Animation of Lip Motions and Facial Expressions Using 3D Homeomorphic Models

Erdem Akagunduz, Ugur Halici
Computer Vision and Intelligent Systems Research Lab,
Department of Electrical and Electronics Engineering,
Middle East Technical University, Ankara, 06531, Turkey
{erdema, halici} @metu.edu.tr
http://vision1.eee.metu.edu.tr/~vision/

Abstract

The aim of this study is the simulation and synchronization of human facial expressions and lip motion for Turkish text using homeomorphic 3D facial models using Java Programming Language and JAVA3D API. For this purpose Turkish text is mapped to related homeomorphic 3D facial models, these models are synchronously morphed to form the animation and this animation is combined with facial expressions, again by using morphing but based on empowered emotion model. The motivation behind using empowered models is preventing the effectiveness of emotions and lip motion while combining them. The synchronization and timing of the animation depends on the letters that are being processed.

Keywords: 3D facial modeling, homeomorphic models, facial animation, lip motion, lip/speech synchronization, facial expression simulation, empowered emotion model.

ACKNOWLEDGEMENTS - This work is partially supported under the research project BAP-2002-07-04-04

1. Introduction

We humans, communicate using our entire body. We use our face, our voice, our body, and our social states to communicate. But among all of these communication gifts, human face and human voice are the most fundamental ones. Essentially, the face is the part of the body that we use to recognize the individuals; we can recognize a face from vast universe of similar faces and are able to detect very subtle changes in facial expression [9]. These skills are learned early in life, and they rapidly develop into a major channel of communication. Actually that's why animators pay a great deal of attention to the face.

In recent years there has been considerable interest in computer-based three-dimensional facial character animation [9]. These studies go back more than 30 years. However with the rapid growth of hardware and software computer technologies during the recent years, the outputs of these studies became more evident. Facial animation, facial expression animation, lip motion for languages and lip/speech synchronization are some of the important applications.

The difficulty of the modeling of human facial motion is mainly due to the complexity of the physical structure of the human face. Not only are there a great number of specific bones, but there is also interaction between muscles and bones and between the muscles themselves [6]. Human facial expressions have been the subject of much investigation by scientific community. Some milestone studies on creating models representing expressions and lip motion are listed below.

Long ago in 1872, Charles Darwin published "The Expression of the Emotions in Man and Animals, where he dealt precisely with these issues. Actually this was the very start of the studies that led us to today's technology in character animation.

In 1972 Frederic I. Parke began with a very crude polygonal representation of the head, which resulted in a flip-pack animation of the face opening and closing eyes and mouth [9].

In 1975 Paul Ekman stated that humans are highly sensitive to visual messages sent voluntarily or involuntarily by the face. Consequently, facial animation requires specific algorithms able to render the natural characteristics of the motion with a high degree of realism. Research on basic facial animation and modeling has been extensively studied and several models have been proposed [4].

Later, Platt and Badler have designed a model that is based on underlying facial structure. The skin is the outside level, represented by a set of 3D points that define a surface, which can be modified. The bones represent an internal level that cannot be moved. Between these levels, muscles are groups of points with elastic arcs [10].

Waters in 1987 represented the action of muscles using primary motivators on a non-specific deformable topology of the face. Two types of muscle were created: linear/parallel muscles that pull and sphincter muscles that squeeze [11].

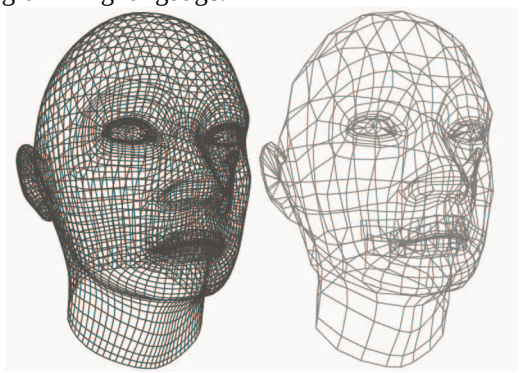
Magnenat-Thalmann et al. defined a model where the action of a muscle is simulated by a procedure, called Abstract Muscle Action procedure (AMA), which acts on the vertices composing the human face figure. It is possible to animate a human face by manipulating the facial parameters using AMA procedures. By combining the facial parameters obtained by the AMA procedures in different ways, we can construct more complex entities corresponding to the well-known concept of facial expression [7].

