3D Photo-realistic talking head for human-robot interaction

Carlos Simplício
Institute of Systems and Robotics – Polo II, Universidade de Coimbra, 3030-290 Coimbra, Portugal
School of Technology and Management - Institute Polytechnic of Leiria, 2411-901 Leiria, Portugal
simplicio@estg.ipleiria.pt

Diego Faria & Jorge Dias
Institute of Systems and Robotics - Polo II Universidade de Coimbra, 3030-290 Coimbra, Portugal
{diego, jorge}@isr.uc.pt

ABSTRACT: This article describes our preliminary work for human-robot interaction. We present the initial stage of a realistic talking head system with facial expressions skills and phoneme-viseme synchronization. Facial expressions are the changes in response to a person's emotional state, intention, or social communication and they are generated by contractions of facial muscles, which results in temporally deformed facial features such as eyelids, eyebrows, nose, lips, and often are revealed by wrinkles and bulges. For the development of a realistic talking head, we have used a laser scanner to acquire the 3D physical structure of a human head. In a realistic talking head is important not only the location of facial actions, but their intensity as well as their dynamics. We are integrating our talking head in a social robot to interact with the humans beings.

1INTRODUCTION

Facial animation has been a source of great interest in the last years. In the reality, this is not a new endeavor; initial efforts to represent and animate faces using computers occurred over 30 years ago (Parke 1974), (Parke 1972).

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 a 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.

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. Each phoneme can be visually represented by means of a viseme – a facial expression related to a certain format of the mouth. With visemes and phonetic representation (e.g., phoneme description, duration and intonation), it is possible to build a facial expression for each small speech segment.

In 1978, Paul Ekman and Wallace Friesen (Ekman & Friesen 1978) carried out a detailed study of the human face and developed the *Facial Action Coding System* (*FACS*). It is a system to measure and classify the different facial expressions – the *Action Unit's* (*AU's*).

Our talking head system can run as a stand-alone application or as a front-end of a social mobile robot. Thus we intend to reproduce the naturalness of facial expressions and to offer a real-time interactive interface.

2MAIN MODULES IN A TALKING HEAD

When developing a talking head different approaches can be exploited. The decisions must be taken in function of the objectives in mind. In this section we briefly justify some of our options: (1) techniques used to model the head physical structure; (2) animation techniques and resources required to do it and (3) speech generation and lip-sync (synchronization between the sounds and the lips movements').

2.1*Techniques to Model the Physical Structure*

When the objective is to obtain a realistic talking head, the physical structure aspect is very important.

Normally, three techniques are used in modeling the head: (1) the definition of a geometric model through a CAD-like application, (2) the utilization of a 3D laser scanner or (3) just the use of video cameras.

We are interested in the two first methods, where, normally, the head model is composed by a mesh of polygons (to increment the realism, textures can be applied to the mesh). As the structures are analogous, we can use the some animation techniques in both.

2.2*Animation Technique and Resources Used*

After the physical structure of the head is modeled, it is necessary to create the animation – a set of dynamic actions which are associated with movements.

Our option to generate the animations was to perform a hybrid technique: a parametric and / or muscles based animation.

Muscle animation methods (Parke & Waters 1996), (Waters 1987) are very attractive by various reasons. Maybe the principal is (after the muscles allocation) the easiness to create expressions; as the muscles used mimics the human anatomy, each of them have a highly specific function. By other way, as each face expression can be generated by the action of just a muscles group, normally a reduced number of issues must be changed to obtain the intended results. When we think in terms of interaction, it is a great advantage in relation to other methods (Alexa et al. 2000), (Wang 1993), (Rydfalk 1987), (Parke 1974), (Parke 1972). With this method, the skin elasticity can be modeled in the muscle action itself. If we considerer that other methods need to solve a set of differential equations to simulate the skin elasticity, an example is the spring mass algorithms (Yuencheng et al. 1995), this is a major gain in computation time.

Moreover, also with the interaction in mind, it is necessary to keep the system as simple as possible. In our particular case the movements of specific head parts (i.e., eyes, jaw and neck) have not performed by muscles – in these cases a parametric animation is used.

Relatively to the resources involved, there are two hypotheses: (1) batch mode or (2) real time mode.

In batch mode the time available to get a result is not critical and, normally, a cluster of computers are used – sometimes this mode is used to generate animation movies. This mode is not considered in this work.

