# Face Expressions Animation in e-Learning

Peter Drahoš
Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava, Slovakia
drahos@fiit.stuba.sk

Martin Šperka
Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava, Slovakia
sperka@fiit.stuba.sk

## ABSTRACT

Learning is collaborative process where teacher/ tutor/ instructor and student work together. Loosing face to face audiovisual communication in electronic and distant learning can be replaced by a video conference. Another possibility is to use embodied conversational agents with the ability to synthesize human speech and simulate facial expressions (talking head). This paper summarizes the state of the art in modeling and animation of human face expressions and shows some results achieved in this field at the Faculty of Informatics and Information Technologies, STU in Bratislava.

## Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Graphical user interfaces.*

## General Terms

Human Factors

## Keywords

Facial expression animation, speech synthesis, avatars.

## 1. INTRODUCTION

Simulating talking persons is a long time goal in the telematic applications [1], [2]. Modeling autonomous agents, which are able to communicate in a natural language, is a long goal of computer scientists.

One of the first programs was Eliza [3] by Joseph Weizenbaum, which was the first chatbot ever created and was based on textual communication. Now there are several systems of this type, but with a more intelligent behavior and with visual representation of virtual talking person. This type of agent belongs to the class of intelligent avatars called invatars. There already are many projects and existing systems with embedded conversational avatars or invatars. Most of these systems use simple 2D or 3D face models.

The goal of our research is not to create intelligent avatars, but we aim to provide a high quality visualization library for facial animation of human heads. Such library or media object could be beneficial to any application in need of facial expression visualization.

Simulating face-to-face communication via Avatars is one of the tasks in building virtual communities in e-commerce, games as well as in e-learning in an asynchronous communication. Another motivation is synchronous communication where Avatars represent real anonymous speakers that communicate via a videoconference. Such method of communication would result in significant reduction of transmitted data, but requires fast simulation of human body and face on one side of the communication channel as well as the tracking and recognition of face and body of speaking person on the opposite side.

In human talking analysis and synthesis there are many partial tasks, which have to be solved simultaneously. Among them we concentrated on talking human head tracking, which is the first step towards mimic recognition. Other tasks are sound synthesis and synchronized lip movement, muscles, eyebrows, head movement and tilt, eye blinking etc. Additionally these tasks depend on geometric model representation, facial structure and shape and available hardware.

**Figure 1. Web based chatbot Halo[4].**

The realistic visualization of the talking person is now the primary topic of our research. We experimented with several approaches of face simulation. First we implemented simple talking face with text to speech conversion and voice synthesis from the written Slovak text [5] based on phonemes. While the problem of speech synthesis is relatively well solved the problem of more fluent and natural speaking including the prosody is still the topic of research. In our experiments we tried various approaches including 3D model with variable geometry and static textures [6], static geometry with animated textures created by joining video sequences into one audiovisual stream.

Currently our efforts are concentrating on hardware accelerated real time animation techniques. For model deformations we use a modified Waters [7] linear muscle model.

# 2. FACE SIMULATION - OVERVIEW

## 2.1 Applications

There are advantages over telepresence with videoconferencing systems since facial animation only needs to transfer expression animation parameters instead of full audiovisual streams. Additionally the visual quality depends only on the quality of the model used by the receiver. For example MPEG 4 video standard already includes facial animation [2].

Human face simulation is used in facial surgery planning, reconstruction and identification of faces from skeletal remains in criminology, archeology and anthropology or realistic animations for movies. Other examples are real time communication and telematics applications such as interactive virtual instructor or city, shopping center or museum guide.

Language-training applications using animated conversational agent is another example of the needed applications, especially in the case when there is a lack of foreign language teachers. Even though the auditory modality is dominant in speech perception, it has been shown that the visual modality increases speech intelligibility, especially in noisy environment [8].

