Co-articulation Generation Using Maximum Direction Change and Apparent Motion for Chinese Visual Speech Synthesis

CHUNG-HSIEN WU, CHUNG-HAN LEE and ZE-JING CHUANG

Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, 701 Taiwan
chunghsienwu@gmail.com

***Abstract*—This study presents an approach for automated lip synchronization and smoothing for Chinese visual speech synthesis. A facial animation system with synchronization algorithm is also developed to visualize an existent Text-To-Speech system. Motion parameters for each viseme are first constructed from video footage of a human speaker. To synchronize the parameter set sequence and speech signal, a maximum direction change algorithm is also proposed to select significant parameter set sequences according to the speech duration. Moreover, to improve the smoothness of co-articulation part under a high speaking rate, four phoneme-dependent co-articulation functions are generated by integrating the Bernstein-Bézier curve and apparent motion property. A Chinese visual speech synthesis system is built to evaluate the proposed approach. The synthesis result of the proposed system is compared to the real speaker. The coarticulation generated by the proposed approach is also evaluated.**

Keywords:Talking head, apparent motion, non-uniform rational B-spline (NURBS)

I. INTRODUCTION

Visual speech synthesis system has been investigated since the early 1980s [1]. Owing to the rising popularity of facial animation technology, agent systems, health care systems and dialog systems [2-4], is increasing in using of 3D virtual characters in human-machine interfaces. In recent decades, the integration of speech synthesis and facial animation has permitted the development of many applications, particularly in visual human computer interfaces using agents [5-7]. Visual speech synthesis involves integrating and synchronizing audio information performance and visual information perception. Accordingly, some approaches collect audio and video data synchronously, and train the relationships between them to prevent mismatches when producing talking head animation [8-10]. However, collecting a new audiovisual training corpus does not benefit the promotion of an existent Text-To-Speech (TTS) system. Excellent results are also produced by image-based method [11-12]. Unfortunately, the image-based method cannot produce a talking head from an arbitrary viewpoint. A further facial animation (FA) system is needed to visualize an existing TTS system. Following the collection of motion parameters of each phone, the FA system selects the corresponding motion parameters according to input text, aligns the parameters to the phone boundaries obtained from TTS system, and derives the coarticulation-of-motion parameters.

The motion parameters are transformed into to facial mesh deformation by the articulatory control model in the FA system. Many articulatory control models have been investigated. The parameterized model [13-16] applies the parameters that control the facial movement to the 2D or 3D face model, and produces face expressions or lip motion. Coarticulation problem is another significant component of the FA system. Many approaches have already been developed to model coarticulation, including Hidden-Markov-Model (HMM) [11], tri-phone model [9], or Active Appearance Model (AAM) [12]. A dominance function is defined by the exponential function and is applied to each synthesis unit [18]. The coarticulation components of a complete sentence are generated by overlapping all successive dominance functions. The dynamic viseme model [17], which uses three different kinds of dominance function for the generation of coarticulation component, has been developed to solve this problem.

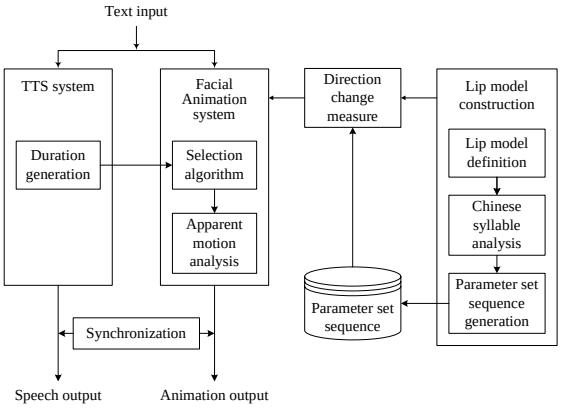

Figure 1. System block diagrams.

Fig. 1 shows the block diagram of a Visual Speech Synthesis system for Mandarin Chinese presented in this paper. The diagram comprises three main components, the TTS system, the FA system and the motion parameter database generation system. The TTS system is pre-existent, and can transfer the input word sequence to speech output. With the input of duration information provided by TTS system and motion parameter sequence, the FA system transfers the input word sequence to facial animation output.