In real time mode, the system must be able to process the input information and generate the animation and speech in "simultaneous".

2.3*Speech Generation and Lip-sync*

In a talking head system, the speech is directly related with the audio that is reproduced together with the facial animation (Edge & Maddock 2001), (Waters & Levergood 1993), (Pelachaud 1991). Basically, there are two approaches that must be considered: (1) the audio can be captured recording the speech of a human being or (2) synthesized. In our system we use synthesized speech – when compared to the first approach it quality is normally inferior, but it versatility is better.

A speech-synthesizer, or text-to-speech synthesizer (*TTS*), is a computer-based system that should be able to read text aloud (Dutoit 1997). Normally two principal units form this system: (1) the Natural Language Processing sub-module (*NLP*) and (2) the Digital Signal Processing sub-module (*DSP*). The NLP is capable of producing a phonetic transcription of the text together with the desired intonation and rhythm (often referred to as prosody). The DSP sub-module transforms the symbolic information it receives from the NLP sub-module into audible sounds – the speech itself.

Actually, instead of simple phonemes, some variations are frequently used, such as diphones and triphones, which helps lip-sync.

3BUILDING A 3D PHOTOREALISTIC HEAD

In our system the first physical structure of a head was a mesh developed in a CAD-like application. In the reality it is only a face, and it is just an adaptation of the Waters' model (Waters 1988), (Waters 1987).

To obtain a photorealistic physical structure we used a 3D laser scanner – the VIVID 910 (*VI-910*) – and the respective software – the Polygon Editing Tool, version 2.10 (*PET*); both from KONICA MINOLTA (Majid et al. 2004). In the next section, the method used to obtain a realist physical structure is explained and some results are presented.

3.1*Scanning a Head*

The physical structures were acquired with the 3D scanner VI-910 using a TELE, f = 25 mm, lens.

A turning chair and some landmarks attached in the background panel were used to help the subject to keep his head fixed during the acquisitions from distinct angles. Its use will decrease the errors introduced by twists or judders of the subject's neck. These movements, even with small amplitude, leads to ghosts or artifacts when blending different acquisitions together.

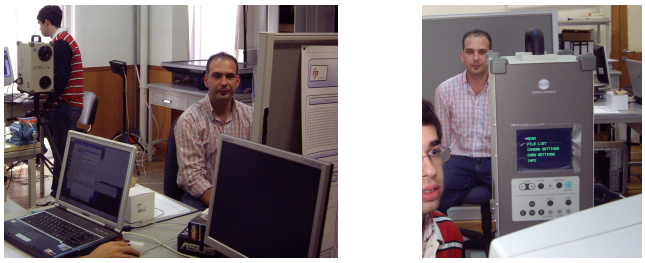


Figure 1. Experimental setup used to capture the 3D model of a subject head.

But even with the help of our setup, subtle movements of facial muscles are unavoidable (just a scan can take more than 10 s). This leads also to ghosts or artifacts.

The Figure 2 shows a 3D model of a mannequin head in different poses and the last image in the right

has texture. This 3D model was reconstruct using four elements (4 scanning), the face, the right and left side taking a part of the back and the top of the head. For this experimental setup we have used only a turning chair (the landmarks in background panel are not necessary) to capture the different poses of the mannequin. The first scanning was the face, the second and third one were the right and left side respectively, taking some areas of the back part of the head. Taking as reference the ears, we approached the face and the sides and we also achieved the back part of the head without to do a scanning of all back of the head. The last scanning was the top of head. All elements were approached and then we did the registration and merging of the elements.

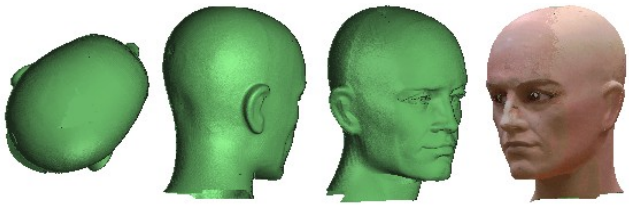
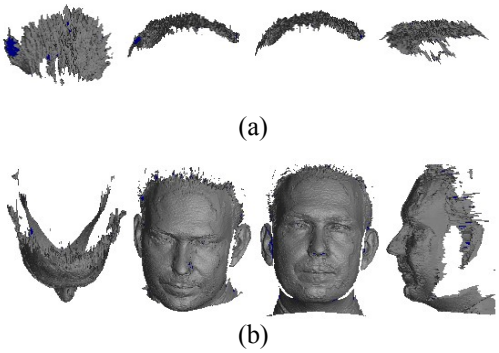


