Expressive Talking Heads:
A Study on Speech and Facial Expression in Virtual Characters

PAULA SALGADO LUCENA[1], MARCELO GATTASS[1], LUIZ VELHO[2]

[1]Departamento de Informática – PUC-Rio - Rua Marquês de São Vicente, 225, Gávea, 22453-900, Rio de Janeiro, RJ, Brasil
Caixa Postal 38097 Telefone: (55 21) 3114-1001
{pslucena, gattass}@inf.puc-rio.br
[2]IMPA–Instituto de Matemática Pura e Aplicada - Estrada Dona Castorina, 110, 22460 Rio de Janeiro, RJ, Brasil
lvelho@visgraf.impa.br

Abstract. This article presents the work carried out for a Master thesis and its main contributions. The thesis focuses on facial animation, more specifically on facial animations with synchronization between speech and facial expressions. First, the motivation behind this work and some new concepts are presented. Then, a taxonomy used to analyze talking head parameters is proposed. Finally, the implementation of a new talking head system is described.

1 Introduction

Facial animation has been a source of great interest in the last years. This is not a new endeavor; initial efforts to represent and animate faces using computers occurred over 20 years ago. But why should we animate the human face?

The human face is interesting and challenging because of its familiarity. Essentially, the face is the part of the body we use to recognize individuals. As a consequence, human facial expressions have been the subject of research by the scientific community. However, the ability to model the human face and to animate the subtle nuances of facial expressions is still a challenge in Computer Graphics.

As well as the human face, speech is an important element for human communication. It can be naturally described through phonetic properties. *Phonemes* are distinct sounds of a language, such as [*w-uh-n*] of the English word *one*. Each phoneme can be visually represented by means of a facial expression related to a certain format of the mouth, called *viseme*. With visemes and phonetic representation (e.g. phoneme description, duration and intonation), it is feasible to build a facial expression for each small speech segment.

Among the diverse types of facial animation systems developed, those that involve the facial animation of the virtual character combined with speech synchronization are distinguished as directly related to this work. These kinds of systems are known as *talking head* or *talking face*.

For the development of a talking head system, it is necessary to identify the possible approaches for speech and face modeling. The models used will influence not only the way the animation is performed, but will also affect the system's interactivity. We have developed an application that receives as input a text composed by the character's speech and gender, as well as language and emotion parameters, and generates as output, in real time, the animation of a virtual character uttering the input text with speech synchronization and expressiveness. This system, called "Expressive Talking Heads", explores the naturalness of facial animation and seeks to offer a real-time interactive interface. The Expressive Talking Heads system can run as a stand-alone application or connected to Web browsers. It was designed and developed to provide a platform- and operating system-independent solution.

2 Taxonomy for a Talking Head System

In this section we will propose a taxonomy for talking head systems, defined according to speech, face and execution parameters. Our goal is to explore these parameters in order to understand the different approaches that can be used to develop a talking head system. With this taxonomy it will be possible to classify talking head systems and to verify whether there is an ideal approach for each parameter in order to build a talking head system with speech synchronization and expressiveness.

2.1 Speech

The speech in a talking head system is directly related with the audio that is reproduced with the facial animation. There are two approaches that must be considered: the audio can be synthesized or captured.

In the first approach (synthesized voice), the audio is generated by a system known as *speech-synthesis* or *text-to-speech*. A text-to-speech synthesizer (TtS) [4] is a computer-based system that should be able to read any text aloud. Two basic modules form this system. The first one is the Natural Language Processing module (NLP), capable of producing a phonetic transcription of the text together with the desired intonation and rhythm (often referred to as *prosody*). The second one is the Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech.

The second approach is based on a captured voice. The audio is captured by either recording the speech of a person talking or by reusing a recorded audio. This approach tends to provide a more realistic effect than the synthesized voice.

Independently from the approach used for the voice, a common strategy in animation has been to analyze the phonemes that compose the speech. Phonemes are valuable information, mainly for lip synchronization (*lip-sync*), because they allow a precise synchronization between lip movements and visemes.