In the coarticulation generation, a Bernstein-Bézier curve [19] is adopted to simulate the coarticulation component, and the parameter sets in the coarticulation component are modified according to the simulated curve. Furthermore, this study concentrates on the system re-development with different speech performers and variable speaking rate. To synchronize the synthesized speech and lip motion, a maximum direction change algorithm is proposed to select key parameter sets from the parameter set sequence in the synthesis units. Since the parameter sets are selected based on the duration information, the selection result under a fast speaking rate may contain only a few important key frames. A rapid variance may occur on the lip motion when the system estimates the coarticulation component according to these key frames. To avoid this problem, the parameter sets within the coarticulation are further smoothed based on Apparent Motion theory, which is the approximate real motion of an alternate frame sequence received by a human visual system [19-20] based on psychophysics. Finally, to avoid complex calculations and to achieve real-time processing, the system specifies a parameterized articulatory control model.

This paper is organized as follows. Section 2 describes the construction of motion parameters. Section 3 describes the calculation of the direction change for the synchronization of lip motion. Section 4 analyzes and adopts the effect of apparent motion property to improve the smoothness in transition. Finally, Section 5 summarizes the simulation results, followed by conclusions.

II. MOTION PARAMETERS CONSTRUCTION FOR MANDARIN

A. Unit Selection for Visual Speech Synthesis

Mandarin Chinese has 16 vowels and 21 consonants. The vowels are clustered into three categories based on acoustic phonetics as Table 1. The 21 consonants are also categorized based on the position of articulation as Table 2. Mandarin Chinese has a total of 408 different phones, or consonant-vowel combinations. The 408 phones are manually grouped to 105 visually distinguishable visemes. Accordingly, 105 syllables, each for one viseme, are chosen as the synthesis units for visual speech synthesis.

TABLE I. CATEGORIES OF VOWELS IN MANDARIN SPEECH

Vowel Type	Vowel
Single Vowel	a, o, e, e, er, yi, wu, yu
Diphthong	ai, ei, ao, ou
Nasal Vowel	an, en, ang, eng

TABLE II. CATEGORIES OF CONSONANTS IN MANDARIN SPEECH

Place of Articulation	Consonant
Bilabial	b, p, m, f
dental-alveolar	d, t, n, l
Velar	g, k, h
alveolar-palatal	j, q, x
retroflex	zh, ch, sh, r
apical-dental	z, c, s

B. Feature Point Definition

The term "feature point", as distinct from the "control point" used in Bern-stein-Bézier curve is used here to signify the points that control the deformation of facial mesh. Fig. 2 displays 56 feature points, which are defined to characterize the facial motions. As shown in the left part of Fig. 2 six feature points are controlled in half the face, under the assumption that the human face is symmetrical. In the right part of Fig. 2, 47 feature points are defined around the lip line in three different layers. The lip corner can be modeled by the feature points in the outer and middle layers, and the three-layered feature point definition can even model some complex lip motion, such as pouting or furling the lip.

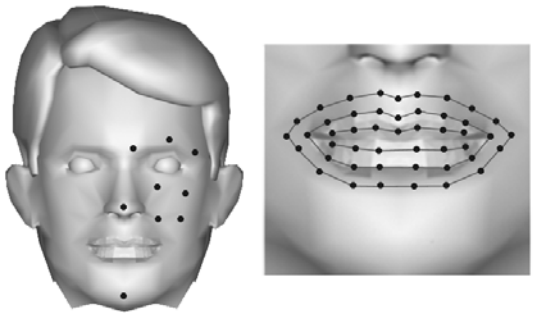

Figure 2. The location of feature points. Feature points used for lip motion are shown in the right part, and the other feature points are shown in left part.

Video footage for the 105 synthesis units was recorded with black markers on the facial feature points of the human speaker. To capture the feature point movement in three dimensions, the video footage was recorded on the front, left and right sides simultaneously. A simple feature tracking method was utilized to trace the black markers in the image sequence automatically. Given a feature point with location (x_t, y_t) at frame t, the location of this feature point at frame $t+1$ is determined using

$$(x_{t+1}, y_{t+1}) = \arg\min_{(x_t+u,\, y_t+v)} \left[\sum_{i=\frac{-m}{2}}^{\frac{m}{2}} \sum_{j=\frac{-m}{2}}^{\frac{m}{2}} \left[I_t(x_t+i, y_t+j) - I_{t+1}((x_t+u)+i, (y_t+v)+j) \right] \right]^2 \quad (1)$$

$$-\frac{b}{2} \le u \le \frac{b}{2}, -\frac{b}{2} \le v \le \frac{b}{2}$$

where $I_t(x, y)$ denotes the intensity of pixel (x, y) at frame t; (x_t+1, y_t+1) denotes the new location of the feature point at frame $t+1$; m denotes the mask width, and b denotes the block size containing possible locations of the feature point at frame $t+1$. The locations of the feature points were assumed to change smoothly in a short period of time, so that only a region of size b is necessary for searching. A feature point sequence is obtained for each synthesis unit by applying the feature tracking technique and normalization.