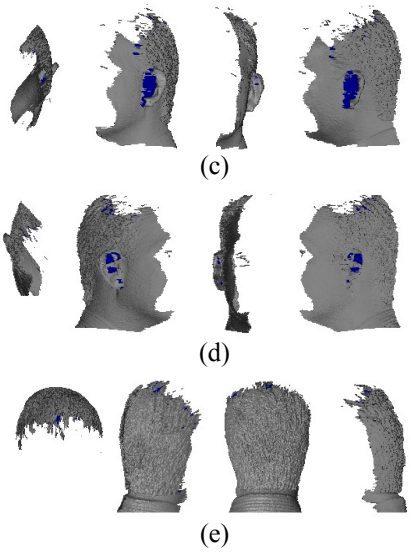
Figure 2. A mannequin head 3D model. We use the VI-910 scanner and the PET software to acquire the model.

The Figure 3 shows 5 elements scanned to reconstruct the 3D model of a human head. The element showed in the Figure 4 (a) is the top of head, (b) is the face, (c) and (d) are the left and right side respectively and (e) the back part of the head, all 5 elements are showed in different poses. Some problems were found during the scanning. A problem found was the movements of the person during the scanner and this causes differences in the position of the head when is done the approaching for the registration (this problem does not happen with the mannequin). Another problem was the hair, the scanner does not capture very well. To solve this problem the person used wet hair with gel and this help us, because the scanner have captured a good part of the hair. Due the hair problem more scanning were necessary.



(a)

(b)



(c)

(d)

(e)

Figure 3. The five scanned elements to acquire the 3D model of a human head: (a) head top; (b) face in different poses; (c) left side; (d) right side and (e) back part of the head.

The Figure 4 shows the 3D model of a human head in different poses, where the 2 images in the right have texture.

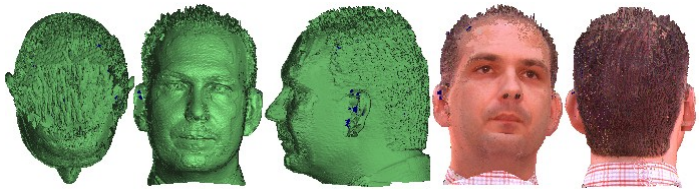


Figure 4. The 3D model of a human head in different poses.

4THE TALKING HEAD ARCHITECTURE

This section presents an overview of our talking head system. It is implemented in C++ using the OpenGL library and has eight modules. They will be described in the following sections. Figure 5 presents a general overview of the system.

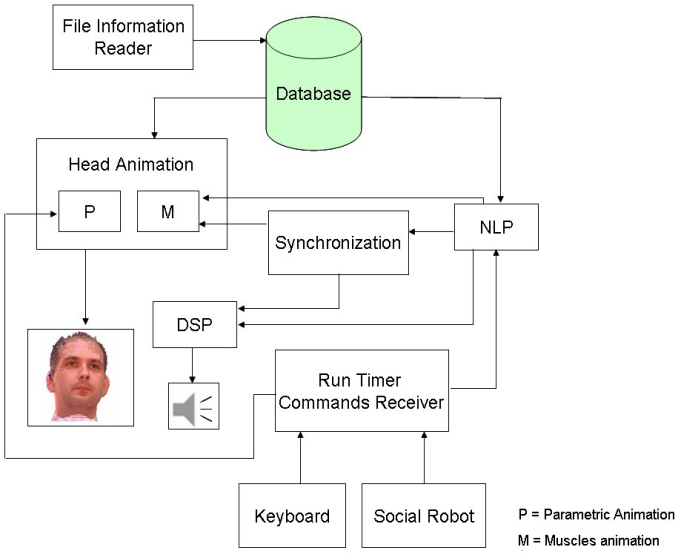


Figure 5. A general overview our talking head system.

