Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEG-4 FAP Features in a Three-dimensional Avatar

Zhiyong Wu[1], *Shen Zhang*[2], *Lianhong Cai*[2] *and Helen M. Meng*[1]

[1] Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N. T., Hong Kong SAR, China
[2] Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China
john.zy.wu@gmail.com, {zhangshen05, clh-dcs}@tsinghua.edu.cn, hmmeng@se.cuhk.edu.hk

Abstract

This paper describes our initial work in developing a real-time audio-visual Chinese speech synthesizer with a 3D expressive avatar. The avatar model is parameterized according to the MPEG-4 facial animation standard [1]. This standard offers a compact set of facial animation parameters (FAPs) and feature points (FPs) to enable realization of 20 Chinese visemes and 7 facial expressions (i.e. 27 *target facial configurations*). The Xface [2] open source toolkit enables us to define the influence zone for each FP and the deformation function that relates them. Hence we can easily animate a large number of coordinates in the 3D model by specifying values for a small set of FAPs and their FPs. FAP values for 27 target facial configurations were estimated from available corpora. We extended the *dominance blending approach* to effect animations for coarticulated visemes superposed with expression changes. We selected six sentiment-carrying text messages and synthesized expressive visual speech (for all expressions, in randomized order) with neutral audio speech. A perceptual experiment involving 11 subjects shows that they can identify the facial expression that matches the text message's sentiment 85% of the time.

Index Terms: visual speech synthesis, expression, MPEG-4

1. Introduction

This paper describes our initial step towards development of an expressive 3D avatar for audio-visual Chinese speech synthesis. Such avatars offer the potential of enhanced human-computer interaction in spoken dialog systems, automatic announcement systems, electronic books, etc. As was discussed in [3], an animated face model has much to offer in addition to synthetic speech, e.g. providing visual cues for dialog turn-taking in a dialog, signaling the system's internal state (i.e. "thinking"), sustaining intelligibility in noisy acoustic environments, etc. Facial expressions can further enhance interaction through nonverbal communication [4]. Our long term plan is to incorporate an expressive avatar for multimodal (audio-visual) response generation in a spoken dialog system. For this purpose, we need to synthesize an expressive avatar (with both audio and visual modalities) in real time as the dialog progresses.

This paper describes work on expressive visual speech synthesis. Our previous implementation of text-to-visual speech (TTVS) synthesis [5] involves direct control of thousands of coordinates in a 3D model to generate Chinese visemes (visual phonemes). This motivates us to seek a simpler parameterization method that aggregates facial model control over a compact set of points. In the following, we will describe the adoption of a standard-based parameterization, our procedures for parameter setting for target visemes and facial expressions, the dominance blending approach for dynamic parameterization that can incorporate *both* viseme coarticulation and facial expression changes, as well as preliminary evaluation results of our expressive Chinese visual speech synthesizer.

2. Background on Animation Platform

We adopt the MPEG-4 Facial Animation (FA) standard [1] for parameterization and use the Xface open source toolkit [2] to define and render these parameters.

The MPEG-4 FA standard defines 84 Feature Points (FP) to describe the shape of the face model which cover eyes, eyebrows, nose, mouth, tongue teeth, etc. The FPs are used in the definition of Facial Definition Parameters (FDP) for calibrating the shape of the head, as well as Facial Animation Parameters (FAP) for characterizing the deformations during animation (see [2] for illustration). FAPs are unitless values and are independent of the face models by virtue of calibration using Facial Animation Parameter Units (FAPUs). There are 68 FAPs in all. The first 2 FAPs are high-level parameters representing visemes and expressions respectively. The remaining 66 are low-level FAPs that define motions (e.g. translations and rotations) of such FPs as right corner of lip, inner corner of left eyebrow, etc. The high-level FAPs can drive animation through controlling the low-level FAPs. Hence for more direct control, our current work focuses on the 66 low level FAPs.

The Xface open source toolkit [2] offers the XfaceEd tool for defining the *influence zone* of each FP. More specifically, each FP is associated a group of points (non-FPs) in terms of animated movements. Xface also supports the definition of a *deformation function* for each influence zone and this function computes the displacement of a point as *influenced* by its associated FP during animation. Hence, a given MPEG-4 FAP values stream, together with corresponding FAP durations can be rendered as influence zones of animated position coordinates in a talking avatar.