III. MAXIMUM DIRECTION CHANGE ALGORITHM

Synchronization of lip motion and speech is significant in visual speech synthesis. To perform the goal of synchronization, the speech signal and lip motion are usually re-corded simultaneously by the same speaker. Although this approach can produce synchronized visual speech synthesis,

it results in some problems when the user changes the speaker or the speaking rate. For example, when the user increases the speaking speed, the system has to shorten the parameter set sequence to fit the new speech duration by eliminating several parameter sets, possibly removing some important parameter sets. One solution to this problem is to annotate the importance level of each parameter set, and select the parameter sets with higher importance level first in the synthesis process. To achieve the goal, a maximum direction change algorithm is proposed to choose the key lip motion parameter sets from the parameter set sequence of the synthesis units.

The definition of the direction change measure is based on the movement of the feature points in the video footage of a human speaker. For each synthesis unit, $P^k = \{p_1^k, p_2^k, ..., p_N^k\}$ denotes the location sequence of feature point *k*, where *N* is the size of the location sequence and $p_t^k = (x_t^k, y_t^k)$ is the *x*-axis and *y*-axis at location *t*. The corresponding direction change measure of the location sequence P^k is $S^k = \{s_1^k = 1, s_2^k, s_3^k, ..., s_{N-1}^k, s_N^k = 1\}$, where s_t^k, which ranges from 0 to 1, denotes the corresponding direction change measure for p_t^k. The first and the last parameter sets, where t=1 and t=n respectively, are assumed to be necessary, and to have the highest direction change measure selected as the animation frames. Hence, the direction change measures for these two parameter sets are set to 1. The direction change measure s_t^k is estimated from the direction change between vectors (p_{t-1}^k, p_t^k) and (p_t^k, p_{t+1}^k). Vector (p_{t-1}^k, p_t^k) is given by a vector from location p_{t-1}^k to p_t^k. Considering that variable θ_t^k is the included angle between vectors (p_{t-1}^k, p_t^k) and (p_t^k, p_{t+1}^k), the direction change measure s_t^k is defined as:

$$s_t^k = \log_{\pi+1}(\theta_t^k + 1),\ 2 \le t \le N-1 \tag{2}$$

According to the above description, the direction change measure is applied to determine which parameter set is chosen as the animation frame, and the others are discarded. In the selection algorithm, $P=\{p_1, p_2, \ldots, p_N\}$ is assumed to be the parameter set sequence of a synthesis unit in which the parameter set pm is selected into the animation frame sequence $F=\{f_1, f_2, \ldots, f_M\}$, where $N > M$. The corresponding direction change measure of pm, which is given as s_t, is the product of all s_t^k. The algorithm for parameter set selection based on the direction change measure is described as follows: (1) Determine all direction change measures of p_t in P; (2) While the number of remaining parameter sets animation frames is greater than M, remove parameter set p_t with the smallest s_t from P; (3) Assign the remaining parameter sets to the animation frames.

IV. Apparent Motion Analysis

In Mandarin speech, the mouth shape of a consonant must be retained because of its occlusive, plosive or fricative characteristics. However, for the vowels, only some diphthongs with occlusive property need to keep the mouth shape. Therefore, the diphthong vowels with occlusive property and the consonants are grouped as class U1, and the other vowels are grouped as class U2.