4.1*File Information Reader Module*

When the system is initialized, this module is responsible to load some parameters, from a set of specific files: (1) the physical structure of the head (vertices, triangles and textures); (2) the muscles items (localization and the parameters which define their influence region); (3) the information for the parametric animation (centers and axis of rotation); (4) the phrases that will be uttered (in US English, without any other information that an ordinal number) and (5) the gender.

The head physical structure is stored from a frontal viewpoint and with a neutral expression. The eyes are not automatically constructed as in the Waters' (Waters 1988), (Waters 1987) original system – they are also stored in a file.

Still, in the system's initialization, this module loads, also from specific files: (1) the information about the distinct facial expressions (the values of contraction of each muscle); (2) the visemes (values analogous to that of the facial expressions, but only for the mouth region) and (3) the phoneme-viseme mapping.

All that information is loaded in the computer memory to form a database of fast access.

4.2*Head Animation Module*

A hybrid technique is used to create the movements, i.e., we use a parametric animation and muscles. These are used to simulate skin movements, while the parametric animation can be used to simulate the eyes and the head / neck activities.

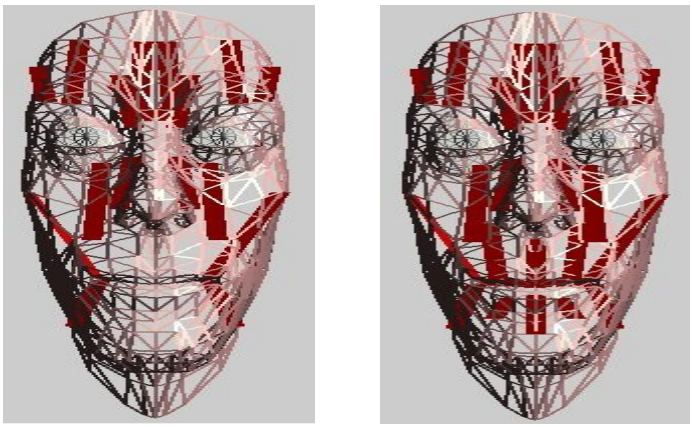
Waters' muscle technique (Parke & Waters 1996), (Waters 1987) was chosen because it is simple and not computationally expensive (e.g., the muscles actions are simulated using cosine functions calculated from vectors dot product). Moreover, the calculations involved in the parametric animation are also undemanding.

Using this hybrid technique approach it is possible to simulate the most important AU's of FACS in real time and with a minimum consumption of resources.

4.3*Muscles Based Animation Module*

Muscles based animation was inspired in the Waters' system (Parke & Waters 1996), (Waters 1987), but some improvements are made.

The muscle set has been chosen in a way to move the more noticeable regions of the face skin. This set is composed by 32 muscles (see Figure 6) which are divided in two main groups.



(a) (b)

Figure 6. Set of muscles: (a) 18 muscles in Waters' system; (b) 32 muscles in our system.

The first group (10 muscles) is responsible for moving the eyebrows, eyelids and the forehead. Eyebrows are especially involved in the generation of the facial expressions.

The second group is related with mouth region movements. It controls the lips and, in a more general way, the area underneath the nose. To allow a precise control of the mouth, this second group is constituted by a great number of muscles (22 muscles).

At the moment, the system has only linear muscles; *sphincter* muscles (like the *Orbicularis Oris*, which handles the protrusion of the lips) is not implemented by now. In the following linear muscles conception is explained.

In Water's model, a muscle is linked to the mesh in two vertices (see Figure 7):

- A – the point of attachment;
- I – the point of insertion;

The vector $\overrightarrow{IA}$ can be considered as the muscle itself. When the muscle suffers a contraction it length is remained unchanged (this is, in contrast with the biological systems). The muscle acts like a magnet attracting all the vertices within its zone of influence, thus the skin moves. This zone is an angular sector defined as follows:

- α – an angle;
- $\overline{IA}$ – a radial distance.

In order to give an illusion of skin elasticity, it must have a smooth transition between the movement of vertices that are in the zone of influence and their neighbors. To give this illusion, is necessary to define:

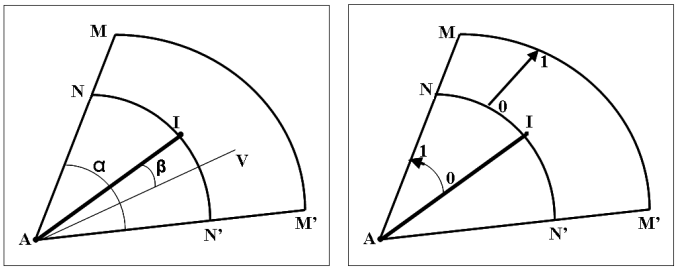
- β – an angle, where $\beta \leq \alpha$;
- S_2 – a sector inside the zone of influence.

