基于数据驱动方法的汉语文本-可视语音合成*

王志明[1+], 蔡莲红[2], 艾海舟[2]

[1](北京科技大学 计算机科学与技术系,北京 100083)

[2](清华大学 计算机科学与技术系,北京 100084)

Text-To-Visual Speech in Chinese Based on Data-Driven Approach

WANG Zhi-Ming[1+], CAI Lian-Hong[2], AI Hai-Zhou[2]

[1](Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing 100083, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62330710, E-mail: wangzhiming@tsinghua.org.cn

Received 2004-05-17; Accepted 2004-09-08

Wang ZM, Cai LH, AI HZ. Text-To-Visual speech in Chinese based on data-driven approach. *Journal of Software*, 2005,16(6):1054−1063. DOI: 10.1360/jos161054

Abstract: Text-To-Visual speech (TTVS) synthesis by computer can increase the speech intelligibility and make the human-computer interaction interfaces more friendly. This paper describes a Chinese text-to-visual speech synthesis system based on data-driven (sample based) approach, which is realized by short video segments concatenation. An effective method to construct two visual confusion trees for Chinese initials and finals is developed. A co-articulation model based on visual distance and hardness factor is proposed, which can be used in the recording corpus sentence selection in analysis phase and the unit selection in synthesis phase. The obvious difference between boundary images of the concatenation video segments is smoothed by image morphing technique. By combining with the acoustic Text-To-Speech (TTS) synthesis, a Chinese text-to-visual speech synthesis system is realized.

Key words: text-to-speech (TTS); text-to-visual speech (TTVS); viseme; co-articulation

摘 要: 计算机文本-可视语音合成系统(TTVS)可以增强语音的可懂度,并使人机交互界面变得更为友好.给出一个基于数据驱动方法(基于样本方法)的汉语文本-可视语音合成系统,通过将小段视频拼接生成新的可视语音.给出一种构造汉语声韵母视觉混淆树的有效方法,并提出了一个基于视觉混淆树和硬度因子的协同发音模型,模型可用

* Supported by the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20010003049 (国家教育部博士点基金); the Scholastic Science Foundation of University of Science and Technology Beijing under Grant No.20040509190 (北京科技大学校内科研基金)

WANG Zhi-Ming was born in 1968. He is a lecturer at the Department of Computer Science and Technology, University of Science and Technology Beijing. His current research areas are visual speech synthesis and image processing. **CAI Lian-Hong** was born in 1945. He is a professor and doctoral supervisor at the Department of Computer Science and Technology, Tsinghua University. His research areas are speech processing and synthesis, biometric recognition. **AI Hai-Zhou** was born in 1964. He is a professor and doctoral supervisor at the Department of Computer Science and Technology, Tsinghua University. His research areas are computer vision and pattern recognition.

于分析阶段的语料库选取和合成阶段的基元选取.对于拼接边界处两帧图像的明显差别,采用图像变形技术进行平滑并.结合已有的文本-语音合成系统(TTS),实现了一个中文文本-视觉语音合成系统.

关键词: 文-语转换系统(TTS);文本-可视语音合成系统(TTVS);视位;协同发音

中图法分类号: TP18 文献标识码: A

1 Introduction

Visual speech means the movements of visible articulatory organs of the speaker, such as lips, tongue, jaw, facial muscles, etc. Both audio speech and visual speech are produced by the movements of articulatory organs, so there is an inherent relationship between them. Visual Speech synthesis by computer can increase the speech intelligibility, whether in a clean or noisy environment. In noise background, visual speech is equivalent to increase 11dB acoustic signal-to-noise ratio[1]. In human computer interaction, text-to-visual speech synthesis can make the human computer interface (HCI) more friendly, especially for hearing-impaired people.

During about 30 years of visual speech synthesis, two different synthesis approaches have been developed. One is the parameter control approach or we can call it model based approach[2−6], and the other is the data driven approach or we can call it sample based approach[7−10]. In the first approach, we need to build a 2D or 3D face model first, then define the visual parameters for the key frames or every viseme, and create these parameters for every video frame. At last, we use these control parameters to drive the face model for visual speech synthesis. The advantages of parametric approach include the easily changing face model, small database size, and universal parameters for different models. The disadvantages include that the synthesized result is always with some artificial effect, and its high computation complexity makes it hard for this approach to simulate the dynamic characteristic of visual parameter.

In the data driven approach, we need to build an image sample database first, select the proper image samples from database based on some energy function, and then carefully concatenate these images together to produce a new visual speech. Data driven approach gives a more realistic result, and it is much like a particular individual. The shortages of the data driven approach include the requirement of building a huge sample database, and the fact that the synthesis quality is proportional to the size of the database. Consequently, if we want to change the face model from one person to another, we must rebuild the entire sample database for that person, which is extremely tedious.

