An Online Speech Driven Talking Head System

Kai Zhao, Zhiyong Wu
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University
Shenzhen, China
zk69052@163.com, zywu@sz.tsinghua.edu.cn

Jia Jia, Lianhong Cai
Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University
Beijing, China
{jjia, clh-dcs}@tsinghua.edu.cn

Abstract—**This paper presents the design and implementation of an online speech driven talking head animation system. The system first recognizes phoneme sequence from the input speech with a Chinese Mandarin speech recognizer. The phoneme sequence is further transformed to a sequence of visemes. The sequence of MPEG-4 facial animation parameters (FAPs) is further derived from the viseme sequence, and is used to drive the facial animations on a 3-dimentional talking head. The architecture and the major features are also presented in the paper, together with the evaluations of the system.**

Keywords—visual speech synthesis; talking head; facial animation parameters (FAPs); viseme

I. INTRODUCTION

Vision and audition are the most important, natural and efficient forms of communication between humans. Visual speech synthesis is to generate visual articulation movements on a talking head accompanying audio speech [1]. Visual speech synthesis can be roughly categorized as text-driven and speech-driven approaches [2]. The ultimate goal of text-driven visual speech synthesis is to create a machine that is able to generate, from a text string, expressive audio-visual speech that is indistinguishable from the speech produced by a human [3]. While speech-driven visual speech synthesis takes the human speech as input, transforms it into facial parameters and then generates facial animation. [4] proposes a method to generate Chinese visual speech and facial expressions using MPEG-4 FAP [5] parameters.

This paper presents an online speech driven talking head animation system that can be accessed through web browser. Details of the system are to be described as below.

II. SYSTEM DESIGN AND IMPLEMENTATION

Architecture of the proposed system is illustrated as in Figure 1. The user visits our online system through web browser. The speech input will be recorded by the client program. The speech input is then transferred to our web server, where the input speech is recognized into a sequence of phonemes by the homegrown speech recognizer. The recognized phoneme sequence is further transformed into a sequence of Chinese visemes. The viseme sequence is further converted to a sequence of FAPs, which are finally used to drive the facial animations on a 3-dimentional talking avatar.

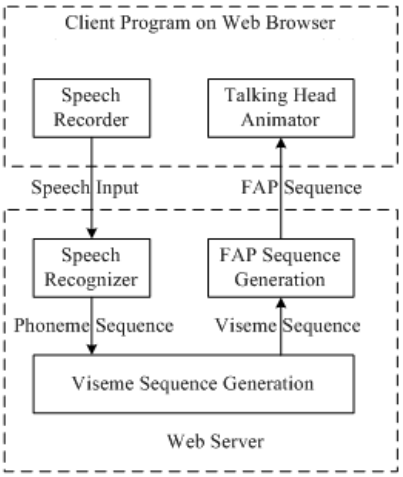

Figure 1. Architecture of the proposed speech driven talking head system.

A. Speech Recognition and Viseme Sequence Generation

Speech driven talking head animation generally involves two steps: 1) to generate viseme sequence from the input speech, and 2) to synthesize facial animations on the talking head model according to the viseme sequence. The results of viseme sequence generation greatly affect the performance of facial animation synthesis.

The problem of viseme sequence generation from input speech can be defined as finding optimal viseme sequence V' for the input speech **W**, which can be formulated as:

$$\mathbf{V}' = \arg\max_{\mathbf{V}} P(\mathbf{V} \mid \mathbf{W}) \quad (1)$$

where $P(\mathbf{V}|\mathbf{W})$ is the probability of observing the veseme sequence **V** given the input speech **W**. The input speech **W** can be represented by a sequence of acoustic feature frames $\mathbf{W}=(w_1,w_2,\ldots,w_m,\ldots,w_M)$, where M equals to the number of feature frames of the input speech and w_m denotes the m-th frame. The viseme sequence **V** consists of a sequence of visemes generated from the input speech, and can be represented as $\mathbf{V}=(v_1,v_2,\ldots,v_l,\ldots,v_L)$, where L is the number of visemes for the speech and v_l denotes the l-th viseme.

By assuming that corresponding phoneme sequence is $\mathbf{O}=(o_1,o_2,\ldots,o_l,\ldots,o_L)$, where o_l denotes the l-th phoneme of the phoneme sequence, where phoneme sequence **O** is produced by the speech recognizer, the problem of phoneme sequence

This work is supported by the National Natural Science Foundation of China (60928005, 60805008, 60931160443), the Upgrading Plan Project of Shenzhen Key Laboratory and the Science and Technology R&D Funding of the Shenzhen Municipal.

generation is to find the corresponding viseme v_l given the phoneme o_l. This work defines phoneme-viseme mapping table (Table 1) for converting phoneme sequence to viseme sequence.