U1 = {an, en, yan, yin, wan, wen, yuan, yun} ∪ {consonants}

U2 = {All the other vowels}

As described above, the transition component is characterized using the standard Bernstein-Bézier curve. The Bernstein-Bézier curve with n+1 control points is defined as:

$$\mathrm{C}(t) = \sum_{i=0}^{n} C_i B_{i,n}(t), 0 \le t \le 1, B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i} \tag{3}$$

where $C=\{C_0, C_1, \ldots, C_n\}$ denotes the control point set, and $B_{i,n}(t)$ denotes the Bernstein polynomial. This study adopts a four-control-point (n=3, $C=\{C_0, C_1, C_2, C_3\}$) Bernstein-Bézier curve. A two-step process is applied to locate the control points. The first step defines the control points C_0 and C_3. The control point C_0 is defined according to the class of the ending of the first syllable, and the control point C_3 is defined according to the class that of the start of the second syllable. Assuming that $P^{t-1} = \{P_0^{t-1}, P_1^{t-1}, ..., P_{NP(t-1)}^{t-1}\}$ and $P^t = \{P_0^t, P_1^t, ..., P_{NP(t)}^t\}$ are two consecutive parameter set sequences, and $NP(t)$ denotes the number of parameters, the control points C_0 and C_3 are manually defined as:

$$C_0 = P_{k_1}^{t-1}, \text{where } k_1 = \begin{cases} \left\lfloor \frac{5}{6} NP(t-1) \right\rfloor & \text{if } P^{t-1} \text{ ends with a unit in U1} \\ \left\lfloor \frac{2}{3} NP(t-1) \right\rfloor & \text{if } P^{t-1} \text{ ends with a unit in U2} \end{cases} \tag{4}$$

$$C_3 = P_{k_2}^{t}, \text{where } k_2 = \begin{cases} 0 & \text{if } P^{t} \text{ starts with a unit in U1} \\ \left\lfloor \frac{1}{6} NP(t) \right\rfloor & \text{if } P^{t} \text{ starts with a unit in U2} \end{cases} \tag{5}$$

The position of the control points C_1 and C_2 is calculated in the next step, and depends on psychophysical analysis. Computer animation, unlike real motion, is generated using alternate frames with little spatial variation in a very short duration. At an appropriate frame rate and a little spatial variation, the alternate frame sequence can be perceived as a real motion by the human visual system, which is called the "apparent motion". From previous investigation of apparent motion in psychophysics, the relationship between spatial variation and frame rate can be reduced, and can be used to avoid the discontinuity problem in computer animation.

Apparent motion is typically investigated by modulating the grating patch. A stimulus moves under a constant velocity, while a fixed grating patch overlaps the stimulus. The stimulus only appears when it passes through the gap of the grating patch. Since the step width, the distance between two gaps in the grating patch, is fixed, the stimulus appears in a fixed spatial and temporal frequency. Given that δ indicates the distance between two gaps, f_x, f_t, and v denote the spatial frequency, temporal frequency and velocity of the stimulus, respectively. The relationship between these variables can be given as

$$f_x = \frac{1}{\delta}, f_t = v \cdot f_x \tag{6}$$

Under a fixed velocity v, f_x and f_t increase when the step width δ decreases. The discontinuous motion of the stimulus is perceived as continuous as f_x and f_t rise above a threshold. From the definition, the term "temporal frequency" equals the term "frame rate".

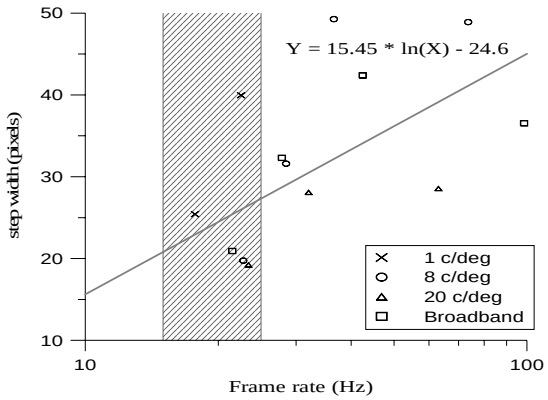

Figure 3. The relationship between frame rate and step width. The area with oblique lines is the region from 15 to 25 frames/sec.

