Design of an MPEG4-Based Text-Driven 3D Facial Animation System

Ming-Shing Su[1,2], Ming-Tat Ko[1], Kuo-Young Cheng[1,2] and Chun-Yen Chen[1]

[1] Institute of Information Science, Academia Sinica, Taiwan.
[2] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.
{simon, mtko, kycheng, ccy}@iis.sinica.edu.tw

Abstract

In this paper, we present a text-driven talking head system, which consists of two major components, a text-to-speech (TTS) unit and a facial animation (FA) unit. We adopt a commercial text-to-speech system, Microsoft Speech API, to be our TTS unit and a feature-point-driven facial animation system based on hypothetical face method to be our FA unit. Given meaningful sentences, the text-to-speech system produces speech signals and speech-driven information that includes 2-D lips shape and timing information of the speech. In this paper, the 3-D facial actions of lips for speech using hypothetical face method are derived from the provided 2-D lips shape information and the facial actions of lips are interfaced with speech-driven information in the text-to-speech system in order to generate a more realistic 3-D lip motion animation for speech. Experiment conducted on thus accomplished text-driven 3D talking head system shows that the proposed method is highly acceptable.

Keywords: MPEG-4, text-driven, hypothetical face, text-to-speech, facial animation, lip synchronization.

1. Introduction

Recently, there exist several facial animation service applications in commercial use on Internet; such applications cover the display of virtual anchors, web-based intelligent virtual agents, and other virtual facial expression generators [1,4,6,8,12]. These applications are all based on the design of a text-driven 3D talking head system to produce vivid facial expressions. In such text-driven talking systems, the input is a meaningful sentence, which may be an arbitrarily user-typed or a computer auto-generated from some intelligent agents, and the output is a text-to-speech synchronized voice as well as lips motion to correspond to the input sentence. Some such systems also generate auxiliary facial expressions such as head nodding, smiling, and eye blinking to produce a realistic simulation result.

Apparently, a text-driven talking head system consists of two major components, a text-to-speech (TTS) unit and a facial animation (FA) unit. Wherein the TTS unit comprises a method to generate speech signals from an input sentence and at the same time to produce speech-driven information, which includes lips shape parameters and timing information, from the input sentence. Wherein the FA unit comprises a method to generate lips motions according to the speech-driven information and at the same time to produce auxiliary facial expressions according to some pre-defined rules. Thus, the graphical objects in a text-driven talking head system are synthetic; and the display of thus generated synthetic objects of the system are virtual. Then we may say that the design of a text-driven talking head system falls into the area of virtual reality in the multimedia application.

There are several methods for the representation of face objects, which may be used in the design of a talking head [9,13,16,17,18,19]. For example, the most often mentioned muscle-based method and the performance-driven method are two of them [18,19]. However, in order to be able to follow the object-based multimedia compression standard MPEG-4 [7], the facial animation unit of a text-driven talking head system would be feature-point-driven. As described in MPEG-4, a face is defined by 84 feature points on it, so that two different faces may be represented by two different distributions of the relative positions of their feature points respectively. Any change of location of the feature points on a face represents a facial action of the face.

Given an input text sentence, in addition to speech signals, most text-to-speech systems produce speech-driven information that includes a sequence of time and phoneme pairs. That is, an input text sentence is decomposed into a sequence of phonemes with timing. If a 2-D lips shape is provided for each phoneme, with the speech-driven information, it is easy to obtain a 2-D lips animation synchronized with the speech by displaying the 2-D lips shapes corresponding to the sequence of phonemes at their specified time. In the design of a text-driven talking head, thus we need a pre-defined 3-D mouth shape corresponding to each phoneme. The pre-

defined mouth shape must be expressed by the location of the MPEG-4 specified feature points. Our research interest is, therefore, to develop a facial animation method, which is MPEG-4 feature-point-driven, and can interface with any text-to-speech system that provides speech-driven information.

The facial animation method developed in this paper is based on the hypothetical face method proposed in [14,15], where a hypothetical face was developed to control the feature-point-driven facial animation. As described in [15], all facial actions that act on a real face model can be parameterized through a hypothetical face. A hypothetical face is basically a combined piecewise cubic Coon's surface, in which each surface piece controls virtually the same real face area bounded by the connected feature points. When feature points are moved, the hypothetical face changes its shape, which in turns deforms the real face model accordingly. On the other hand, a facial action parameter that describes how a real face model is deformed may be expressed in terms of a change of hypothetical face shape.