TABLE I. PHONEME TO VISEME MAPPING TABLE
(where 'Vsm#' means 'Viseme number', 'Phs' means 'Phonemes')

Vsm#	Phs	Vsm#	Phones	Vsm#	Phs	Vsm#	Phs
0	SIL	5	j,q,x	10	ao	15	ou
1	b,p,m	6	zh,ch,sh,r	11	e,eng	16	u
2	f	7	z,c,s	12	ei,en	17	v /yu/
3	d,t,n,l	8	a,ang	13	er	18	i (/zi/)
4	g,k,h	9	ai,an	14	o	19	-i (/zhi/)

B. FAP Sequence Generation

To generate the FAP sequence from the viseme sequence, dominance blending method is adopted, which has also been implemented in our previous work on dynamic visemes [6] according to the speech production theory on articulators [7].

Let p denote the p-th FAP of the current viseme i, the dominance function D_{ip} is then defined as:

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

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

Totally 21 dominance functions are defined in the system, where 19 dominance functions are for non-silence visemes (i.e. viseme 1 to 19 in Table 1), one dominance function for left silence viseme (i.e. viseme 0 in Table 1, representing the state transition from silence to non-silence viseme), and one dominance for right silence viseme (i.e. viseme 0 in Table 1, representing the transition from non-silence viseme to silence).

The FAP sequence is then generated by computing the FAP value $F_p(t)$ for the p-th FAP at time t:

$$F_p(t) = \frac{\sum_{i=1}^{n} \left(D_{ip}(t) \times T_{ip} \right)}{\sum_{i=1}^{n} D_{ip}(t)} \quad (3)$$

where T_{ip} is the target value for FAP p according to the current viseme i, n(=21) is the total number of visemes.

The values of parameters T_{ip}, $\theta_{ip(-)}$ and $\theta_{ip(+)}$ are estimated from the training data. Figure 2 illustrates the variations of the dominance functions and related FAP values over time for visemes corresponding to /er4 ba1/ (for the pronunciation of Chinese digits '2' and '8').

C. Talking Head Animation

Talking head animation is generated by applying the FAP sequence on a three-dimensional (3D) talking head model with the image of a Chinese female speaker (Figure 3). The model was created, which specifies the 3D positional coordinates for talking head animation and rendering. These positional coordinates are connected to form a mesh of triangles that determine the initial coordinates of the model without head animation.

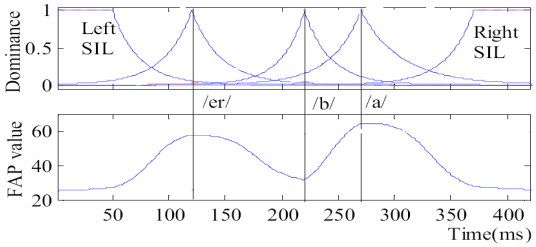

Figure 2. Dominance functions and related FAP values over time for the visemes /er/, /b/ and /a/.

Figure 3. The 3D talking head model with the image of a female speaker.

Each FAP is associated with a set of positional coordinates of the 3D talking head model. The change of the FAP values will lead to the change of the positional coordinates for the mesh of triangles of the model. In this way, talking head animation is achieved by computing the FAPs at a specific time stamp and then applying the FAPs to the 3D talking head mode.

III. CONCLUSIONS

This paper presents the design and implementation of an online speech driven talking head animation system. The system first recognizes phoneme sequence from the input speech with a Chinese Mandarin speech recognizer. The phoneme sequence is further transformed to a sequence of visemes. The sequence of MPEG-4 facial animation parameters (FAPs) is further derived from the viseme sequence and is used to drive the facial animations. In the future, we will port the system to embedded devices.

REFERENCES

[1] G. Bailly, M. Berar, F. Elisei, and M. Odiso, "Audiovisual speech synthesis," J. Speech Technology, Netherlands, 2003, pp. 331-346.

[2] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from HMM," in ICASSP, vol. 6, 1998, pp. 3745-3748.

[3] B. Theobald. "Audiovisual speech synthesis," in ICPhS, 2007, pp. 6-10.

[4] Z.Y. Wu, S. Zhang, L.H. Cai, and H. Meng, "Real time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar," in ICSLP, 2006, pp. 1802-1805.

[5] I.S. Pandzi, and R. Forchheimer, "MPEG-4 facial animation," 2002.

[6] Z.M. Wang, L.H. Cai, and H.Z. Ai, "A dynamic viseme model for personalizing a talking head," in ICSP, 2002.

[7] H.H. Bothe, and R. Frauke, "Visual speech and coarticulation effects," in ICASSP, 1993, pp. 634-637.