Fig. 3 summarizes the relationship between frame rate and step width as determined in this study. In this figure, to obtain the specification used in animation, the original features, spatial frequency (*c*/deg) and velocity (deg/sec) are converted to frame rate (frames/sec) and speed (pixels/sec), respectively. The data denoted by cross symbol are ignored due to the low spatial frequency (1 *c*/deg). The area with oblique lines is the region from 15 to 25 frames/sec. A logarithmic curve is then fitted to describe the function:

$$D(F) = 15.45 \times \ln(F) - 24.6 \tag{7}$$

Because the variation of frame rate in this system is in the range 15–25 frames/sec, the appropriate step width is 17.2–25.1 pixels, assuming that the user is one meter away from the monitor. Therefore, the position of control points [formula] and [formula] is determined according to the limitation in step width:

$$C_1 = P_{k_1}^{t-1} + d \cdot \bar{u}\left(P_{k_1-1}^{t-1}, P_{k_1}^{t-1}\right), C_2 = P_{k_2}^{t} + d \cdot \bar{u}\left(P_{k_2+1}^{t}, P_{k_2}^{t}\right) \tag{8}$$

where $\bar{u}(p_1, p_2)$ denotes the unit vector from point p_1 to p_2, and the variable d is determined such that the slope of vector $\overrightarrow{C_1C_2}$ fits the step width restriction. The slope of vector $\overrightarrow{C_1C_2}$ is determined as

$$s = \frac{y_A + d \cdot y_B}{x_A + d \cdot x_B}, \text{where} \begin{cases} x_A = x_{k_2}^{t} - x_{k_1}^{t-1} \\ x_B = x_{k_2}^{t} + x_{k_1-1}^{t-1} - x_{k_2+1}^{t} - x_{k_1}^{t-1} \\ y_A = y_{k_2}^{t} - y_{k_1}^{t-1} \\ y_B = y_{k_2}^{t} + y_{k_1-1}^{t-1} - y_{k_2+1}^{t} - y_{k_1}^{t-1} \end{cases} \tag{9}$$

Accordingly, under the constraint that the slope should be inside the range 17.2–25.1 pixels/frame, d can be represented as

$$\frac{17.2x_A - y_A}{y_B - 17.2x_B} \le d \le \frac{25.1x_A - y_A}{y_B - 25.1x_B} \tag{10}$$

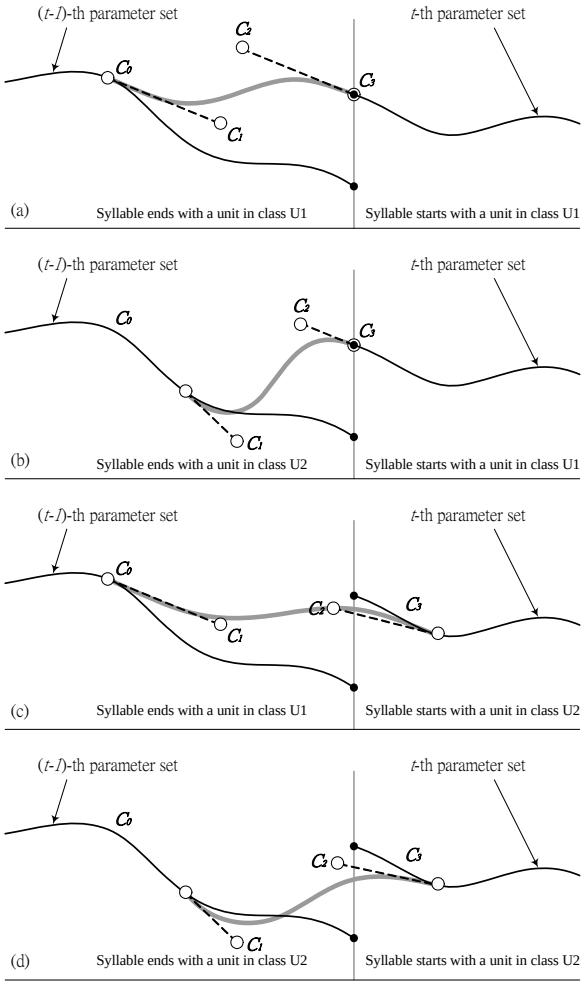

Figure 4. The four types of transitions generated by a four-point Bernstein-Bézier curve.

Fig. 4 illustrates four types of transitions simulated by the Bernstein-Bézier curve for the four phone combinations, which are the combinations of two phones belonging to U1–U1, U1–U2, U2–U1 and U2–U2. For example, Fig. 4(a) illustrates the case for two concatenated syllables where the first syllable ends with a unit in class U1 and the second syllable begins with a unit in class U2.