This paper discusses how the facial actions on lips of a hypothetical face based on MPEG-4 specified feature points are derived and how they are interfaced with speech-driven information in a commercial text-to-speech system, the Microsoft Speech API [11]. After the integration with the text-to-speech system, the result is a text-driven 3D talking head. Experiment conducted on thus accomplished text-driven 3D talking head system shows that the proposed method is highly acceptable.

The remaining sections are divided as follows. A general text-to-speech system and its mechanism for lips synchronization are described briefly in Section 2. A feature-point-driven face model based on hypothetical face control is described in Section 3. The method to generate three-dimension lip motions synchronized with text-to-speech is described in Section 4. Some experiment results are shown in Section 5, which is followed by concluding remarks in Section 6.

2. Brief description of text-to-speech unit

Apparently, a text-driven talking head system consists of two major components, a text-to-speech (TTS) unit and a facial animation (FA) unit. Figure 1 shows the architecture of a TTS unit. The input for the TTS unit is a series of meaningful sentences. The outputs of the TTS unit are the generated synthetic speech as well as the speech-driven information that includes lip shape parameters for the visualization of lip shape, and timing information for synchronizing audio with video results.

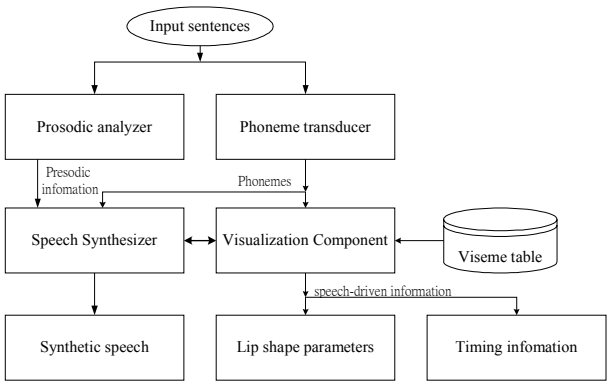


Fig 1. The architecture of a text-to-speech system

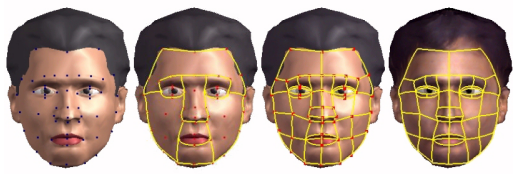
In most cases, the TTS unit is composed of four components, which are prosodic analyzer, phoneme transducer, speech synthesizer, and visualization component. Prosodic analyzer analyzes the input sentence and generates the corresponding prosodic information. Phoneme transducer accepts the text or words as input and produces a phonemic transcription as output. Speech synthesizer accepts both prosodic information and phonemic transcription to produce the synthetic speech, which may be obtained according to LPC-based, or formant-based, or PSOLA-based methods [3,10]. Further, these phonemes can be mapped into visual mouth shapes known as visemes. That is, visualization component can lookup a viseme table to generate lip shape parameters and its corresponding timing information for the purpose to generate and synchronize the visual output of graphic display.

In the following, we shall propose a facial animation model that can be easily controlled and can be integrated with commercial successful TTS unit such as Microsoft Speech API [11]. By integrating the FA unit with a TTS unit, a text-driven talking head system can be established.

3. A feature-point-driven facial animation using hypothetical face control

Facial animation is complicated because of complex biophysical structure of human face. Usually, the deformation of its skin is nonlinear. To achieve the intuitive and easy manipulation of facial expression, we adopt the hypothetical-face-based facial animation method [14,15]. For clarity and completeness, we describe this method briefly in the following. This method connects face feature points into a net and divides the whole face into regions that can be mathematically represented as a low order hypothetical surface. The shape control of a hypothetical surface is through manipulation of movement of the feature points in the surface. All the hypothetical surfaces form the hypothetical face. Then the deformation

of the hypothetical face drives the deformation of the corresponding face model. In this model, we adopt MPEG-4 specified facial feature points for the compatibility of the standard. The selected feature points in model are shown as Figure 2(a) and the constructed hypothetical face model is shown as Figure 2(d).