3. Facial Animation Parameters (FAPs) for Chinese Visual Speech and Facial Expressions

3.1. Three-dimensional (3D) Chinese Avatar Model

We created a 3D avatar model with the image of a Chinese female spokesperson (Figure 1) using the software 3D Studio

Max. The avatar model specifies the 3D positional coordinates for animation and rendering, normal coordinates for lighting effects as well as texture coordinates for texture mapping. Both lighting and texture enhance the appearance of the avatar. The positional coordinates are connected to form a mesh of triangles that determine the *neutral coordinates* of the model.

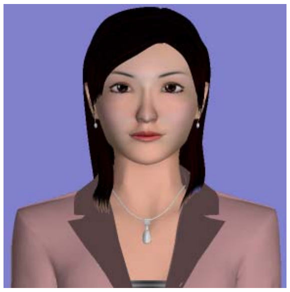

Figure 1: The 3D avatar model with the image of a Chinese female spokesperson.

3.2. FAPs for Chinese Visemes

As mentioned above, an MPEG-4 FAP value specifies the displacement of its corresponding FP from its neutral position. In order to animate a Chinese-speaking avatar, we need to define the FAP values corresponding to each Chinese viseme. Our current work focuses on Chinese Mandarin, a monosyllabic language/dialect where each syllable can be divided into an optional initial (i.e. a consonant) and final (i.e. a core vowel followed by an optional consonant). There are 21 initials and 38 finals in Mandarin, some of which share the same viseme. Hence, this work defines 20 basic visemes in all. In Table 1, initials/finals sharing the same viseme are grouped together. There is a viseme for "silence" (SIL), i.e. the natural state of the face without speaking.

Table 1: Definitions of basic visemes for Chinese Mandarin.

Initial		Final		
b, p, m	g, k, h	a, ang	ou	i
f	j, q, x	ai, an	e, eng	u
d, t, n	zh, ch, sh, r	ao	ei, en	v (/yu/)
l	z, c, s	o	er	SIL

In order to get FAP values for these Chinese visemes, we adopted a data-driven approach based on our previous work [6]. 3D FAPs were estimated from a video recording of orthogonal (frontal and side) views of a Chinese female speaker – this is achieved by placing a mirror next to the speaker's face (Figure 2). The recorded corpus covers all initials and finals in Chinese Mandarin. FPs around the nose and lip regions are tracked by a series of image processing techniques. More specifically, the frontal image is used for tracking the nostrils and the outer lip contour (from which the inner lip parameters were estimated), while the side image is used for tracking the protrusion of the upper and lower lip, the thrust and openness of the jaw.

We estimated values for 28 FPs related to mouth and jaw from the video recording to calculate FAPs for the creation of 20 Chinese visemes based on the 3D neutral avatar model. Figure 3 shows two example visemes – /b/ and /o/. The former viseme is identical to the neutral state (Figure 1), while the latter involves negative values for FAP4 (lower top middle lip) and FAP5 (raise bottom middle lip) etc to create rounded lips and positive values for FAP16 (push bottom lip) and FAP17 (push top lip) etc to create pushed lips in side view.

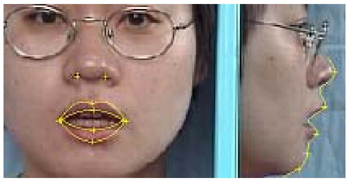

Figure 2: Tracking FAP values for Chinese visemes from both frontal and side views in a video recording (source: [6]).

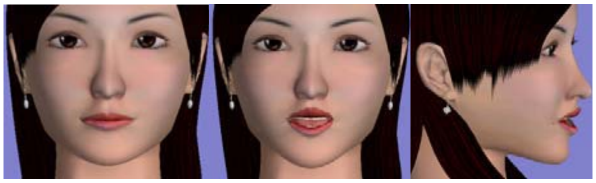

Figure 3: Chinese viseme models created for the 3D avatar. In order from left to right: viseme /b/ (frontal view), viseme /o/ (frontal view) and viseme /o/ (side view).

3.3. FAPs for Facial Expressions

MPEG-4 also defines 6 basic facial expressions – anger, disgust, fear, joy, sad and surprise. We set the FAP values manually by means of the XfaceEd tool (as described in Section 2). Various value settings are tried until the expression created for the avatar model matches those from the Japanese female facial expression database (JAFFE) [7]. Figure 4 shows the 6 basic expression models created for the 3D avatar. In total, we defined 48 FPs for facial expressions. These subsume the 28 FPs used for visemes since facial expressions involve the eyes and eyebrows, in addition to the mouth. The definitions also include FAPs for open and closed eyes.