V. Simulation Results

To evaluate the performance of the proposed approach, a visual speech synthesis system was implemented on a personal computer with Pentium IV CPU, 512 MB memory and a general display card. For the real-time requirement, the system can produce the output animation in real-time with 25 frames per second. Fig. 5 displays some image sequences of synthesis results. Two synthesis results with 25FPS and 15FPS frame rates are displayed in the second and third rows, respectively. The image sequence of real speaker is also displayed in the first row. The upper part of the second and the third row displays the synthesis result using Maximum Direction Change algorithm, while the lower part displays

the result without using Maximum Direction Change algorithm.

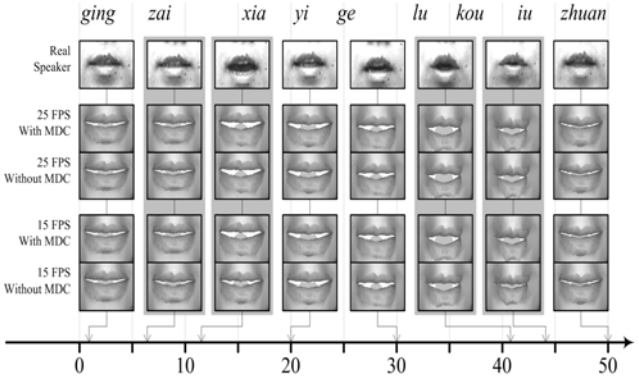

Figure 5. Image sequences of synthesis result.

A. Effect of the selection based on Maximum Direction Change

This experiment attempted to reveal the benefit of applying the direction change measure in the synchronization process. The experimental result is shown by the comparison of the rotation angle of chin, which can be estimated from the movement of the feature points below the mouth. Fig. 6 compares the parameter change contours using the same sentences.

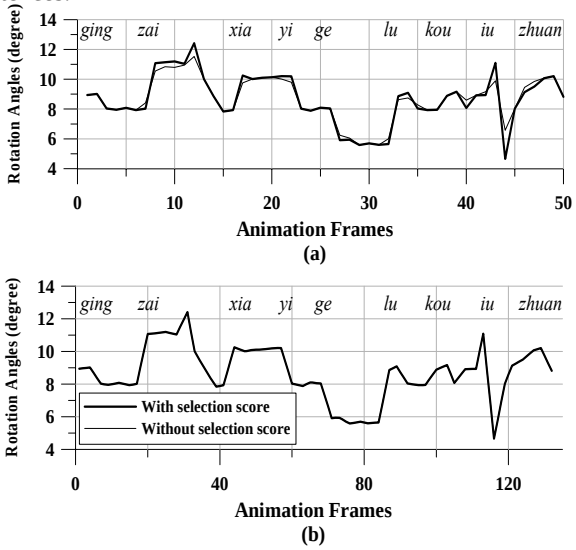

Figure 6. Comparison of parameter set sequences with and without maximum direction changes, respectively. The contours show the rotation angles for the synthesized result animated in (a) a normal speaking rate and (b) a low speaking rate. In (b), the two contours are the same and overlap completely.

Since the benefit of direction change measure is significant when the speaking rate is high, the error was compared in two different speaking rates. The maximum number of parameter sets in the previously defined parameter set sequence was set to 13, i.e., about 0.5 seconds per syllable. Hence, if the speaking rate is below 2 syllables per second, the effect of direction change measure is unobvious. The speaking rate is typically about 4 to 5 syllables per second, and the minimum number of parameter sets in the parameter set sequence is 4. Therefore, the measurement of direction change plays an important role in realistic synthesis output in most situations. Fig. 6(a) depicts lip animation generated with a general speaking rate, i.e., about 4 to 5 syllables per second. The result demonstrates that the difference between two curves is significant in some tips. Fig. 6(b) depicts the results of slowing the speaking rate down to about 0.67 syllables per second. Due to the slow speaking rate, the two curves overlap each other, and are almost the same.

The visualized synthesis result of this sentence is presented in Fig. 5. The images in the same column represent the synthesis results at the same temporal position, which are shown by the arrow lines. The temporal positions are chosen to reveal the benefit of using Maximum Direction Change algorithm. The images in the first, forth, fifth, and eighth columns show that there is no difference between synthesis results with 25FPS and 15FPS, while the images in the second, third, sixth, and seventh columns (columns with gray backgrounds) show that the synthesis result under 15FPS without using Maximum Direction Change algorithm cannot perform the lip shape as precise as the other methods.