Two sectors are defined:

S_1 *defined by vertices* $\lfloor M\ M'\ N'\ N \rfloor$

S_2 *defined by vertices* $\lfloor A\ N\ N' \rfloor$

Inside S_2 sector, vertices movements are faded as the β raises to α. Moreover, inside S_1 sector, vertices movements are faded by the conjunction of two factors: the angular and the radial (see Equations 1-4). This gives the illusion of skin elasticity.



(a) (b)

Figure 7. Water's based muscle model (a) and contraction fading (b).

Equations 1 and 2 explain how to calculate, respectively, the angular and radial factors to fad the vertices movements:

$$\delta_\alpha = \frac{\cos\beta - \cos\alpha}{1 - \cos\alpha} \tag{1}$$

$$\delta_R = \cos\left(\frac{\overline{VN}}{\overline{MN}}\frac{\pi}{2}\right) \tag{2}$$

Equations 3 explain how to calculate the total fad ($\vec{\delta}_v$) in a vertex *V*:

$$\vec{\delta}_v : \begin{cases} \delta_\alpha\ \delta_R\ \vec{IA} & \text{if } v_{old} \text{ inside sector } S_1 \\ \delta_\alpha\ \vec{IA} & \text{if } v_{old} \text{ inside sector } S_2 \end{cases} \tag{3}$$

For a vertex *V*, it new position (v_{new}) is calculated as the sum of it old position (v_{old}) and the total fad:

$$v_{new} = v_{old} + \vec{\delta}_v \tag{4}$$

Then, the maximal displacement is gotten when vertex *V* is located at the insertion point *I*.

4.4*Parametric Animation Module*

This module is responsible for the movements of the head itself, as well as, of some head components. Those movements are independent of the actual expression. Actually, there are four modes: no movement, eyes movement only, head movement only, eyes and head movement.

Eye rotations are triggered by a biological need. To simulate theses behaviors this module incorporates an "automatism" responsible to trigger movements with random values (time and angles). To perform head movements, it has another "automatism" similar to that of the eyes. Eyes movements are simple rotations around, their centers. In simulating neck movements, the head could only be turned over a vertical axis.

The head can also to reproduce the effects of jaw movements, that is, a rotation of the chin.

4.5*Decoder Module*

Actually, the fundamental element of this module is the sub-module NLP, the application eLite. The Decoder module is the responsible for analyzing the input text (the phases to utter) and for providing a data structure which contain the phonemes that compose the speech and also, the respective fundamental frequency and the timing data. With this information, it is possible that other modules generate the sounds and the lips movements in a synchronized way.

4.6*Text-to-Speech Module*

The Text-to-Speech (*TTS*) module is composed by two sub-modules: The NLP and the DSP.

The NLP sub-module integrates also the decoder module (see Section 4.5 for an explanation of it functionality).

The DSP sub-module is fundamentally composed by the application MBROLA (Dutoit 1997). It takes the phonetic description provided by the NLP sub-module and, when receives a command from the synchronization module, sends the synthesized sounds to the loudspeakers.

This TTS synthesizer offers a multilingual platform of acceptable quality, where the language (actually only US English) and the gender can be selected by the user (see Section 4.1).

4.7*Synchronization Module*

This module is responsible for a fundamental activity: the synchronization between speech and muscles actions. It analyzes the information provided by the Decoder module and sends signals to the Animation module and to the DSP sub-module side the TTS module.

4.8*Run Time Commands Receiver*

At run time the Talking Head system can receive some commands to change its behaviors.

These commands can be supplied by a user, through the computer keyboard, to modify, e.g., the monitor luminosity or to select one of the four functioning modes (see Section 4.4).

The commands can also be received from one of ours social robots. The Talking Head and the robots communicate via a TCP / IP connection. The robot sends only a number which corresponds to the phrase to be uttered.