(a) (b) (c) (d)

Fig 2. A hypothetical face constructed from a neutral face model: (a) feature points, (b) 8 regions, (c) 48 subregions and (d) a constructed hypothetical face model.

Mathematically, each hypothetical surface can be expressed by a Coon's bilinear interpolation surface [2],

$$H(u,v)=\begin{bmatrix}1-u & u\end{bmatrix}\begin{bmatrix}B_2(v)\\ B_4(v)\end{bmatrix}+\begin{bmatrix}B_1(u) & B_3(u)\end{bmatrix}\begin{bmatrix}1-v\\ v\end{bmatrix}-\begin{bmatrix}1-u & u\end{bmatrix}\begin{bmatrix}P_{10} & P_{40}\\ P_{20} & P_{30}\end{bmatrix}\begin{bmatrix}1-v\\ v\end{bmatrix}, \quad (1)$$

where P_{10}, P_{20}, P_{30}, P_{40} are four corner (feature) points; $B_1(u)$, $B_2(v)$, $B_3(u)$, $B_4(v)$ for $u,v\in[0,1]$ are four boundary curves represented by piecewise Hermite cubic curves described below. For two adjacent feature points, P_i and P_{i+1} with tangents T_i and T_{i+1} respectively, the cubic Hermite curve is

$$C_i(u) = P_iH_0(u) + P_{i+1}H_1(u) + T_iH_2(u) + T_{i+1}H_3(u),$$

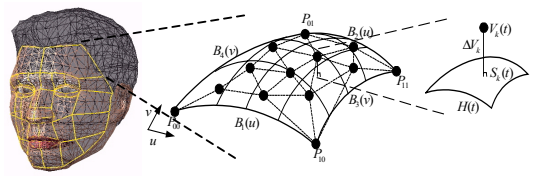
where $H_i(u)$, i = 0, 1, 2, 3 are Hermite cubic polynomial functions over $u\in[0,1]$ [2]. Let $C(u)$ be the boundary curve with $n+1$ feature points P_i, for $0 \le i \le n$, where P_i and P_{i+1} are adjacent for for $0 \le i \le n$-1, and let C_i be the given cubic Hermite curve between P_i and P_{i+1} for $0 \le i \le n$-1. Then $C(u)$ is defined as

$$\textit{if} \quad n=0, \quad C(u)=P_0 \quad u\in[0,1]$$

else

$$C(u)=\begin{cases} C_0(n\cdot u-0) & \textit{if } u\in[0,\frac{1}{n})\\ C_1(n\cdot u-1) & \textit{if } u\in[\frac{1}{n},\frac{2}{n})\\ \vdots & \vdots\\ C_{n-1}(n\cdot u-(n-1)) & \textit{if } u\in[\frac{n-1}{n},1]. \end{cases} \quad (2)$$

From Equations (1) and (2), the shape of a hypothetical surface is determined by the locations of its feature points and the tangents along the boundary curves on these feature points. It facilitates to control the deformation of a hypothetical surface simply by specifying the locations of feature points and their tangents along time.



(a) (b) (c)

Fig 3. Reconstruction of face model from hypothetical surface

Let $H(u,v)$ denote a hypothetical surface of the neutral face model. Let $H_t(u,v)$ denote a hypothetical surface and V^t denote a polygon vertex on the portion of the real model corresponding to the hypothetical surface at time t. Let $H(u_k,v_k)$ be the corresponding shortest distance projection point of V_k on the hypothetical surface $H(u,v)$, and ΔV_k the displacement vector between V_k and $H(u_k,v_k)$ shown in Figure 3(c), then the following equality holds:

$$V_k = H(u_k,v_k) + \Delta V_k. \quad (3)$$

Once the values of $\{V_k\}$ are obtained from a reference face model, say from a neutral face, then these values are remained unchanged during the reconstruction process of the deformed real face model. In other words, $\{V_k\}$ are pre-fabricated and pre-stored initial residual vectors. During the process of deformation, as time passed from t' to t and $H_{t'}(u_k,v_k)$ changed to $H_t(u_k,v_k)$, the new location of each V_k on real model at time t is determined by the following equation:

$$V_k^t = H_t(u_k,v_k) + \Delta V_k. \quad (4)$$

4. Integration between text-to-speech unit and facial animation unit

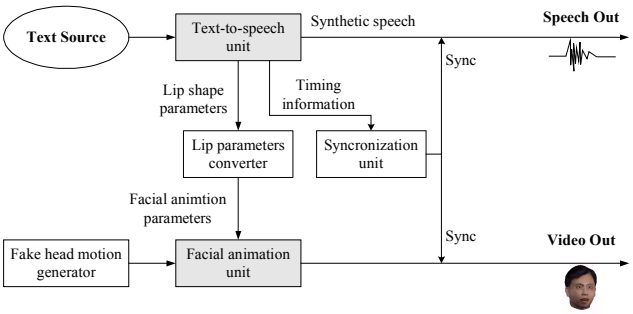


Fig 4. The architecture of text-driven talking head system

By integrating TTS unit and FA unit, we can establish a text-driven talking head system. Figure 4 illustrates the architecture of text-driven talking head system. For most existing text-to-speech commercial software, the timing information is provided. Although the TTS unit also generate lip shape parameters, there still exist several problems for producing three-dimensional facial expressions with these parameters. For example, a TTS provided lip shape parameters are only two-dimensional and portion of lip shape. In the following, we use Microsoft Speech API [11] as our TTS unit and the hypothetical-face-based model as the FA unit to discuss how a 3-D mouth shape can be generated from the captured 2-D information.

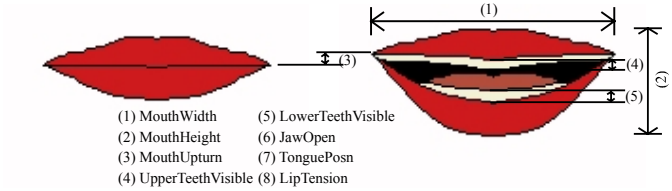


Fig 5. The lip shape parameters from TTS unit

The lip shape parameters of Microsoft TTS unit includes eight parameters, *MouthHeight, MouthWidth, MouthUpturn, TeethUpperVisible, TeethLowerVisible, JawOpen, TonguePosn,* and *LipTension*. Several parameters are shown in Figure 5.

In order to generate a more realistic 3-D lip motion animation, we need to consider:

a) when lip corner moves, how to change the cheek shape and how to estimate the depth variation of lip corner;
b) when lips extrude, how the depth of lip center changes; and
c) when in the state of mouth openness, how the jaw rotates.

4.1. Estimation of cheek variation

As shown in Figure 6, two muscles: the *zygomaticus* major and depressor *anguli oris* pull lip corner and determine how the cheek changes, wherein the *zygomaticus* major pulls lip corner upward to produce smiling expression and the depressor *anguli oris* pulls lip corner downward to produce sad expression. We shall base on this observation to simulate the variation. We can use the linear muscle as proposed in Waters' muscle model [13,19] to formulate their muscle action, which can be expressed as

$$P' = P + f \cdot akr \frac{\overrightarrow{PM}}{\left\|\overrightarrow{PM}\right\|}, \qquad (5)$$

where feature point P' is the new location from the displacement of P, f is the muscle strength, M is the fixed end point of muscle, a is the angular displacement parameter, r is radial displacement parameter, and k is a fixed constant representing the elasticity of skin. In this simplified muscle model, we assume the final displacement of the feature points is determined by accumulating the displacement of each muscle effect. As shown in Figure 6, we express this displacement relationship between feature points and muscles as

$$\Delta P_1 = P_1' - P_1 = f_1 \cdot a_1 k_1 r_1 \frac{\overrightarrow{P_1M_1}}{\left\|\overrightarrow{P_1M_1}\right\|} + f_2 \cdot a_2 k_2 r_2 \frac{\overrightarrow{P_1M_2}}{\left\|\overrightarrow{P_1M_2}\right\|}, \text{ and (6-1)}$$

$$\Delta P_2 = P_2' - P_2 = f_1 \cdot a_1 k_1 r_1 \frac{\overrightarrow{P_2M_1}}{\left\|\overrightarrow{P_2M_1}\right\|} + f_2 \cdot a_2 k_2 r_2 \frac{\overrightarrow{P_2M_2}}{\left\|\overrightarrow{P_2M_2}\right\|}, \qquad (6\text{-}2)$$

where P_1 and P_2 are the feature points on lip corner and cheek respectively, M_1 and M_2 are the fixed end points of the muscles, and f_1 and f_2 are the forces applied on these muscles. The muscle parameters, a_i, k_i, and r_i for i=1,2, can be determined by the characteristic of the muscle and the relative position of the feature points with respect to the muscle.