In 1991 Prem Kalra, Angelo Mangili et al. introduced SMILE: A multilayered facial animation system. They described a methodology for specifying facial animation based on a multi-layered approach. Each successive layer defines entities from a more abstract point of view, starting with phonemes, and working up through words, sentences, expressions, and emotions. Finally, the high level allows the manipulation of these entities, ensuring synchronization of the eye motion with emotions and word flow of a sentence [6].

In 1994, the first version of VRML, virtual reality modeling language was presented. The lack of animation and interaction was leading to VRML 2.0 that became the ISO standard VRML-97. In succeeding years VRML-97 was added to the MPEG-4 standard which is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group), the committee that also developed the Emmy Award winning standards known as MPEG-1 and MPEG-2. Specific extensions were related to the animation of artificial faces and bodies. The facial animation object in MPEG-4 can be used to render an animated face. Face animation in MPEG-4 provides for highly efficient coding of animation parameters that can drive an unlimited range of face models [8].

Among these studies several 3D facial geometry representation techniques were introduced and are still being introduced. Facial polygonal models are basically the 3D facial models constructed using polygonal meshes (Figure 1)

In a project conducted in our research laboratory we have developed a system for 3D animation of Turkish speech, human facial expressions and synchronization with a Turkish Speech engine using JAVA programming language, JAVA3D API and Java Speech API. We developed a 3D animation model that was able to simulate Turkish speech together with visual and audio components. To generate the simulation we have used OpenGL via JAVA3D API classes and interfaces. Also we have used GVZ Speech Technology's Turkish Speech engine GVZ SDK for human voice synthesis. Using these two interfaces we have constructed the animation and synchronization software in Java programming language.



(a) **(b)**

Figure 1. Facial Polygonal Meshes: a) 22984 vertices, 11492 faces. b) 3070 vertices, 1490 faces.

The implementation block system has JSML Turkish text input and it has the output of real-time 3D Turkish speaker animation.

We have written a simple JSML parser code to parse the markup text. For our application we simply needed to use SENT tag and defined our basic six emotion tags, namely happiness, fear, sadness, surprise, anger and disgust. Details of this code can be found in [2] together with details of the simulation software.

In order to model Turkish lip motion we have first defined the Turkish visual phoneme by inspection. We have mapped all the letters of Turkish alphabet to a single or multiple numbers of 3D visual phoneme models (Figure 2.) In this mapping operation we have taken the syllable structure of Turkish language in consideration. For automatic extraction of syllables from a Turkish sentence, we have defined an algorithm.

Finally we have synchronized the animation with synthesized Turkish speech on a word-by-word synchronization basis. The simulation we have constructed can be animated with any suitable set of 3D homeomorphic set of 3D models.

In this paper we explain the approach that we used for animation of lip motions and facial expression in our system. Our approach uses weighted morphing for animation of facial expression and lip motion. However we use empowered emotion models

instead of traditional emotion models in order to overcome the problem of loosing effectiveness of target models when more than one target models are combined.

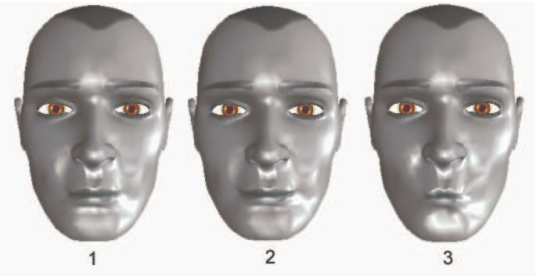


Figure 2 Different Visual Phonemes for the letter "m" in different syllables

2. Weighted Morphing For Animation of Facial Expressions And Lip Motions

Animating facial expression is the most challenging aspect of facial animation. When we animate facial expressions, several factors are taken into consideration, like personalities or motion or weight of the emotion or the kind of the emotion. Actually this is a subject of 3D modelers, cartoon animators, graphical designers. However, we will try to briefly understand the mechanical process behind simulation of 3D facial expressions.

