Facial Animation Framework for Web and Mobile Platforms

Engin Mendi
Computer Science Department
University of Arkansas at Little Rock
Little Rock, AR, 72204
esmendi@ualr.edu

Coskun Bayrak
Computer Science Department
University of Arkansas at Little Rock
Little Rock, AR, 72204
cxbayrak@ualr.edu

***Abstract*—In this paper, we present a realistic facial animation framework for web and mobile platforms. The proposed system converts the text into 3D face animation with synthetic voice, ensuring synchronization of the head and eye movements with emotions and word flow of a sentence. The expression tags embedded in the input sentences turn into given emotion on the face while the virtual face is speaking. The final face motion is obtained by interpolating the keyframes over time to generate transitions between facial expressions. Visual results of the animation are sufficient for web and mobile environments. The proposed system may contribute to the development of various new generation e-Health applications such as intelligent communication systems, human-machine interfaces and interfaces for handicapped people.**

I. Introduction

Audiovisual interaction is an important design factor for human-computer communication systems. Visual channel in the speech communication may significantly improve the speech intelligibility and speech perception by hearers. This makes the face the most effective tool of human communication. In recent years, 3D face animation systems have become a popular subject in various fields including video games, human computer interaction and virtual reality. These systems can also serve as assistive tools for children with learning disabilities such as dyslexia, auditory/visual processing or nonverbal learning disorders [1], [2]. Building an animatable, moderately sophisticated human face can help such children in improving their reading, writing or listening skills.

The emotion of the speech is not only affected by the words it uses, but also by the way the speech is said [3], [4]. The vocal nonverbal component is more efficient than the verbal content for communicating information about the speaker's state or attitude [5]. To realize a more efficient and pleasant human-computer communication, vocal cues should be included in the synthetic speech, especially for robots in social situations [3], [4].

The following specific features of speech may contribute to convey emotional information [3], [4], [6]:

- Pitch and duration play an important role in speech emotion. In particular, the interaction of pitch with loudness and with the grammatical features of the text seems to be critical. In some conditions, pitch and duration are sufficient to distinguish between neutral speech, joy, boredom, anger, sadness, fear and indignation.
- Loudness alone may not be important but the correct synthesis of loudness can help to deliver emotional information.
- Spectral energy distribution and spectral structure can carry much of the affective information.
- Voice quality is also significant in showing the affective information.

In this paper, we present a facial animation framework for web and mobile platforms. The system converts any source of given text to the facial animation talking the text. Our system generates lip movements with emotional expressions corresponding to speech. The text-to-speech engine converts the input sentence into phonemes. The phonemes are then used to create a speech wave. The phonemes are mapped into visemes and sent to the face model to realize lip movements. As the movements finish, the next text input is processed in the same way. The proposed system may contribute to the development of various e-Health applications such as intelligent human-machine interfaces. The system can also serve as an assistive tool for children with learning disabilities that hinders their ability to read, write, listen or speak despite having normal intelligence. The paper is arranged as follows: Section 2 provides brief overview of the keyframe based and MPEG-4 animation. Section 3 describes the components of the system. Section 4 shows our results. Section 5 concludes the paper.

II. Facial Animation

A. Keyframe based Animation

In keyframe based animation, the face motion is obtained by interpolating the key frames for different emotions and visemes (mouth shape) over time to obtain the face shapes between keyframes. A keyframe is a deformed version of a face shape. Each viseme corresponds to a phoneme which is the smallest part of a spoken word. Phonemes are dependents on the spoken language. English has 40 different phonemes [7]. Mapping the phoneme sequence with the visemes, visemes are located on the starting utterance frames for each phoneme. Although most phonemes correspond to a single keyframe,

978-1-61284-696-5/11/$26.00 ©2011 IEEE

some require a linear combination of two or more keyframes [8]. Interpolation of keyframes using a function (linear or cubic) produces the smooth final animation.

B. MPEG-4 Animation

MPEG-4 is an ISO standard developed by MPEG (Moving Picture Experts Group) in 1999. [9], [10] The standard defines numerous tools for representing rich multimedia content. According to MPEG-4 facial animation specification, 84 feature points (FP) are specified on human face. FAPs are used for defining animation parameters as well as animating faces of different sizes and proportions. Figure 1 shows the set of FPs.

