In order to overcome this problem, we introduced a “linear synchronization”. The talking head presented in this chapter is a real-time interface based on Java3D with full 3D capabilities and it is embedded into a simple conversational agent whose knowledge base is accomplished by pattern-matching rules. It uses the Java3D Morphing technique provided by the animation engine of Java3D and exploits the linear model to perform the Synchronization task.

2 State of the Art

In this paragraph we give an overview of the main works related to talking heads and conversion from text inputs to audiovisual streams. As an example, a 3D head model interactive interface to express a sequence of phonemes and emotions is presented in [10]. Rational Free Form Deformations simulate abstract muscle actions. This technique moves from Free Form Deformation (FFD) depicted in [11]; rational basis functions are included in the analytical formalisation. The face mesh is subdivided into regions and control points impose the deformation. A basic facial motion parameter is called Minimum Perceptible Action (MPA). Each MPA is associated to a set of parameters. Such parameters modify the face mesh. An MPA can be thought as an atomic action unit like the Action Unit (AU) of the Facial Action Coding System (FACS) [12], but it also includes non-facial muscle actions, such as eyes movements.

In [13] a 3D head model is constructed starting from a cylindrical acquisition performed with a Cyberware scanner. A generic face mesh is fitted on the scanner data thanks to an image analysis technique aimed at discovering local minima and maxima of the sampled data. Some facial features (like nose, eyes etc.) are