Other applications are speech therapy (one of the goals in speech synthesis is to have a talking head that articulates as clearly as or even clearer than human talkers), real-time control of user-interface agents and avatars in 3D chat systems and for accessibility purposes.

Modern web presentations can already use existing chatbots to provide additional information or gather user feedback from vsitors. One such example is chatbot Halo[4] which is capable of visual speech feedback, ambient movement and limited expression visualization. These chatbots are limited by the technology available in web presentations and cannot believably visualize human expressions.

## 2.2 Face anatomy and geometry

The skull, particularly frontal, nasal, zygmatic and mandible bones determine the overall proportions of the head and face. There are 40 muscles with various shapes, which intermediate between the bones and skin. More facial characteristics are determined by individual appearance of eyebrows, eyelids, eyeballs, lower-nose, lips and neck.

The animation of synthetic facial tissue, which is subject to force-based deformations through the action of contractible muscles such as wrinkling and furrowing, is desirable because people can be very sensitive to subtle skin deformations when interpreting facial expressions.

The main physically based component of the face model is layered tissue, including epidermis, the dermis and muscle layer. Dermal tissue is relatively deformable under low stress, with the increasing stress becomes stretch resistant. The first physical computer models used tetrahedral elements to simulate epidermis and hexahedral elements for two lower layers. The edges of polyhedral represented springs with linear approximated non-linear strain stiffness.

Other phenomena like bulging and dimpling of skin, caused by incompressibility of cutaneous ground substance and subcutaneous fatty tissue were modeled.

A physically-based (bio-mechanical) dynamic model of a face which considers mass, stiffness and damping computations, enhances level of realism compared to purely geometric schemes, but is much more computationally expensive than purely geometric models. Such models, which often require volumetric model representation, are not suitable for real-time and on-line applications.

From a geometric modeling point-of-view, the most popular model representation is the polygonal mesh, because the hardware rendering accelerators and Application Program Interfaces libraries support triangle and polygonal meshes directly. Polygonal mesh can be generated manually or automatically from set of points. The cloud of points can be received from 3D digitizer, scanner, structured light or other digitizing techniques based on several photographs or video frames. Adaptive (non-regular) mesh techniques can reduce the amount of data or to adjust the model complexity dynamically.

There exist several data banks where geometry models of human body parts can be obtained.

# 3. EXPRESSION MODELING

Facial expressions contribute to the observer comprehension of the formal and emotional content of the speech. Generation of continuous and realistic transitions between different facial expressions is one of the main problems in synthesizing visually acceptable digital avatars.

Expression is a group of facial parameter values that together transform the neutral face into an expressive face. Expressions simulation and animation can be divided into two groups.

The first one are animated or recorded visemes (and corresponding phonemes), which act on the region of the lips and the mouth and form segments of words. The second one - emotions act on any part of the face.

Contrary to other regions of human face, the lips are mainly characterized by their contours. Several models of lips were created. One of them uses algebraic equations for contours approximation.

From a multidimensional analysis of a real speaker’s gestures, visemes and corresponding phonemes can be extracted.

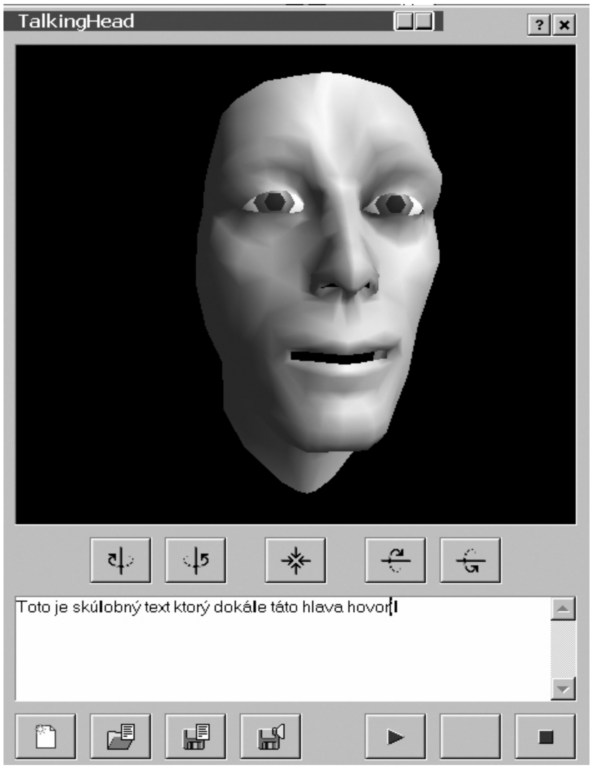
After lips, the most visible moving part of the face by speaking is jaw. Jaw kinematics can be modeled manually or automatically from the data recorded by video, mechanical or opto-electronic sensors. In realistic image synthesis, superpositions of jaw transformations and lips shapes must be integrated with emotions.