5CONCLUSIONS AND FUTURE WORK

This work provides three main contributions. The first is the presentation of a complete procedure to obtain a photorealistic 3D physical structure of a human being head. The second is the presentation of a hybrid way to perform the animations. Finally we describe our Talking Head system – its principal modules and respective interrelations are explained in some detail.

A complete procedure to construct a photorealistic human head, based in the data provided by a 3D laser scanner is presented. Several involved actions, such as the acquisition procedure (and respective layout) and the tasks required to convert the raw data in a 3D head structure have been described with some detail.

We wish to emphasize that the scanning of a human being is much more complicated than that of an inanimate object, i.e., the mannequin we have used. We will continue to scan human beings to build a dataset.

The animations are executed in a hybrid way: we use principally muscles but, eyes, jaw and neck are undergone to a parametric animation. Using this technique, it is possible to generate the more important AU's of FACS with a moderate use of resources. However, the AU's responsible for tongue movements are missing as the model does not have nor tongue nor teeth. Also, the swelling of the cheeks are not carried out. All those are tasks to develop in the next months.

Finally, the third, and main contribution of this work, is the development of a Talking Head system.

The Viseme - Phoneme Synchronization module is not described in detail because it is not complete. The facial animation with synchronized speech is not realistic (the humans beings, are very sensible and critics to fails in this subject) if the frame rate is too slow. Considering that a phoneme takes about 60ms, the frame rate should not be less than 15 images / s. It is a fundamental objective to finish briefly It is a fundamental objective to finish as soon this module this module.

6ACKNOWLEDGES

We are grateful to Prof. Helder Araújo (University of Coimbra) for allowing the access to the 3D scanner and the space in the Vision Lab for the acquisition of the 3D human head structure. We also would like to thank the students João Pacheco and João Salgado for the assistance in the scanning.

This work is partially supported by FCT-Fundação para a Ciência e Tecnologia Grant #30655/2006 to Diego Faria and by the BACS-project-6th Framework Programme of the European Commission contract number: FP6-IST-027140, Action line: Cognitive Systems.

7REFERENCES

Alexa, Marc; Behr, Johannes; and Müller, Wolfgang 2000. TheMorph Node. *Proceedings of Web3D/VRML*. Monterey, CA, USA.

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

Edge, J. & Maddock, S. 2001. Expressive Visual Speech Using Geometric Muscle Functions. *Proc. Eurographics* UK, pp. 11-18.

Ekman, P. & Friesen, W. 1978. *Facial Action Coding System*. Consulting Psychologists Press, Palo Alto

Majid, Zulkepli; Setan, Halim & Chong, Albert 2004. 3D Modeling of Human Face with Noncontact Three dimensional Digitizer. *International Symposium and Exhibition on Geoinformation 2004*, Kuala Lumpur, pp. 21-23.

Parke, Frederic I. 1972. *Computer Generated Animation of Faces*. Master's thesis, University of Utah, Salt Lake City, UT.

Parke, Frederic I. 1974. *A Parametric Model For Human Faces*. PhD thesis, University of Utah, Salt Lake City, UT.

Parke, Frederic I. & Waters, Keith 1996. *Computer Facial Animation*. England. A K Peters, Ltd.

Pelachaud, C. 1991. *Communication and Coarticulation in Facial Animation.* PhD Thesis, University of Pennsylvania, Philadelphia.

Rydfalk, M.1987. CANDIDE - A parameterized face, Technical Report LiTH-ISY-I-0866, Linköping University, Sweden.

Wang, Carol Leon-Yun; Langwidere 1993. *A Hierarchical Spline Based Facial Animation System with Simulated Muscles*, Ph.D. thesis, University of Calgary.

Waters, Keith 1987. A Muscle Model for Animating Three-Dimensional Facial Expression. P*roceedings of Siggraph* (Anaheim, California).

Waters, K. 1988. *The Computer Synthesis of Expressive Three Dimensional Facial Character Animation*. PhD Thesis, Middlesex Polytechnic

Waters, K. and Levergood, T.M. 1993. DECface: An Automatic Lip-Synchronisation Algorithm for Synthetic Faces, Digital Equipment Corporation

Yuencheng, Lee; Terzopoulos, Demetri & Waters, Keith 1995. Realistic Modeling for Facial Animation. *Proceedings of Siggraph*, Los Angeles, California.