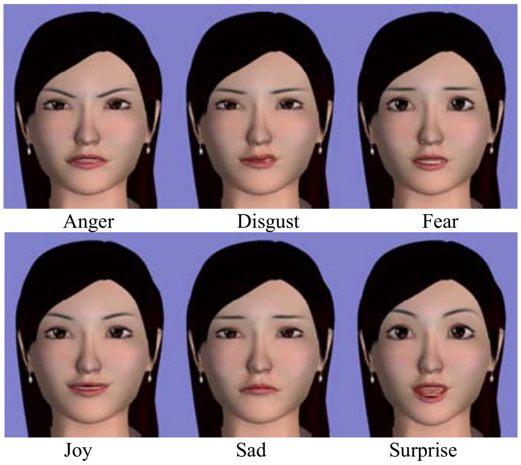

Figure 4: Six expression models created for the 3D avatar.

3.4. Definition of Influence Zone and Deformation Function

As mentioned in Section 2, each FAP corresponds to a set of FP and in turn, each FP corresponds to an influence zone of non-FP points. We used the XfaceEd tool to define influence zones for each FP in the eyes, eyebrows, and mouth regions. For example, FP2.3 (middle point of inner bottom lip) is related to FAP5 (raise bottom middle lip) as well as FAP16 (push bottom lip). FP2.3 is shown as the cross in Figure 5a and its influence zone is shown in terms of big dots. Similarly, FP4.1 (left inner eyebrow) is related to FAP31 (raise left inner eyebrow) and FAP37 (squeeze left inner eyebrow). FP4.1 is shown as the cross in Figure 5b and its influence zone as the group of big dots.

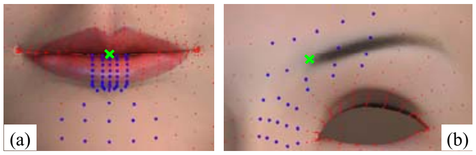

Figure 5: (a) Influence zone of FP2.3 (middle point of inner bottom lip). (b) Influence zone of FP4.1 (left inner eyebrow).

We need to define deformation functions to specify the displacements of points in the influence zone, in correlation with its FP movements. Xface offers a default deformation function based on a raised cosine function. We adopt this default for animation of the eyes and eyebrows. However, we adopt Radial Basis Functions (RBF) (see Equation 1) as the deformation function for the lip region (FPs 2.2, 2.3, 8.1, and 8.2), since RBFs offer a better fit for the shapes of the lips:

$$k(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}\|^2}{2\sigma^2}\right), \qquad (1)$$

where $\boldsymbol{x}$ is the 3D position coordinate of a vertex in the influence zone, $\boldsymbol{c}$ is the position coordinate of the feature point (FP), and σ is the distance between feature point (FP) and the farthest vertex in the influence zone. $k(\boldsymbol{x})$ is the displacement of the vertex $\boldsymbol{x}$ in the influence zone along the FAP direction.

4. Dynamic FAPs that superpose Visemes and Expressions with Dominance Blending

FAP values defined in the previous section refer to the target values for each Chinese viseme and facial expression. In order to animate a talking avatar in real-time, we need to generate FAP values for all time instants automatically and *dynamically* according to coarticulatory effects and changes of expressions.

We propose to leverage the *dominance blending* method previously implemented in our work on dynamic visemes [6]. The approach was first proposed by [8] according to gestural theory of speech production [9].

We extend the dominance blending method to encompass not only visemes but also facial expressions, and allow them to superpose during blending. Let p denote the FAP# of either a viseme or an expression i. The dominance function D_{ip} is:

$$D_{ip} = \begin{cases} e^{-\theta_{ip(-)}|\tau|}, & if \quad \tau \geq 0 \\ e^{-\theta_{ip(+)}|\tau|}, & if \quad \tau < 0 \end{cases}, \tau = t_{ci} - t, \qquad (2)$$

where t is current time; t_{ci} is the time of attaining target FAP values based on current viseme or (changed) expression; $\theta_{ip(-)}$ and $\theta_{ip(+)}$ represent the exponential decay before and after t_{ci}.