More recently, some people try to combine the advantages of parameter control approach and data driven approach[11−13]. Face model is animated by control parameters, but these parameters are obtained by the data-driven approach. These approaches could preserve the dynamic properties of face movement, but still need a realistic face model and a good animation algorithm.

Chinese language is syllable-based, which is quite different from phone-based western languages. Phonemes in Chinese language include initials and finals, and they have clearly different roles in pronunciation. Some Chinese text-to-visual speech synthesis systems have been built in recent years[14−18]. But up to now, all of them are built with model based approaches. In this paper we present a Chinese text-to-visual speech based on data-driven approach. First we construct two visual confusion trees for Chinese initials and finals respectively. Then we give each phoneme a hardness factor and propose a co-articulation model for data-driven visual speech synthesis. Third, we optimize the unit selection cost function and design the recording corpus according to our co-articulation model. Finally, by combining with our acoustic text-to-speech synthesis, we realize a Chinese text-to-visual speech synthesis system.

This paper is organized as follows. Section 2 describes how to estimate the viseme parameters and how to

construct the visual confusion tree for Chinese phonemes. Section 3 gives a co-articulation model based on visual distance and a hardness factor for corpus design and unit selection. Section 4 gives the framework of text-to-visual speech synthesis system based on data-driven approach. In Section 5, we give some experimental results, and finally the conclusion is given in Section 6.

2 Visual Confusion Tree

In order to find the similarity between every phoneme to another, we estimate MPEG-4 defined Facial Animation Parameters (FAPs) for all Chinese phonemes. Considering the different role with initials and finals, we measure parameter distances and construct visual confusion tree within each group.

2.1 Estimation of FAPs from orthogonal view

A viseme is a visual correlate to a phoneme and defined in MPEG-4 as follows: Viseme is the physical (visual) configuration of the mouth, tongue and jaw that is visually correlated with the speech sound corresponding to a phoneme[19]. MPEG-4 has also defined 68 Facial Animation Parameters (FAP), which can be used to describe almost any facial expression in our daily life. We select 28 parameters from MPEG-4 defined FAPs (FAP#3~14, FAP#16~17, FAP#44~47, FAP#51~60) to describe Chinese visemes[18].

To estimate FAPs for Chinese visemes, we extract Chinese static visemes for initials and finals according to their roles in syllable[20]. Because the initial phoneme is always in the beginning of a syllable and has very short time duration, we extract the static viseme for initial on the beginning of the acoustic speech. On the other hand, final phoneme has the time duration almost as long as the whole syllable and has a steady state in the middle of the pronunciation, so we extract the static viseme for the final on the middle of acoustic speech.

In visual speech analysis and synthesis, some people use 2D parameters, ignore the 3D information[21,22]; some people extract 3D viseme parameters by 3D face model, but constructing a 3D face model by multi-camera technique has many unsolved problems, such as the synchronous problem and the difficulty to guarantee reconstruction precision[23,24]. In order to estimate 3D FAPs from video and avoid the difficult 3D re-construction, we use a mirror to acquire two orthogonal frontal and profile views simultaneously[25]. The 3D FAPs are estimated from the tracking results of feature points in both views. In frontal views, the nostrils, two points in glasses, are tracked as reference points to estimate the face pose. Outer lip contour is tracked by deformable template and the inner lip parameters are estimated based on the outer lip parameters. Combing the statistic learning method and rule based method, precise tracking results are obtained for mouth contour and facial feature points based on facial color probability distribution and priori knowledge on shape and edge. The high frequency noise in reference points tracking is eliminated by low-pass filter, and the main face pose is estimated from four reference points to remove the overall movements of the face. In the profile view, points represent the nose tip; the protrusion of upper and lower lip, the thrust of jaw, and the openness of jaw are to be located. The correlative FAPs are estimated from the positions of these feature points. More details could be found in Ref.[23]. Four inner mouth parameters (FAP#44~47) are estimated by watching X-ray speaking video and phonetics knowledge. Figure 1 shows a typical tracking result.