The simulation process depends on the kind of animation you use. You may be using a 3D face with control parameters of virtual muscles. You may be using the method morphing with homeomorphic models [1]. Each method has its own advantages and disadvantages. There may be other known methods but in this section we will examine this two important methods. The first method is named *control parameterization* and it is the most commonly used method. The other method is *morphing*, which is the one we have used in this study.

Control parameterization: In this method the development of facial animation may be viewed as two independent activities: the development of control parameterizations and associated user interfaces, and the development of techniques to implement facial animation based on these parameters [9]. Basically in this method the movement on face is modeled in relation to some

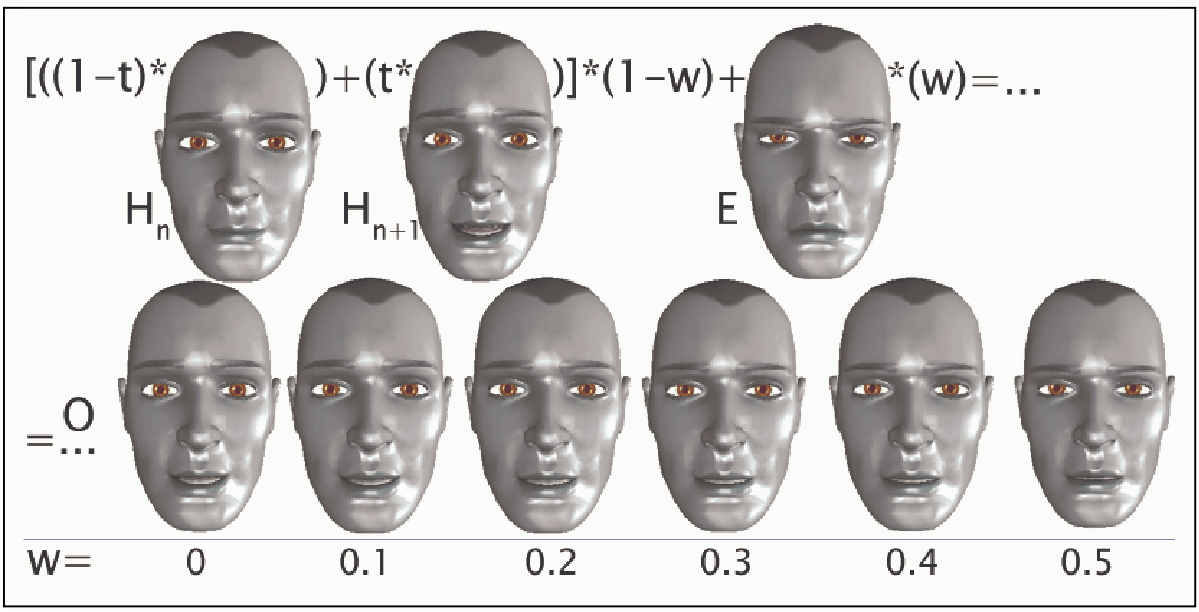


Figure 3-Weighted Morphing based on Traditional Emotion Model: In this figure the simple logic behind weighted morphing is explained. This captures are taken from the simulation of the sentence "Merhaba." (Merhaba in Turkish means Hello). The sentence is in "anger" emotion package. The jsml input is *"<SENT> <ANGRY> Merhaba </ANGRY> </SENT>"* The simulation position is between the first and second letters of the sentence. $\mathbf{H_n}$ denotes the 3D model of visual phoneme "m" $\mathbf{H_{n+1}}$ denotes the 3D model of visual phoneme "e". **E** denotes the 3D traditional emotion model. Finally **O** is the output. At the instant of the capture the parameter *is*: **t = 0.199.** The results are shown for different **w** values.

criteria. These criteria may be the movement of facial muscles, or the elastic movement of the facial skin. The main idea can be described as understanding the motion capabilities of the face and by extracting every independent motion on the face, implementing these parameters on a virtual face model.