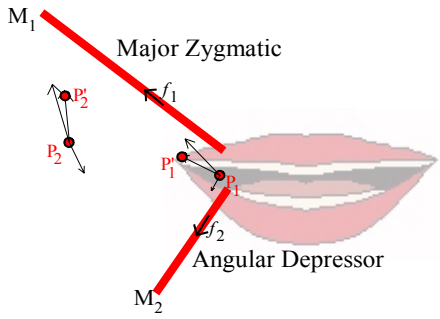


Fig 6. The relationship between lip corner, cheek, and facial muscles

The procedure to estimate lip corner depth and cheek 3D movement variation is described as follows:

a) get 2-D lip corner displacement (ΔP_{1x}, ΔP_{1y}) from the TTS unit provided 2-D lip shape parameters;
b) let x- and y- components of the displacement of P_1 of Equation (6-1) be equal to ΔP_{1x} and ΔP_{1y} respectively to obtain the muscle strength f_1 and f_2; and
c) substitute f_1 and f_2 back to Equations (6-1) and (6-2) to obtained the depth variation of P_1 and the 3-D displacement of P_2.

4.2. Estimation of lip extruding depth

As shown in Figure 7(a), assume the lip length before extruding is l, and for simplicity, assume lip center is a

line. So that when extruding occurs, the line shrinks to l_s = $s*l$, where l_s is the lip length after extruding, s is the lip centerline shrinking ratio, Δz is the height displacement, and l_p is the projection length of l_s onto *xy* plane.

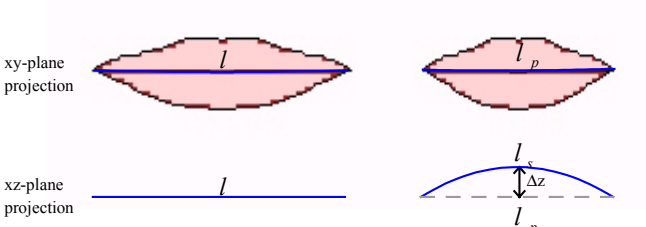


Fig 7. The relationship between lip corner, cheek, and facial muscles

It seems reasonable to let the shrinking ratio s depend on the projective length l_p and fall in $[\frac{l_p}{l},1]$. When s is set to be 1, it means the length l_s of lip centerline keeps the same as the length l during lip extruding. In this condition, we have the maximum depth variation of lips. On the other extreme condition, if $s=\frac{l_p}{l}$, then the lip length l_s becomes l_p during lips extruding, which means and the depth variation of lips is zero. Thus, we assume the shrinking ratio $s=(\frac{l_p}{l})^{1-\alpha}$ and express depth variation Δz as:

$$\Delta z \cong \frac{1}{2}\sqrt{l_s^{\,2}-l_p^{\,2}} = \frac{1}{2}\cdot l_p \cdot \sqrt{(\frac{l}{l_p})^{2\alpha}-1} \qquad (7)$$

where α falls in [0,1] and is a pre-defined constant so that the larger the α, the more tendency of lips out extruding. In our implementation, we choose α as 0.075.

4.3. Estimation of jaw rotation

When in the state of mouth openness, the jaw may be rotated. As shown in Figure 8, we assume θ is the jaw rotation angle, h_2 is the teeth separation gap when mouth open, h_1 is the perpendicular distance between upper teeth and jaw center, and r is the radius of the jaw. Then, we have $\theta=(\theta+\phi)-\phi=\tan^{-1}(\frac{h_1+h_2}{r})-\tan^{-1}(\frac{h_1}{r})$.