Actually, instead of simple phonemes, some variations are frequently used, such as *diphones* and *triphones*. Diphones are short audio sequences which sample the transitions between one phoneme and the next. Triphones are sets of three phonemes. The reason for using diphones or triphones instead of phonemes derives from the need to obtain a visual speech dynamics that forces co-articulation. Sometimes lip positions for the same phoneme change according to the context in which the phoneme is inserted.

2.2 Face Animation

In a talking head system, the face and animation modules are complex and have significant importance. There are some factors that must be taken into account for human face reproduction, such as the way the face is represented and the style, directly implying the degree of apparent realism of the face.

Similarly, a possible classification for human face representation is whether the face is generated by means of a captured image or defined by a geometric model.

In the first approach, captured image models could be two-dimensional or three-dimensional. For the second approach, a geometric model is defined to represent the face. In addition, it is possible to make use of textures, which can reproduce a human face in greater detail.

Once the face is modeled, the next step consists of its animation. The animation - not only facial animation but the animation of any object - is directly associated with movement, i.e., it is a dynamic mechanism. For a face defined by captured images, it is possible to produce the facial animation by means of image operation techniques, such as morphing. On the other hand, for the polygonal face model, animation is possible by applying facial muscle techniques [2].

The face style can be realistic or caricatural. In the realistic style, the face seeks to be similar to the human face. In contrast, the caricatural style distorts or exaggerates the face, usually for cartoon characters.

2.3 Execution

There are two approaches for system execution. The first approach is to execute the application in real time. For talking head systems, this approach usually involves user interaction. The main idea is that the user supplies the input speech and the system generates a facial animation of the virtual character enunciating the input text. It is important to point out that, in this form of execution, the input data must be able to be computed in parallel with the animation produced as output.

The second approach is to execute the application in batch mode. Unlike the real-time approach, this method is not associated with interactivity; batch mode is a passive approach. The main idea is that the user supplies an input and the system is responsible for capturing and processing it. The output generated by this first step will be the input for the lip-sync module. Finally, a video stream is produced for the talking head animation. The resulting video can be presented when desired.

3 Expressive Talking Heads

This section presents an overview of the Expressive Talking Heads system. The system has three modules: input synthesis, face management and synchronization. Figure 1 illustrates the overview of the system with these modules. The system is implemented in Java Language Programming and some functions are implemented in Scheme. There are also other subsystems incorporated to some modules of Expressive Talking Heads.

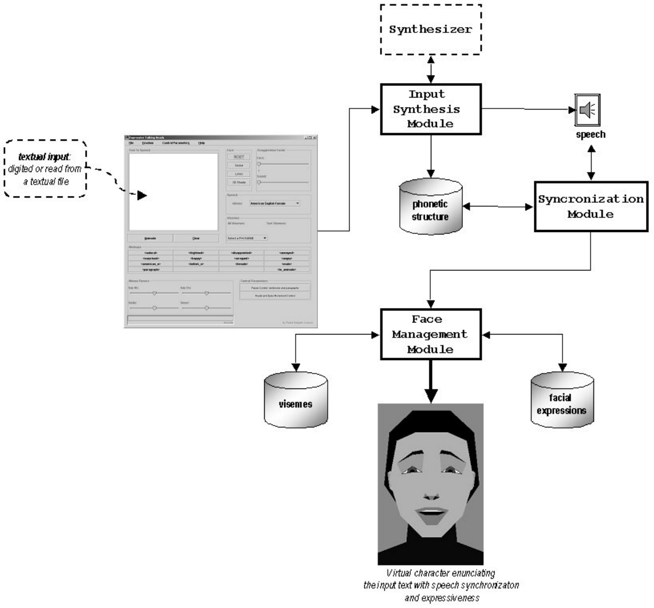
Synthesizer
Input Synthesis Module
speech
textual input: digited or read from a textual file
phonetic structure
Syncronization Module
Face Management Module
visemes
facial expressions
Virtual character enunciating the input text with speech synchronizaton and expressiveness

Figure 1: An overview of “Expressive Talking Heads”.