Primary emotions are for example surprise, fear, anger, happiness, sadness. Basic expressions and their variants can be defined using snapshots of the computer or real model of the actor.

## 4. ANIMATION

One of the most effective approaches (for example in entertainment applications) to date has been the use of 2D morphing between two static (photographic) images.

Another method that we experimented with is joining video sequences of visemes into one video stream, accompanied by sound [5]. Disadvantage of this approach is this method is not general and requires recording and extracting all possible visemes for the individual person. This method is satisfactory for applications with limited number of used words or sentences. In comparison with using static texture with dynamic model, only few (at least one) photographs of the talking person are necessary.



Toto je skúlobný text ktorý dokále táto hlava hovorl

**Figure 2. User interface of the program for audiovisual text to speech synthesis according to [8].**

On the other hand, 3D animation is more general and flexible, because of the head/face can be viewed from any angle and placed in a 3D virtual environment, making it much more flexible than conventional video.

The combination of 2D and 3D methods is very effective in producing more realistic looking animation with less computation compared to the exact 3D simulation.

A number of approaches have been developed to model and animate realistic facial expressions in three dimensions.

The simplest method in 3D simulation is to use static 3D geometry model with dynamic - texture sequences which is similar to the 2D animation.

Prior works used keyframing and geometrical parametric models, for example morphing between corresponding face models.

Morphing between polygonal meshes is a difficult one, since it requires a set of correspondences between meshes. This task is easier when all face meshes are identical. Together with the geometric interpolation, we need to blend the associated textures, which require pair-wise correspondence between image features.

Similarly to the anatomical face, which is shaped by muscle contractions – virtual muscles can be used to animate the computer model directly.

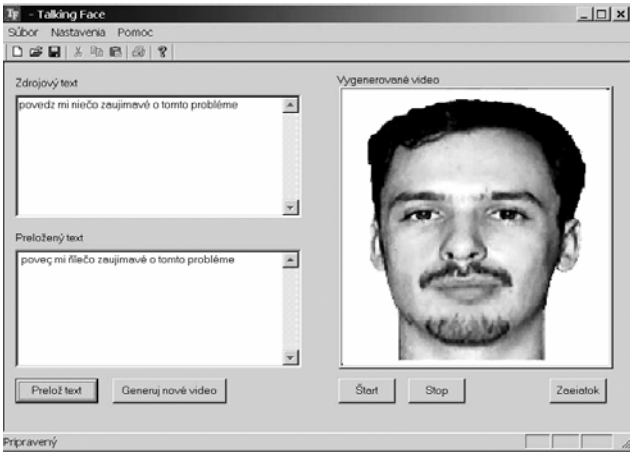
Facial animation control model can be taught from computer vision of real facial behavior, incorporating vocal and facial dynamics such as co-articulation (the interaction between nearby speech segments). Facial muscle contractions and inputs can be analyzed from video sequence of a subject performing expressive articulations. From these we can estimate animation control parameters.