B. Generation of transition component

This experiment evaluates the generation of the coarticulation part, especially for the application of apparent motion and Bernstein-Bézier curve. As in the previous experiment, the result is shown by the contours of mouth height in the transition component.

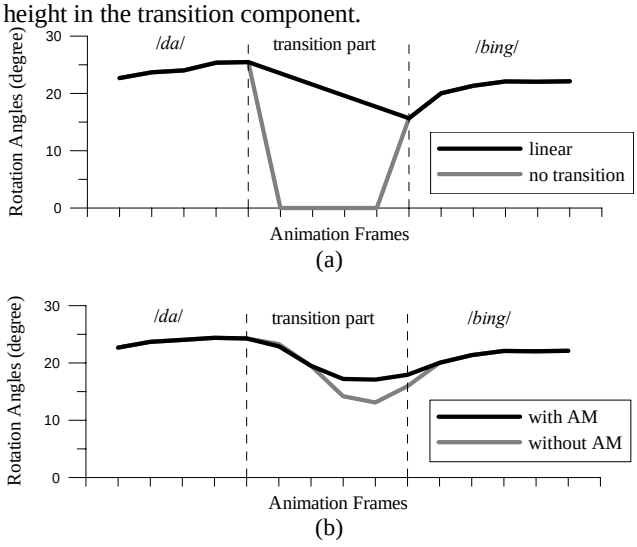

Figure 7. Comparison of parameter set sequences with different transition generation approaches. The transition part is generated using (a) conventional linear interpolation and (b) the Bernstein-Bézier curve with (black curve) and without (gray curve) apparent motion (AM).

Fig. 7 illustrates the contours for parameter change, which show the transition part between two concatenated synthesis units. Fig. 7(a) illustrates the contour for parameter change with standard liner interpolation, and clearly shows an abrupt change in Fig. 7(a). Fig. 7(b) illustrates the

synthesis results after applying the Bernstein-Bézier curve. The gray line in Fig. 7(b) illustrates the transition contour without the constraint of apparent motion. However, the slope of the contour in the middle of the transition is too sharp to generate a suitable apparent motion. Finally, applying both the Bernstein-Bézier curve and the apparent motion constraint to generate the transition part significantly improves the smoothness, as demonstrated in the black contour in Fig. 7(b).

To clarify the result of the proposed method, Table 3 lists the numerical comparison between the proposed approach and other baseline approaches. The numerical comparison of the proposed method and real speaker is also displayed here. Score 1 and Score 2 in Table 3 denote the averages of the normalized rational angles and rational angle offsets, respectively. To compare the difference of these methods, the test is repeated under different frame rates, and the parameters of real speaker are linearly interpolated directly from the captured data sequence. The final result reveals that the proposed approach can produce the most similar scores to the real speaker under both 25FPS and 15FPS.

TABLE III. COMPARISON BETWEEN DIFFERENT COARTICULATION GENERATIONS

Articulation Model	Score 1		Score 2	
	25 FPS	*15 FPS*	*25 FPS*	*15 FPS*
Real Speaker	0.5199	0.5199	-0.0366	-0.0366
Linear interpolation	0.2144	0.2012	-0.1290	-0.1107
Bézier curve without AM	0.5230	0.5210	-0.0433	-0.0441
Bézier curve with AM	0.5193	0.5191	-0.0329	-0.0325

VI. CONCLUSION

This study proposes automated lip synchronization and smoothing for Chinese visual speech synthesis. According to the characteristics of Mandarin Chinese, 105 syllable-based synthesis units are defined to represent the lip motions of Mandarin speech. The parameter set sequences for each synthesis unit are estimated and applied for animation. The direction change measures for each parameter sets are estimated to preserve the significant frames in lip motion. In the generation of animation output at a high speaking rate, the direction change measure is employed to determine the mapping between parameter sets and the animation frames. Finally, the "apparent motion" property based on psychophysics is adopted to constrain the slope of Bernstein-Bézier curve and generate a realistic transition between two consecutive synthesis units. Restraining the slope of Bernstein-Bézier curve causes the two different parameter set sequences to be smoothly concatenated.

Some problems still remain unresolved in these applications. For instance, the selection algorithm tends to select parameter sets with larger direction change measures, regardless of the duration information. In some extreme cases, when the speaking rate is too fast, the algorithm discards all the parameter sets without direction change, causing the synthesis to be unconnected. To solve this problem, the estimation of direction change measure has to involve the duration information.