3.1 Input Synthesis Module

The input synthesis module is responsible for capturing and treating the input text provided by the user, and generating as output a data structure containing the fundamental units (phonemes, duration, emotion, etc.) to generate the facial animation corresponding to the input text.

The input text is provided through a markup language that includes information about the character's emotion, the voice's gender (currently, male and female adults) and the speech's language (currently, American English and British English). This text can be supplied by means of user input or through a file. The synthesis module interprets the markup text by means of a *parser*, separating speech and control information.

With the text and the markups analyzed by the parser, the system's synthesizer uses the features provided by the subsystems Festival [1] and MBROLA [5] to obtain a phonetic description and the digitalized audio corresponding to the character's speech. In the combination of these two synthesizers, Festival works as the Natural Language Processing unit (NLP) and MBROLA works as the Digital Signal unit (DSP). The advantage of this union is the acquisition of a TtS synthesizer that offers a multilingual platform (MBROLA's contribution). This flexibility is important to Expressive Talking Heads because the language and the gender can be system parameters selected by the user for the text synthesizer, enhancing the system's expressiveness.

The Festival-MBROLA subsystem is used in server mode. The Expressive Talking Heads system establishes a TCP connection with the TtS synthesizer. Through this connection, Expressive Talking Heads sends *Scheme* commands to the synthesizer. Because most of the requests are sets of Scheme commands, groups of functions were developed in this language and put together with the Festival-MBROLA synthesizer.

Still in this module, depending on the user-selected option concerning the system's interface, the synthesizer adds a especial treatment for pauses among sentences in order to produce a more realistic generated speech audio.

We have defined and implemented three approaches for pause treatment: block-to-block, sentence-to-sentence and sentence-block. The first approach uses the Festival standard treatment, where all sentences have the same pause duration value. The other two approaches randomly change the pause duration among sentences. This is done by intercepting Festival's output and editing the phonetic description before sending it to the audio file generation. These approaches differ on the time of the request for the audio generation: the sentence-to-sentence approach sends each sentence individually both to Festival and MBROLA, while the sentence-block approach sends sentences to Festival, then these sentences are concatenated forming a text block and, finally, this block is sent to MBROLA.

3.2 Face Management Module

The face management module is responsible for the face control of Expressive Talking Heads. It links the system's graphic interface and synchronization module (Section 3.3) with the Responsive Face subsystem [3].

In the system's initialization, this module provides the database of visemes and facial expressions, as well as a special database containing the phoneme-viseme mapping. During the animation processing the synchronization module makes requests to the face controller to get information from such databases. The face controller also receives requests to activate facial muscles.

The face in the Expressive Talking Heads system is modeled by a three-dimensional polygonal mesh that is an inheritance of the Responsive Face project [3]. The facial muscles are built by means of vertex grouping. There are twelve muscles, five controlling eye movements, three controlling head positions and four controlling mouth movements. These muscles are used to animate the face by applying a value in the interval of [-1.0, +1.0] for each muscle.

The Expressive Talking Heads character has a simple modeling, with minimal controls, but it supports rich expressiveness. In addition, Expressive Talking Heads produces an animation of the face enunciating the text with all components synchronized.

We have defined a group of sixteen visemes to represent lip positions during the speech. As already mentioned, four muscles specify mouth (lip) positions, each one having a specific acting area.

Expressiveness is an element that enriches facial animation. In the Expressive Talking Heads system this element is explored in a preliminary and simplified manner.

From the already defined facial expressions, the value of each muscle is captured for each individual expression, and during the animation these values are applied on the facial muscles. Figure 2 (a) illustrates the frightened emotion and the Figure 2 (b) illustrates the polygonal mesh for this emotion showing the muscles for the eye, mouth and head components for the frightened expression.