We defined 28 dominance functions in all – 19 for non-silence visemes, one for left silence viseme, another for right silence viseme, one for the neutral face and 6 for the facial expressions. Weighted dominance blending follows Equation 3 for FAP p at time t:

$$F_p(t) = \left(\sum_{i=1}^{n} D_{ip}(t) \times T_{ip}\right) \Big/ \sum_{i=1}^{n} D_{ip}(t), \qquad (3)$$

where T_{ip} is the target value for FAP p according to the current viseme or expression i; $n(=28)$ is the total number of visemes and expressions.

As described in previous sections, the values of parameters T_{ip}, $\theta_{ip(-)}$, and $\theta_{ip(+)}$ can be estimated from or set with reference to training data. Figure 6 illustrates the dominance functions and related FAP values over time for visemes corresponding to /er4 ba2/ (i.e. Chinese digits '2, 8') superposed on a change of expression from "neutral" to "surprise". However, our current training data is very limited. We need to collect adequate data to sufficiently model the various contexts of coarticulation with expression changes.

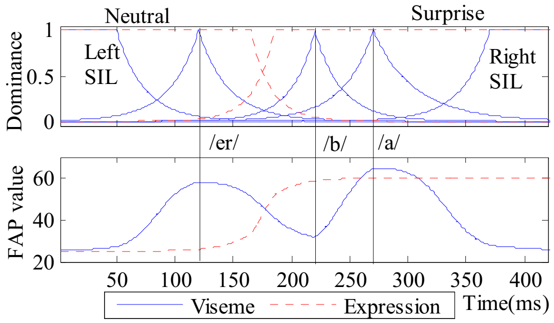

Figure 6: Variations of dominance functions and related FAP values over time for viseme and expression.

5. Integrated System for Expressive Chinese Visual Speech Synthesis

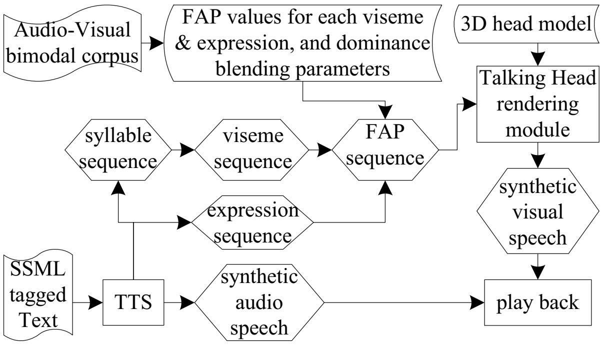

Figure 7: Process control flow for the integrated expressive Chinese visual speech synthesis system.

The process control flow of the integrated system for our expressive Chinese visual speech synthesis is shown in Figure 7. System input is specified using a proposed extension for the SSML (Speech Synthesis Markup Language) [10]. For example:

好久没有见到的老朋友又碰上了,真是<expression="joy">太高兴了.</expression>
Translation: He was delighted to meet old friends and said: <expression ="joy">"It's so nice to see you!!"</expression>

The input Chinese text is first translated into Chinese syllable sequence and expression sequence with timing information. The syllable sequence is further mapped to visemes. The dominance blending method is then used to augment the target FAPs based on the input's visemes and expressions to generate a sequence of dynamic FAPs. These are then sent to the rendering module to animate expressive Chinese visual speech, which is played back with the synthesized audio speech.

6. Evaluation of Expressivity in Visual Speech

We designed a set of perceptual experiments to assess the expressivity of the synthesized visual speech. We selected six sentences whose messages carry inherent emotions (Table 2):

Table 2: Text messages used carrying inherent emotions.

两次!三次!四次!五次!六次!第六次了!!大周末的,还让不让人睡啦!气死我啦!一把抓起听筒: “你有毛病啊!” Coarse translation: angry about a very noisy telephone
令人讨厌的表弟又来打扰自己,见到他一股厌恶之情就涌上心头:“你怎么又来了!” Coarse translation: disgusted by an annoying cousin
一个人走在深夜里,周围死一般的沉寂.突然传来一阵“呜呜”的呻吟声,真是太恐怖了! Coarse translation: fear about walking in the dark
好久没有见到的老朋友又碰上了,真是太高兴了. Coarse translation: joyous in seeing a long-missed friend
听到一个老朋友生病过世了,真是太伤心了. Coarse translation: sadness due to a friend passing away
我张大了嘴巴,眼珠子差点没跳出来.天哪!太不可思议了!金字塔竟然这么大! Coarse translation: surprise due to the size of the pyramid