As this method depends on the control parameters, only one facial model is kept at memory during software simulation. For this reason we may say that the method is a *memory-friendly* implementation. On this model the desired animation is achieved by the controlling the parameters. But when it comes to the processor performance, it is not the same. Every frame requires parameter calculation, which requires extra CPU usage. This method is widely used for realistic and artistic animations. Today's Hollywood movies use this method to animate their computer-generated characters. Needless to say that real-time rendering is avoided in this method. There are real-time rendered examples of this implementation method, but the reality and artistic view of the animation is highly reduced in those examples.

Morphing: Animation of three-dimensional shapes involves the change of vertex attributes over time. Morphing represents such changes over time as the interpolation of two given shapes. We can extend this concept to more than two base shape and use morphing to produce blends of several objects. The interpolation function can be summarized as follows:

Let **I** be a geometry array formed of the vertices $\mathbf{I_i}$, the orthogonal coordinates being represented as $\mathbf{I_{ix}}$, $\mathbf{I_{iy}}$ and $\mathbf{I_{iz}}$.
Let **T** be the target object, **I** the initial object, the **O** the output object.
Let $\boldsymbol{\alpha}$ be the morphing weight, where $0 \le \alpha \le 1.0$

$$O_i = \alpha \cdot I_i + (1-\alpha) \cdot T_i \qquad (1)$$

For this method we have to find the suitable set of necessary shapes for the animation. These suitable sets depend on the kind of animation. If the animation is speech animation of a certain language, the set is chosen in accordance with the visual phonemes of that language.

To morph various targets successively to form a complete animation, the method '*key-framing*' is used. Key-framing is the method of interpolating some key models successively due to some certain time.

The use of key poses and interpolation is among the earliest and still the most widely used schemes for implementing and controlling facial animation. The basic idea and the control parameterization for interpolation are very simple and also very limited.

This method requires extreme usage computer memory. Because for a high performance real-time animation, the entire key poses must be kept in computer memory. However the processor usage is lower in comparison to control parameterization. Except for the rendering calculations, which are mainly calculated in video cards processors, the only calculation is the simple interpolation function.

Weighted morphing is the ability to morph a base or anchor model into two or more target models simultaneously [5]. This is a major advantage when you are creating lip synch animation that includes both dialog and facial expressions.

The idea behind weighted morphing is very similar to that of single morphing. The only difference is the target model is not a single model but a weighted sum of some number of different models. The weighted morphing function can be described as follows:

Let $\boldsymbol{I}$ *be a geometry array formed of the vertices* $\boldsymbol{I_i}$*. The orthogonal coordinates are represented as* $\boldsymbol{I_{ix}}$, $\boldsymbol{I_{iy}}$ *and* $\boldsymbol{I_{iz}}$.
Let $\boldsymbol{T_n}$*, n=1,2…N, be the target objects,* $\boldsymbol{I}$ *the initial object, the* $\boldsymbol{O}$ *the output object.*
Let $\boldsymbol{\alpha_n}$ *be the morphing weights, where*

$$O_i = (1-\alpha) \cdot I_i + \sum_n (\alpha_n \cdot T_{n,i}) \qquad (2)$$

where

$$0 \le \alpha = \sum_n \alpha_n \le 1.0$$

An example of this weighted morphing is seen in Figure 3. The main problem of weighted morphing is as the number of morph targets increase, weight of each target decrease and the affect of these morph events become less evident in the final result. Let's give an example of morphing animation of speech with one single emotion. The speech requires lip motion that is morphing of visual phonemes to each other in time. But if you require an angry lip motion in speech you need to morph the whole animation with an emotion model corresponding to anger. Let's assume that the total lip motion sequence is weighted by the real number *w*, where 0<w<1. Then naturally the emotion model will be weighted with *1-w*. If w is very close to 1 then the emotion will be very insignificant. On the other hand as we decrease w, the weight of emotion increases but the weight of lip motion sequence decreases and the animation quality of the lip motion decreases too. In other words the animation loses its understandability. To

better understand this phenomenon visually, Figure 3 explains the affect of using different emotion weight w values, for traditional emotion models. To avoid loosing understandability, segmented morphing, which morphs separate areas of the face individually, can be used. In this way the morphing of the brows does not affect the morphing weight around the lips. Segmented morphing has two advantages over straight morphing; the ability to build a large number of expressions from a smaller set of targets and the ability to animate changing expressions while the character is talking. However, it is somehow difficult to segment the model.