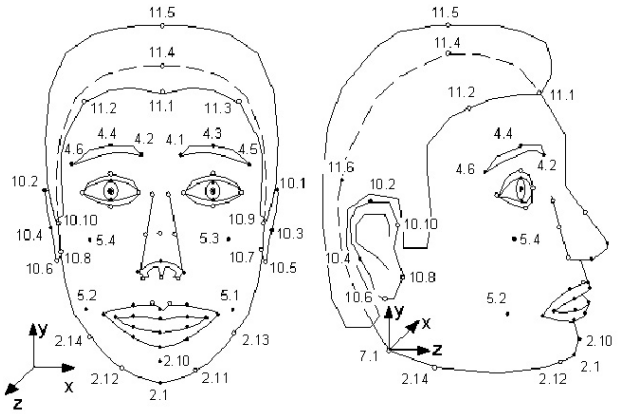

Fig. 1: MPEG-4 feature points [10].

The facial animation is controlled by 68 Facial Animation Parameters (FAPs) driving the animation on the FPs.

III. System Description

In our system we use keyframe based animation. This approach is less CPU intensive [7] and the visual results of this animation are sufficient for web and mobile platforms. An overview of the proposed system is depicted in Figure 2.

First, an input text that can be annotated with emotion tags controlling the 3D face model is converted into phoneme sequence and speech signal via a text-to-speech engine. Then, phoneme sequence is mapped with the visemes. Each phoneme is corresponded to one or multiple appropriate visemes to generate facial motion. Finally, the resulting facial motion is smoothly applied on the facial model to produce realistic speech-synchronized animation.

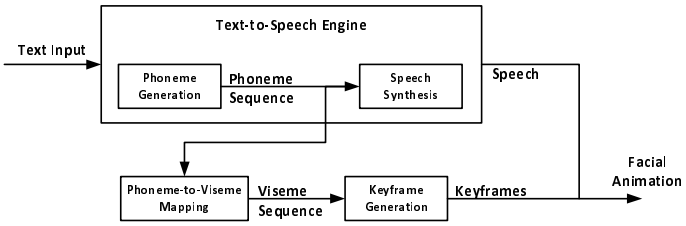

Fig. 2: System overview.

A. Text-To-Speech Conversion

Conversion of text to speech is realized by a text-to-speech engine. Our system uses Microsoft Speech API 5.1 (MSSAPI 5.1) [11] and Java Speech API 2.0 (JSAPI 2.0) [12] for web and mobile platforms, respectively. They allow incorporating speech synthesizing into the applications aimed at web and embedded devices. Given an input string, text-to-speech engines produce the corresponding synthetic speech data as well as side information in the form of phonemes along with their duration. We use Acapela prepared voices [13] which provide great realism.

B. Viseme Generation

To animate the motion of the face that corresponds to speech, visemes are constructed by mapping from the set of phonemes. Once the visemes for each time frame is created on the fly by blending process, they are interpolated and synchronized with the timing and phonetic parameters obtained from the speech data. For the interpolation, our system relies on linear interpolation using:

$$\nu(t_r) = \nu(t_0).(1-\omega) + \nu(t_1).\omega \tag{1}$$

where ω is an arbitrary weight such that $\omega \in [0,1]$, $\nu(t_0)$ and $\nu(t_1)$ are the vertices of previous and next visemes respectively, and $\nu(t_r)$ designates resultant viseme interpolated using these two.

C. Face Modeling

The proposed system is based on keyframe interpolation [14] that face motion is obtained by interpolating the visemes over time. Given a set of n facial expressions and corresponding face meshes $M = \{M_0, M_1, ..., M_n\}$, the resultant facial expression R is computed by blending different amounts of the original meshes M_i:

$$R = M_0 + \sum_{i=1}^{n} [\omega_i (M_i - M_0)] \tag{2}$$

where ω_i are arbitrary weights and M_0 denotes to a neutral expression. The distinctive facial features in face M_i become more exaggerated in R when ω_i gets larger. Setting up the weights such that the individual weights as well as their sum are between $[0,1]$ avoids such exaggerated expressions. For our system we use the FaceGen editor [15] to generate realistic 3D faces with numerous facial expressions. Fig. 2 shows a set of 3D models we created.

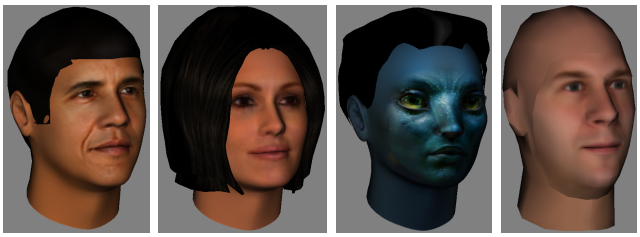

Fig. 3: A set of 3D models.