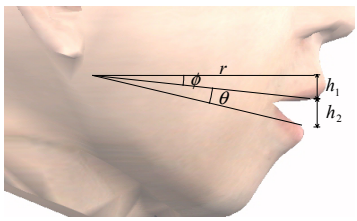


Fig 8. The rotation of jaw part

4.4 Generation of 3D MPEG4 facial animation parameters

Because the hypothetical-face-based model is controlled by MPEG-4 FAPs parameters, to animate this face model, we need to encode each 3D lip action using MPEG4 FAPs parameters. According the methods we have discussed above, the displacements of the facial feature points, including lip depth information and those on the cheek regions, for each lip action can be calculated from the text-to-speech unit generated lip shape parameters. Then, we encode these displacements of the facial feature points using the MPEG-4 FAP parameters in terms of facial animation parameter units (FAPUs). FAPU units are defined in order to allow interpretation of the FAPs on any facial model in a consistent way.

Figure 9 depicts the lip shape control parameters for the phoneme /w/ (sounds as woo). Figure 9(a) shows its eight parameters generated by Microsoft text-to-speech object. The constructed virtual lip object according to these eight parameters is shown in Figure 9(b). Figure 9(c) shows the encoded MPEG4 FAPs parameters for the same phoneme /w/. In the following, we show the encoding procedure for the parameter *push_t_lip* only. For the other FAP parameters, the encoding procedure is similar. The parameter *push_t_lip* stands for depth displacement of top middle lip. In the encoding, *push_t_lip* is expressed as

$$push\_t\_lip = \frac{1}{2}l_p \cdot \sqrt{(\frac{l}{l_p})^{2\alpha}-1}\cdot\frac{1024}{l} \quad \text{and}$$

$$\frac{l_p}{l} = 1+\frac{2}{5}\cdot\frac{MouthWidth-128}{255}, \qquad (8)$$

where l is the natural lip width, l_p is the lip width for phoneme /w/, and α is a pre-defined constant which is set as 0.075 in our implementation. In Microsoft text-to-speech lip shape parameters, *MouthWidth* ranges from 0 to 255, which stand for minimum and maximum lip width respectively. When *MouthWidth* is 128, the lip width l_p is equal to the natural lip width l. In our implementation, we take the minimum lip width as about $\frac{3}{5}$ natural lip width

and the maximum lip width as about $\frac{7}{5}$ natural lip width. After we get the ratio $\frac{l_p}{l}$, we can calculate the depth displacement of middle lip using the Equation (7) and encode it as the *push_t_lip* parameter in terms of FAPU units such as Equation (8). Because the FAPU unit for the *push_t_lip* parameter is *MW*, which is the length of $\frac{1}{1024}$ natural lip width *l*, we have $\frac{1024}{l}$ in Equation (8).

Lip parameters	Value
MouthWidth	0
MouthHeight	243
MouthUpturn	153
JawOpen	32
TeethUpperVisible	0
TeethLowerVisible	0
TonguePosn	0
LipTension	48

(a)

(b)

ID	FAP name	Value	Unit
3	open_jaw	182	MNS
4	lower_t_midlip	-143	MNS
5	raise_b_midlip	-143	MNS
6	stretch_l_cornerlip	-101	MW
7	stretch_r_cornerlip	-101	MW
12	raise_l_cornerlip	-34	MNS
13	raise_r_cornerlip	-34	MNS
14	thrust_jaw	-145	MNS
15	shift_jaw	0	MW
16	push_b_lip	75	MNS
17	push_t_lip	75	MNS
39	pull_l_cheek	-50	ES
40	pull_r_cheek	50	ES
44	raise_tongue	-197	MNS
45	thrust_tongue	-32	MW
46	raise_tongue	-147	MNS
51	lower_t_midlip_o	0	MNS
52	raise_b_midlip_o	-237	MNS
53	stretch_l_cornerlip_o	-101	MW
54	stretch_r_cornerlip_o	-101	MW
59	raise_l_cornerlip_o	-34	MNS
60	raise_r_cornerlip_o	-34	MNS

* **t: top, b: bottom, l: left, r: right, o:outer**

(c)

Figure 9: Generation of phonemes, (a) values of text-to-speech lip parameters for /w/, (b) the constructed 2D virtual lip object, and (c) values of FAP parameters for /w/.