The important and difficult task is synchronization between expression animation and sound. It is necessary to control start time and duration of expressions. These events, which can run simultaneously, are phonemes, words, sentences and their corresponding minimal perceptible actions, visemes, face expressions, emotions.

In animation, flow of events can be described by some formal methods, for example scripts, speech and emotion description languages. In performance driven animations and applications like videoconferencing, the synchronization of audio and video is controlled by the sound and image recognition procedures.

The multilayer approach is extensible and independence of each layer allows the modification of elements without the impact on the others. The animation is based on the interpolation between expressions.

A phoneme snapshot is a particular position of the mouth during a sound emission and can be described by an expression. Expression snapshot is a particular appearance of the face at a given time.



Zdrojový text

povedz mi niečo zaujimavé o tomto probléme

Preložený text

poveç mi ňiečo zaujimavé o tomto probléme

Vygenerované video

**Figure 3. User interface of a program for audiovisual text to speech synthesis [5].**

## 4.1 Real time animation

The creation of facial animation control parameters is a tedious process. Many parameters have to be specified and coordinated (synchronized) to create realistically looking animation sequences.

Bad timing can cause wrong interpretation of expressions (for example slow playback can cause that instead of surprise the expression look like yawn). The automation of this process can be solved by the performance driven (controlled) animation.

This type of animation uses correspondence between video record or real time capturing of an actor, and between the model, using intrusive schemes; for example by tracking markers placed on the actor's face.

Example is Facesnatcher technology, developed by Eyetronics[10], which allows create dynamic models at the video frame rate. The method is based on reconstruction of face from two camera images with structured light projected on the real face from video projector (no markers are involved).

Other methods are capable of automatic real-time track of movements from video sequences and translate them into animation control parameters for the synthesis of facial expressions. This process requires analysis, recognition and interpretation of visually distinguishable facial movements in real time.

The research in autonomous avatars includes intonation, verbal behavior, answering of spoken questions, responding to non-verbal behavior, hand gestures-based on a grammar and communicative context, synthesis of hand gestures based on an understanding of pragmatic information, planning of mixed-initiative dialog including conversational storytelling, etc. Simulation of activities, such as breathing and eye blinks must be automated during the session without the user's intervention.

The ultimate goal in realistic animation is to model in real time human facial anatomy and movement so that it will satisfy structural, functional, as well as audiovisual reality.

A parallel to Alan Turing test, originally proposed for the recognition natural intelligence from the artificial one, can be used to validate results.

## 5. EXPERIMANTAL TALKING FACE

Summarizing the above-mentioned methods we can divide 3D face animation (with polygonal surface models) into several groups:

- Static geometry, dynamic texture.
- Dynamic geometry, no texture.
- Dynamic geometry, static texture.
- Dynamic geometry, dynamic texture.

Considering the rendering and computational capabilities of modern PC hardware real time animation with hardware accelerated geometry and dynamic texturing is possible.

Our current model consists of a polygonal model, set of linear muscles, independent eyes and a jaw manipulator. Dynamic geometry deformations are based on virtual muscles, which are derived form the Waters linear muscle model [6].

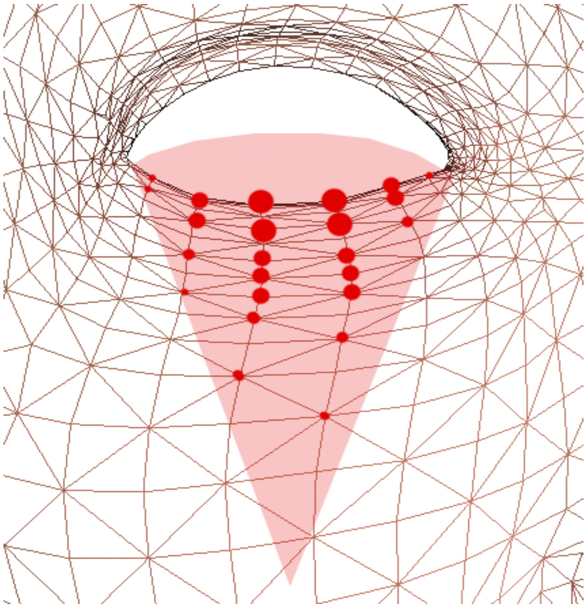


**Figure 4. Illustration of a linear muscle influencing a polygonal mesh. Red dot sizes indicate deformation intensity.**