3. Empowered Emotion Model

In our system, instead of segmented morphing, we used a method, which is based on empowering emotion key-frame poses. In this method the model is not segmented. The interpolation algorithm works for the whole model. But in this method instead of using traditional key-frame poses for the emotion models we use empowered models.

To better understand this method let us return back to our question. The main problem was that as the emotion weighting increases the speech lip motion becomes insignificant (For example when w=0.2). In this method the lip motion weight is detained at a considerable value like %80 or more. The remaining %20 or lesser weight is used for the emotion models. Unfortunately %20 is a low weight with which the emotion of face will be insignificant. Instead of using traditional emotion models we use empowered emotion models. A traditional emotion model has the 3D shape of the face having an emotion as illustrated in Figure 4.b. But empowered emotion model is somehow exaggerated version of this model, Figure 4.c. To obtain this model we make a simple difference calculation. To achieve an empowered emotion model we need the neutral facial model. Neutral facial model is the one with no lip motion, no emotions and no expressions as depicted in Figure4.a.

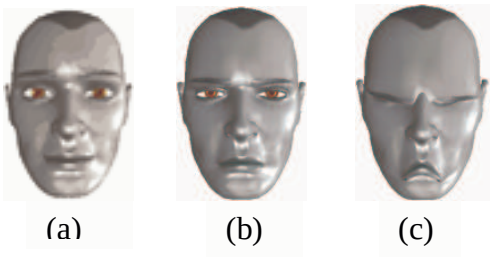


Figure 4 a) Neutral Face Model, N **b)** Emotion Model for "anger", E **c)** Empowered Emotion Model P for "anger"

The method to acquire this empowered model can be briefly described as:

Let **N** be the neutral model and $\mathbf{N_i}$ be the $\mathbf{i^{th}}$ vertex

Let **E** be the emotion model and $\mathbf{E_i}$ be the $\mathbf{i^{th}}$ vertex