REFERENCES

[1] N. Magnenat-Thalmann, N. E. Primeau, D. Thalmann, "Abstract Muscle Actions Pro-cedures for Human Face Animation," *Visual Computer*, Vol. 3(5), 1988, pp. 290-297.

[2] F.D. Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, B.D. Carolis, "From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent," *International Journal of Human-Computer Studies*, Vol. 59, 2003, pp. 81-118.

[3] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, K. Alvarez, "Developing multimodal in-telligent affective interfaces for tele-home health care," International Journal of Hu-man-Computer Studies, Vol. 59, 2003, pp. 245-255.

[4] T. Takahashi, C. Bartneck, Y. Katagiri, N.H. Arai, "TelMeA-Expressive avatars in asynchronous," *International Journal of Human-Computer Studies*, Vol. 62, 2005, pp. 193-209.

[5] G. Breton, C. Bouville, D. Pelé, "FaceEngine: A 3D Facial Animation Engine of Real Time Applications," in: *Proceedings of ACM SIGWEB 2001*, 2001, pp. 15-22.

[6] J.D.R. Wey, J.A. Zuffo, "InterFace: a Real Time Facial Animation System," in *Proceedings of International Symposium on Computer Graphics, Image Processing, and Vision. Rio de Janeiro*, 1998, pp. 200-207.

[7] J. Piesk, G. Trogemann, "Animated Interactive Fiction: Storytelling by a Conversa-tional Virtual Actor," in *Proceedings of Virtual Systems and MultiMedia'97*, 1997, pp. 100-108.

[8] G. Bailly, M. Bérar, F. Elisei, M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, Vol. 6, 2003, p. 331-346.

[9] T. Okadome, T. Kaburagi, M. Honda, "Articulatory movement formation by kinematic triphone model," in *Proceedings of IEEE International Conference on Systems Man and Cybernetics*, 1999, pp. 469-474.

[10] M. Tamura, S. Kondo, T. Masuko, T. Kobayashi, "Text-to-audiovisual speech synthe-sis based on parameter generation from HMM," in *Proceedings of European Confer-ence on Speech Communication and Technology*, 1999, pp. 959-962.

[11] T. Kuratate, H. Yehia, E. Vatikiotis-Bateson, "Kinematics-based synthesis of realistic talking faces," in *Proceedings of Auditory-Visual Speech Processing Conference*, 1998, pp. 185-190.

[12] T.F. Cootes, G.J. Edwards, C.J. Taylor, "Active Appearance Models," in *Proceedings of European Conference on Computer Vision 1998*, 1998, pp. 484-498.

[13] L.S. Chen, T.S. Huang, J. Ostermann, "Animated talking head with personalized 3D head model," in *Proceedings of IEEE First Workshop on Multimedia Signal*, 1997, pp. 274-279.

[14] S. Shan, W. Gao, J. Yan, H. Zhang, X. Chen, "Individual 3D face synthesis based on orthogonal photos and speech driven facial animation," in *Proceedings of IEEE In-ternational Conference on Image Processing 2000*, 2000, pp. 238-241.

[15] E. Cosatto, H.P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Transaction on Multimedia*, Vol. 2(3), 2000, pp. 152-163.

[16] T. Guiard-Marigny, A. Adjoudani, C. Benoit, "3D Models of the Lips and Jaw for Visual Speech Synthesis," in *Proceedings of Santen, J. et al. (Ed.), Progress in Speech Synthesis*, 1996, pp. 247-258.

[17] J. Kleiser, "A fast, efficient, accurate way to represent the human face," in *Proceed-ings of ACM RAPH '89 Course Notes 22: State of the Art in Facial Animation*, 1989, pp. 37-40.

[18] M.M. Cohen, D.W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech," in *Proceedings of Thalmann*, 1993, pp. 139-156.

[19] R.H. Bartels, J.C. Beatty, B.A. Barsky, *An Introduction to Splines for Use in Com-puter Graphics and Geometric Modelling*, Morgan Kaufmann Publishers Inc., San Francisco, CA , 1998.

[20] Y. Zhuo, T.G. Zhou, H.Y. Rao, J.J. Wang, M. Meng, M. Chen, C. Zhou, L. Chen, "Contributions of the visual ventral pathway to long-range apparent motion," *Science*, Vol. 299, 2003, pp. 417-420.