Linear muscle is a very simple polygonal mesh deformator, which transforms all model vertices in its influence according to a simple mathematical formula. The amount of deformation is determined by a contraction deformation parameter k.

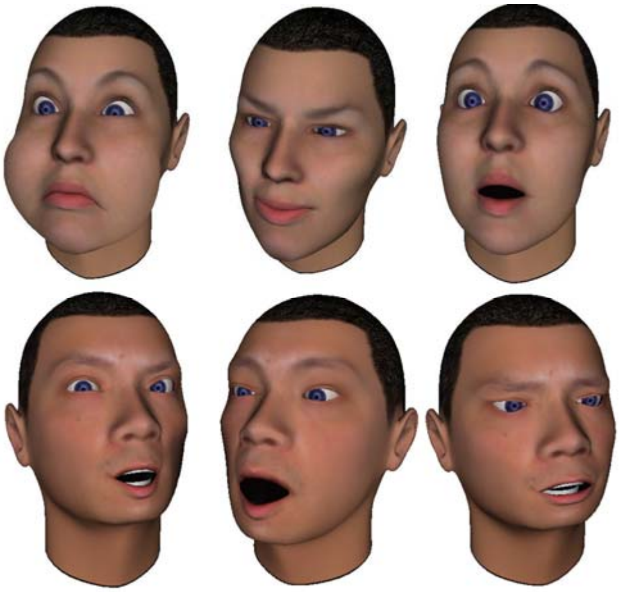


**Figure 5. Example of generated expressions.**

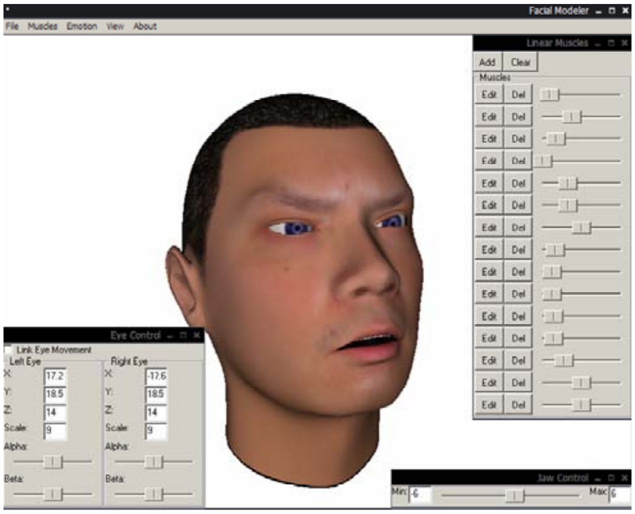
To get around overlapping and performance issues we use accumulative displacements that are applied to the polygonal model just before rendering and static muscle influence calculations [11].

Using this very simple type of muscle, we are able to implement all geometric deformations using modern graphic hardware when available. Additional elements such as eyes are modeled independently as a set of textured spheres with a set diameter, position and normal vector, which defines orientation.

The reason behind using linear muscles is that they provide an abstraction layer between the polygonal model and expressions. Animation is achieved by slightly modifying the set of k contractions between animation frames. For example visemes can be represented using multiple sets of k values and animation can be achieved by simple linear interpolation between these sets.

This approach allows us to animate models of any polygonal structure (not only human faces) using any number and types of virtual muscles. The muscle abstraction layer also enables us to interchange models for simpler or more complex ones depending on available hardware.