IV. Results

The mobile component of the 3D face animation system was implemented using Java ME with OpenGL ES (OpenGL for Embedded Systems) [16] support. Xface [17], an open source toolkit for creation of MPEG-4 and keyframe based 3D talking heads, is used for the web component of the system.

Our framework uses a tagged text input and produces the corresponding facial animation. The 3D face model speaks the input sentence with the indicated emotions. This virtual face can show six expressions: anger, disgust, joy, fear, surprise, sadness (Fig. 4 and Fig. 5) apart from neutral expression. The expression tags (i.e. <joy>Hello<\joy>) in the input turn into given emotion on the face while the face model is speaking. The keyframes in Fig. 4 and Fig. 5 correspond to <anger>, <disgust>, <joy>, <fear>, <surprise> and <sadness> tags in web/desktop and mobile environments, respectively. Fig. 6 and Fig.7 show a sequence of snapshots of 3D face model during speech.

V. Conclusion

A 3D facial animation system for web and mobile platforms is proposed. Our system generates facial movements with emotional expressions corresponding to speech. The input sentence is converted into speech and phonetic information. The phonemes are mapped into visemes and sent to the face model to realize facial movements. Final animation sequence is constructed by synchronizing the visual stream with the speech and interpolating the keyframes over time. The visual results of the animation are sufficient for web and mobile environments. The proposed system offers an affordable quick solution for various e-Health applications that require virtual actors speaking text in which human-machine interfaces can profit.

References

[1] "Assistive technology for children with learning difficulties," 2000, bridges To Reading, 2nd Edition.

[2] "Learning differences," association of Specialized and Cooperative Library Agencies, American Library Association.

[3] B. M. X. Li and C. I. Watson, "Expressive facial speech synthesis on a robotic platform," in *IEEE International Conference on Intelligent Robots and Systems, St. Louis, MO*, 2009, pp. 5009–5014.

[4] A. I. Xingyan Li, C. I. Watson and B. MacDonald, "Expressive speech for a virtual talking head," in *Australasian Conference on Robotics and Automation (ACRA), Sydney, Australia*, 2009.

[5] R. L. K. Scherer and K. Silverman, "Vocal cues to speaker affect: Testing two models," *Journal of the Acoustical Society of America*, vol. 76, no. 5, pp. 1346–1356, 1984.

[6] C. W. S. Roehling and B. Mac-Donald, "Towards expressive speech synthesis in english on a robotic platform," in *Australasian International Conference on Speech Science and Technology*, 2006, pp. 130–135.

[7] L. K. J. Danihelka and J. Zara, "Reduction of animated models for embedded devices," in *In WSCG 2010 Communication Papers Proceedings*, 2010.

[8] M. E. A. Wang and P. Faloutsos, "Assembling an expressive facial animation system," in *Proceedings of the 2007 ACM SIGGRAPH symposium on Video games*, New York, NY, USA, 2007, pp. 21–26.

[9] ISO/IEC 14496-1:1999. Information technology - Coding of audio-visual objects - Part 1: Systems. ISO, Geneva, Switzerland.

[10] ISO/IEC 14496-2:1999. Information technology - Coding of audio-visual objects - Part 2: Visual. ISO, Geneva, Switzerland.

[11] Microsoft, Inc. Speech SDK 5.1. www.microsoft.com/speech/download/sdk51.

[12] JSR 113: Java Speech API 2.0. http://jcp.org/aboutJava/communityprocess/final/jsr113/index.html.

[13] Acapela Group. http://www.acapela-group.com/index.html.

[14] J. Noh and U. Neumann, "A survey of facial modeling and animation techniques," University of Southern California, Tech. Rep. 99-705, 1998.

[15] FaceGen Modeller. http://www.facegen.com.

[16] Khronos Groups. OpenGL ES - The Standard for Embedded Accelerated 3D Graphics. http://www.khronos.org/opengles/.

[17] K. Balci, "Xface: Mpeg-4 based open source toolkit for 3d facial animation," in *Working Conference on Advanced Visual Interfaces*, Gallipoli, Italy, May 2004, pp. 399–402.

Fig. 4: Emotion keyframes: anger, disgust, joy, fear, surprise, sadness.

Fig. 5: Emotion keyframes rendered in mobile emulator: anger, disgust, joy, fear, surprise, sadness.

Fig. 6: Fragments from animation sequence during a speech.

Fig. 7: Fragments from animation sequence in mobile emulator.