4.5 Generation of 3-D lip motion animation synchronized with text-to-speech

Consequently, we can generate 3-D lip motions animation synchronized with the text-to-speech system by the following steps:

1. constructing a virtual lip object using TTS provided lip shape parameters;
2. generating 3-D facial actions of lips for speech according to the above procedures;
3. encoding the facial actions using MPEG-4 facial animation parameters (FAPs) by observing the movement of the specified facial feature points; and
4. using MPEG-4 FAPs parameters and TTS provided timing information to generate the 3-D lip motion animation based on hypothetical face method.

5. Experiments & Results

The performance of the proposed 3D talking head system depends on the details of the 3D models and the 3D rendering ability of graphics cards. Normally, the performance can achieve about 15-40 frames on a Pentium family PC with a 3D accelerator graphic card. So it can be applied in most real time applications.

Figure 10 shows the front view of the computer-generated lip shape when talking /h/ (sounds as he). In such action, the more the mouth is stretched open, the more the cheek is pull sideward according to Equation (6-1) and (6-2). However, it is difficult to present the effect of subtle change such as cheek variation in two-dimensional static images. To illustrate this effect more clearly, we draw the variation of the hypothetical surface and the feature point on the cheek onto the computer-generated face image. The blue lines (or dark lines) and point A are represented for the neutral face; the yellow lines (or light lines) point B are represented for the variation when talking /h/.

Figure 11 shows the side view of lip shape when talking /w/. To emphasize the extruding effect, this figure is synthesized by overlaying the lip figure without considering lip-extruding depth (marked with A) onto the one with considering lip-extruding depth (marked with B) using our estimation rule of Equation (7). In our video result, this slight extruding effect makes the synthesized lip motion animation more realistic.

Figure 12 shows several rendering results of lip motions in 3D talking head which is driven by text-to-speech.

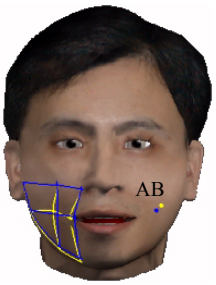


Fig 10. The front view of the lip motion for phoneme /h/

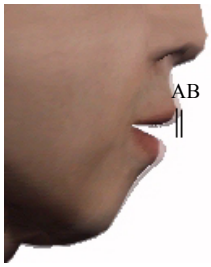


Fig 11. The side view of the lip extruding for the phoneme /w/

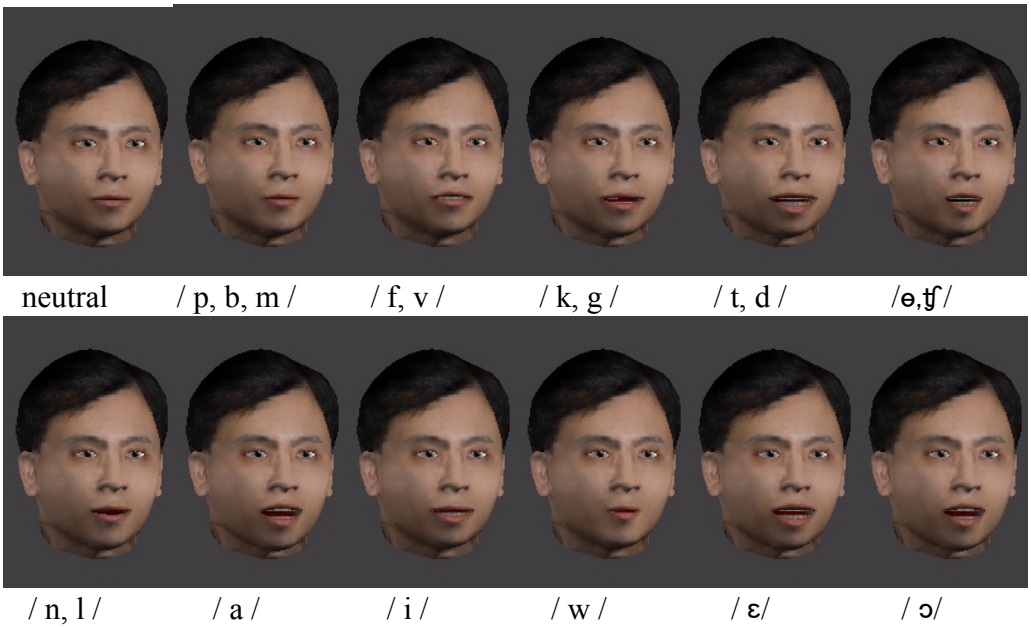


Figure 12: Several computer-generated lip motions of basic phonemes.