We implemented this animation method in a simple reusable library which is capable of real-time rendering of models with more than 20 muscles and more than 20 000 polygons on common PCs without using powerful graphics hardware.



Left Eye: X: 17.2, Y: 18.5, Z: 14, Scale: 9, Alpha:, Beta:
Right Eye: X: -17.6, Y: 18.5, Z: 14, Scale: 9, Alpha:, Beta:
Min: -6 Max: 6

**Figure 6. Model creator user interface.**

## 6. CONCLUSIONS

All projects described in this paper are just the first steps towards creating of multimodal avatar and interpreter, as mentioned in the introduction. Currently the integration of audio and visual part of the talking face is in progress.

The model we developed is capable of complex facial expression animation and visualization on common household computers. For more realistic results the model needs to be enhanced by tongue, hair and realistic texturing and lighting.

Final model complexity and performance depends on the available 3D acceleration hardware. More complex models and animations could be achieved by using more powerful 3D hardware without modifications to the animation software. Many available accelerators could also enhance the quality of the model skin by adding animated skin effects tied to muscle contractions.

In future the integration of facial animation with body and hand gestures is planed with the final goal of creating Avatars with ability of verbal and non-verbal communication.

Full-featured pedagogical agents will require collaboration of many researchers such as linguists, psychologists and graphic specialists. Such effort could result in a reusable multimedia object for facial animation[12] which could be used by existing applications similarly to using any other media object.

## 7. ACKNOWLEDGMENTS

We would like to thank master degree students Vladimir Bohmer who implemented sound synthesis module, Peter Flimmel who implemented Slovak speech synthesis module based on phonemes and diphones and Branislav Pariza who implemented video sequence animation system.

This work is supported by the Slovak Cultural and Education Agency (KEGA) grant 3/3206/5 Interdisciplinary Study Program: Design of Interactive Digital Media and (VEGA) 1/3103/06: Information infrastructure for knowledge mining in distributed environments.

## 8. REFERENCES

[1] Terzopulos, D. Waters, K.: *Techniques for Realistic Facial Modeling and Animation.* In: Thalmann, N., Thalmann D. (eds.): Computer Animtion´91, Springer-Verlag, Tokyo 1991, pp. 59-74

[2] Ostermann, J.: *Animation of synthetic faces in mpeg-4*, Computer Animation, pp. 49-51, 1998.

[3] Weizenbaum, J.: *ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine*. Communications of the Association for Computing Machinery 9 (1966): 36-45.

[4] Advanced Chatbot Solutions: http://www.daden.co.uk/chatbots/. Last visited on 20.04.2006

[5] Flimmel, P.: *Program for the talking face speech synthesis.* Master thesis. Faculty of informatics and information technologies STU. Bratislava 2004.

[6] Pariza, B.: *Program for the simulation of the talking face with video sequences.* Master thesis, Faculty of Informatics and Information Technology, STU, Bratislava 2004.

[7] Waters, L.: *A muscle model for animating three-dimensional facial expressions*, Computer Graphics SIGGRAPH'87, vol. 21, pp. 17–24, 1987.

[8] Guiard-Marigni, T., Adoudani, A., Benoit Ch.: *3D Models of the Lips and Jaw for Visual Speech Synthesis.* Handout given at the IST Conference and exhibition, Vienna, November 1998.

[9] Regec, K., Boehmer, V., Sperka, M.: *Virtual face model and visual speech perception.* Proceedings of the Spring Conference on Computer Graphics (posters). Comenius University, Bratislava, 2001. Pp. 40-41.

[10] Eytronics: http://www.eyetronics.com. Last visited on 10.11.2005

[11] Drahos, P.: *Human Face Animation*. Master thesis. Faculty of Informatics and Information Technologies, STU Bratislava 2005.

[12] Arya., A, Dipaola, S.: *Face as a multimedia object*. In Proceedings of WIAMIS, Lisbon, Portugal. (2004), pp. 21–23.