For each textual message, we generate the *neutral* synthetic audio speech (expressionless), together with *expressive visual* speech for all the six facial expressions but in randomized order. We also applied a random number generator to create blinking effect while the avatar is speaking. The audio-visual synthetic speech outputs for each textual message are then presented to a group of 11 subjects in a quiet conference room with audio-video projection facilities. Each subject is asked to identify the synthesized facial expression that is most appropriate for the textual message. A subject is allowed to select more than one facial expressions for a given textual message as long as they are found to be fitting.

Results indicate that the percentage of correct perception is 85%. Major confusions include: perceiving anger for disgust; fear for sadness; and surprised for joy. These may be understandable because the confusable emotion pairs share some commonalities between the pair.

7. Conclusions and Future Work

This paper describes our initial work in developing a real-time audio-visual Chinese speech synthesizer with a 3D expressive avatar. The 3D avatar model is parameterized according to the MPEG-4 facial animation standard. This standard includes 66 low-level facial animation parameters (FAPs) corresponding to movement of 84 feature points (FPs) for the entire face model. In particular, this work focuses on 48 FPs in the eyes, eyebrows and mouth areas. These FPs are used during animation to realize 20 Chinese visemes and 7 facial expressions in the 3D avatar model. Each FP corresponds to an influence zone of finer points for face animation, which we defined manually by means of the Xface open source toolkit. A deformation function is used to displace points in the influence zone in relation to the FP movement. Target FAP values for the Chinese visemes are estimated from a small set of syllable recordings, while those for facial expressions are adjusted from the JAFFE database. We extended the dominance blending approach in order to effect animation for coarticulated visemes superposed with expression changes. Perceptual experiment on the evaluation of expressivity in visual speech shows that the matching facial expression can be identified 85% of the time by 11 subjects. In the near future, we plan to collect an audio-visual bimodal corpus to support better modeling of viseme coarticulation, dynamic expression changes and especially expressive visemes (e.g. in hyperarticulation). We also plan to integrate expressive visual speech synthesis with expressive audio speech synthesis.

8. Acknowledgments

This work is support by the research fund from the National Natural Science Foundation of China (NSFC) under grant No. 60433030 and the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) under grant No. 60418012 and N-CUHK417/04. This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies. We thank Mr. L.H. Cheung for creating the 3D avatar model.

9. References

[1] Motion Pictures Expert Group, ISO/IEC 14496-2: 1999/Amd. 1: 2000(E). International Standard, Information Technology – Coding of Audio-Visual Objects. Part 2: Visual; Amendment 1: Visual Extensions.

[2] Balci, K., "Xface: MPEG-4 based Open Source Toolkit for 3D Facial Animation", *Proc. Advance Visual Interfaces*, 399-402, 2004.

[3] Cassell, J., "Embodied Conversational Agents: Representation and Intelligence in User Interface", *AI magazine*, 22(3): 67-83, 2001.

[4] Branstrom, B. and House, D., "Multimodality and Speech Technology: Verbal and Non-verbal Communication in Talking Agents", *Proc. Eurospeech*, 2901-2904, 2003.

[5] Wang, J.Q., Wong, K.H., Heng, P.A., Meng, H., and Wong, T.T., "A Real-time Cantonese Text-to-Audiovisual Speech Synthesizer", *Proc. ICASSP*, 653-656, 2004.

[6] Wang, Z.M., Cai, L.H., and Ai, H.Z., "A Dynamic Viseme Model for Personalizing a Talking Head", *Proc. 6th Int. Conf. on Signal Processing*, 2:1015-1018, 2002.

[7] Lyons, M.J., Akamatsu, S. et al, "Automatic Classification of Single Facial Images", *IEEE Trans on Pattern Analysis and Machine Intelligence*, 21(12):1357-1362, 1999.

[8] Cohen, M.M. and Massaro, D.W., "Modeling Coarticulation in Synthetic Visual Speech", *Models and Techniques in Computer Animation*, 139-156, 1993.

[9] Lofqvist, A., "Speech as Audible Gestures", *Speech Production and Speech Modeling*, 289-322, 1990.

[10] SSML: http://www.w3.org/TR/speech-synthesis/.