Let ρ be the empowering constant

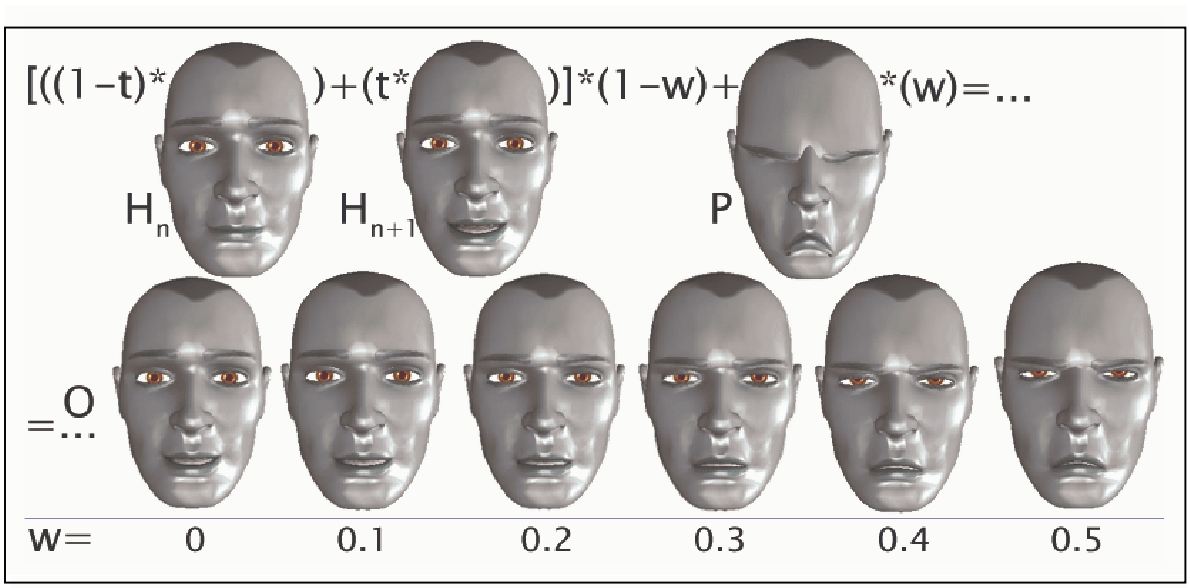


Figure 5-Weighted Morphing based on Empowered Emotion Model: Here $\mathbf{H_n}$ and $\mathbf{H_{n+1}}$ and all other parameters are the same as given in Figure 3. However empowered emotion model **P** is used instead of traditional model **E**.

Then Empowered model vertex $\mathbf{P_i}$ can be calculated as

$$P_i = N_i + \rho \cdot (E_i - N_i) \quad (3)$$

where $\rho > 1$

When these models are used in the animation as the empowered vertices affect the animation even with a small weight value. And the rest of the simulation, the lip motion, is still weighted with a high value and the lip motion is still understandable. In our system we have chosen the empowering constant ρ as 3. The resulting outputs based on the empowered model are given in Figure 5. From the results it can be easily seen that by using empowered emotion model better emotion features are generated even with small w values.

4. Conclusions And Future Studies

In our system lip motion is synchronized and morphed with emotions and synchronized with speech. Since the input was text and the output is lip motion with emotions and synthesized speech we used a markup language, which should handle emotions, delay and similar natural speech event during simulation. We have constructed the whole simulation in java programming language. In order to build the 3D virtual world we have used JAVA3D API. In addition to 3D lip motion simulation we have synchronized this animation with synthesized speech. We have used Turkish speech synthesis engine of GVZ Company. As an interface between this speech engine and our simulation program, we have used CloudGarden Java Speech API [3]. Since our entire application is written in JAVA programming language, we have chosen JAVA Speech Markup Language (JSML) to use in our application.

Through out the animations we have achieved a frame rate over than 40 frames/seconds. We have constructed an implementation software with a fixed size memory requirement of 267 MB. We have used synthetic 3D homeomorphic models with 20784 vertices.

In our approach we used empowered emotion models, which provided us combining emotions into animation together with lip motion without loosing the effectiveness of emotions in the expression.

Currently we are working on extracting 3D homeomorphic model sets of real people using multiple camera images, so that we will be able to apply our empowered model on 3D models of real people.

References

[1] E. Akagündüz and U. Halici. Simulation and Synchronization of Human Facial Expressions and Lip Motion for Turkish Syllables. *In Proc. TAINN*, 2003

[2] E. Akagunduz. *Simulation and Synchronization of Turkish Lip Motion and Facial Expressions in a 3D Environment with a Turkish Speech Engine*. M. Sc.Thesis, Middle East Technical University, January. 2004

[3] CloudGarden JSAPI. `www.cloudgarden.com`

[4] P. Ekman and W. V.Friesen. *Unmasking The Face – A guide to recognizing emotions from facial clues.* Malor Books, 2003

[5] B. Fleming and D. Dobbs. *Animating Facial Features And Expressions.* Charles River Media, 1999

[6] P. Kalra, A. Mangili, N. Magnenat-Thalmann, D.Thalmann, SMILE: A Multilayered Facial Animation System, Proc. *IFIP Conference on Modeling in Computer Graphics*, Springer, Tokyo, Japan, 1991

[7] N. Magnetat-Thalman, E. Primeau, D. Thalmann. Abstract Muscle Action Procedures for Human Face Animation, *The Visual Computer*, Vol.3, No.5, 1988

[8] International Standardization Organization (ISO) IO JTCI SC29/WG11. MPEG-4, http://www.mpeg-4.com/, 1998.

[9] F. I. Parke and K. Waters. Computer Facial Animation, A K Peters, 1996

[10] J Platt, and N. Badler. Animating Facial Expressions. In Proc. SIGGRAPH, 1981

[11] K. Waters. A Muscle Model for animating Three- Dimensional Facial Expression. In Proc SIGGRAPH, 1987.