6. Conclusions

In this paper, we have presented a text-driven talking head system. The talking head system has been developed by integrating a facial animation unit and a text-to-speech unit. In the facial animation unit, we adopted the hypothetical face method that is feature-point-driven using MPEG4-specified feature points. In the text-to-speech unit, we adopted a commercial system, Microsoft Speech API, which provides speech information including lips shape information and timing information. The proposed integration method works whenever the text-to-speech unit can provide the above speech information. The integration includes (1) generation of each 3D lip action from the 2D lip shape information provided by the text-to-speech unit, and (2) encoding each 3D lip action using the MPEG-4 facial animation parameters. The 3D lip motion animation of the talking head is then generated by the facial animation unit according to these MPEG-4 facial animation parameters and the text-to-speech provided timing information. The experiment conducted on the accomplished text-driven 3D talking head system shows that this integration method is acceptable.

12. References

[1] Ananova Ltd., *Ananova*, http://www.ananova.com

[2] G. Farin. *Curves and Surface for Computer Aided Geometric Design – A Practical Guide.* Academia Press, Inc., San Deigo, CA, 1988.

[3] D. Bigorgne, O. Boëffard, B. Cherbonnel, F. Emerard, D. Karreur, J.L. Le Saint-Milon, I. Metayer, C. Sorin, and S.

White. Multilingual PSOLA text-to-speech system. *Proceeding of the InternationalConference on Acoustics, Speech and Signal Processing'93*, Mainneapolis, 2:187-190, 1993.

[4] R. Cole, T. Carmell, P. Connors, M. Macon, J. Wouters, J. de Villiers, A. Tarachow, D. Massaro, M. Cohen, J. Beskow, J. Yang, U. Meier, A. Waibel, P. Stone, G. Fortier, A. Davis, and C. Soland. Intelligent Animated Agents for Interactive Language Training, http://cslu.cse.ogi.edu/tm/ilt.html

[5] W.I. Hallahan. DECtalk Software: Text-to-Speech Technology and Implementation, http://www.europe.digital.com/info/DTJK01/

[6] Haptek Inc., *VirtualFriend Chat,* http://www.haptek.com/

[7] ISO/IEC 14496-2:1998, *Information Technology - Generic Coding of Audio-Visual Objects*, Part 2: Visual. 1998.

[8] Lucent Ltd., *Face2Face,* http://www.f2f-inc.com/

[9] P. Kalra, A. Mangili, N.M. Thalmann and D. Thalmann. Simulation of facial muscle actions based on rational free form deformation. *Eurographics*, 11(3):C59-69, 1992.

[10] D.H. Klatt. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America,* 67(3):971-995, 1987.

[11] Microsoft Co., *Microsoft Speech Technology SAPI 4.0 SDK*, http://www.microsoft.com/iit/sapisdk.htm

[12] Microsoft Co., *Research efforts related to Lifelike Computer Characters*, http://www.research.microsoft.com/research/ui/persona/related.htm

[13] F.I. Parke, and K. Waters. *Computer Facial Animation.* Wellesley Press, MA, 1996.

[14] M.S. Su, M.T. Ko, and K.Y. Cheng. Controlling 3-D lip motion using hypothetical surface functions. Submitted to *Journal of Information Science and Engineering*. Available at http://www.iis.sinica.edu.tw/VISUAL/Publications.

[15] M.S. Su, M.T. Ko, and K.Y. Cheng. Control of feature-point-driven facial animation using a hypothetical face. *Proceeding of the 8th Pacific Conference on Computer Graphics and Applications, Hong Kong,* 359-368, Oct 2000, Available at http://www.iis.sinica.edu.tw/VISUAL/Publications.

[16] H. Tao, and T.S. Huang. Bézier volume deformation model for facial animation and video tracking. *CAPTECH'98 Lecture Notes in Artificial Intelligence (LNAI) 1537*, 242-253, 1998.

[17] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. *Computer Graphics*, 29(4):55-62, August 1995.

[18] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(4):235-242, 1990.

[19] K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics*, 21(4):17-24, July 1987.