Independent from the expressiveness element, we also developed in the Expressive Talking Heads system a function to treat and control the movement of facial components during the speech mechanism. There are five approaches for this: no movement, eye movement only, head movement only, eye and head movement without synchronization, and eye and head movement with synchronization.

3.3 Synchronization Module

The synchronization module is responsible for the synchronization between speech and face components. This mod-

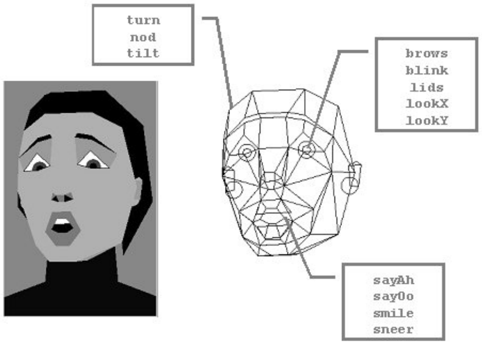

Figure 2: An emotion and the facial muscles.

ule has the greatest complexity, because Expressive Talking Heads intends to be a system where the facial animation allows user interactivity in real time.

The operation of this module depends on the output generated by the input synthesis module (audio and animation data structure) and also on fundamental units of the facial management module's databases (visemes, facial expressions and phoneme-viseme mapping). The main idea behind this module is that, in parallel with the audio reproduction, the system must determine for each time instant the phoneme that has been uttered and the state of the character's emotion. With this information, it is possible to match the phoneme with the respective viseme and to map in the corresponding facial expression. With such data it is possible to generate the animation of facial muscles synchronized with speech.

Once the audio presentation starts, the audio controller sends this information to the animation controller, so the face controller starts the phoneme treatment. In parallel with the animation controller execution, the movement of facial components is occurring. At the end of the interaction, the animation controller verifies the values for other facial muscles and applies the transformations to the facial structure.

4 Conclusions and Future Works

Facial animation is an important area of Computer Graphics and is continuously growing. In particular, the work for the Master thesis presented here has defined a set of parameters that can be used for classifying talking head systems and developed techniques for building a talking head system.

The thesis provides three main contributions. The first is the research and analysis of talking head basic components: speech, face and animation. The second contribution resulted in a taxonomy for talking head systems. Finally, the third and main contribution of this work is the development of the Expressive Talking Heads system.

The system developed has brought some technical advances. One was the integration of existing subsystems: Festival, MBROLA e Responsive Face. As a consequence of this integration, another contribution was the configuration of the Festival and MBROLA synthesizers to work together, the first as an NLP unit and the second as DSP. Finally, other advances include pause treatment to obtain more realism on the generated speech, a simplified extension of the HTML language to define a new markup language for character emotion, voice gender and speech language, the definition of different approaches to treat the movement of facial components, and the development of a viseme database.

There are two main topics for future works: the development of applications based on the Expressive Talking Heads system, such as 3D chats, virtual online shopping and distance learning, and a deeper research and implementation of expressiveness aspects, such as the incorporation of emotions in the speech and the use of other approaches for speech, such as speech recognition.

Acknowledgements

The first author would like to acknowledge the Tecgraf/PUC-Rio laboratory for technological support.

References

[1] Alan Watt, Paul Taylor, *The Festival Speech Synthesis System*, University of Edinburgh (1997)

[2] Frederic I. Parke and Keith Waters *Computer Facial Animation*, A K Peters, Ltd. (England, 1996).

[3] Ken Perlin *Responsive Face*, Media Research Lab, New York University, http://mrl.nyu.edu/ perlin/demox/Face.html (1997).

[4] Thierry Dutoit, *A Short Introduction to Text-to-Speech Synthesis*, TTS Research Team, TCTS Lab, Facult Polytechnique de Mons (Belgium, 1997).

[5] Thierry Dutoit and et al., *The MBROLA Project*, TCTS Lab da Facult Polytechnique de Mons, http://tcts.fpms.ac.be/synthesis/mbrola.html, (Belgium, 1997).