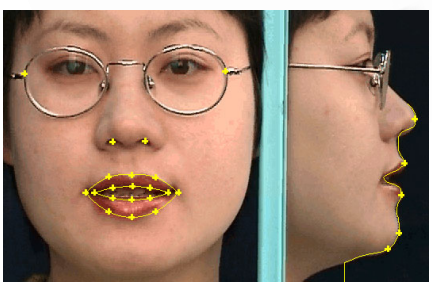


Fig.1 FAP extraction from orthogonal views

2.2 Construction of visual confusion tree

The difference of every viseme to other viseme is represented by the square of Euclidian distance between their FAPs:

$$d_{i,j}=\sum_{n}(Fap(i,n)-Fap(j,n))^2 \tag{1}$$

where $Fap(i,n)$ is the nth FAP value for viseme i.

For different person, the FAP values may have some difference, but the distance between corresponding viseme should be similar. So we normalize FAPs from different person by the maximum distance between all pairs of visemes, and the normalized distance becomes:

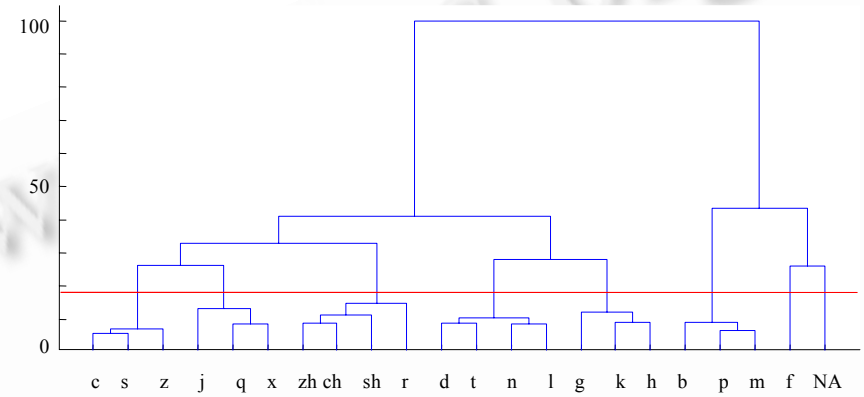
$$D_{i,j}=d_{i,j}/\mathrm{Max}(d) \tag{2}$$

Visual confusion tree is constructed as follows. First, every phoneme is defined as the leaf of the tree. Then the two elements that results in the least increase of total square error are merged. The error increase of merging branch i and j is defined as:

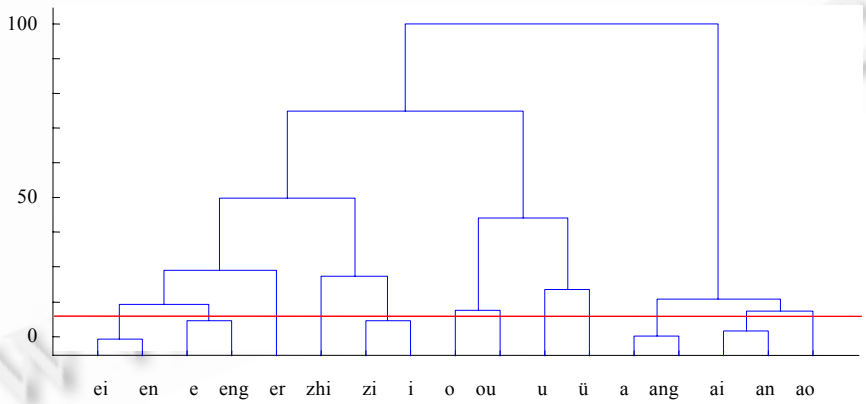
$$Err_{i,j}=n_i n_j D_{i,j}/(n_i+n_j) \tag{3}$$

The merging procedure is repeated until all of the branches are merged to only one tree trunk.

The resulting visual confusion tree for the initial visemes and final visemes are show in Fig.2.



(a) Initial



(b) Final

Fig.2 Visual confusion tree of Chinese phoneme

3 Co-Articulation Model Based on Visual Distance and Hardness Factor

The visual confusion tree is constructed as follows. First, every phoneme is treated as the leaf of the tree. Then the two elements that results in the least increase of total square error are merged. The error increase of merging branch i and j is defined as:

In data-driven visual speech synthesis, the co-articulation model is different from that used in model based visual speech[2]. Because in data-driven approach, we must select proper image samples from sample database, so we cannot change the parameters continuously. Cosatto[9] gives a co-articulation model based on video frame, but it is hard to decide the frame length of co-articulation for every phoneme, and the large sample size makes the unit selection very slow. Huang[26] gives a co-articulation model based on tri-phone, but some times the co-articulation effect runs longer than tri-phone, especially in Chinese. Figure 3 shows two different mouth shapes for final /e/ in the same tri-phone environment. This is because they have different preceding and upcoming finals out of the tri-phone.

(a) wu-de-xiu (b) da-de-xiong

Fig.3 Different mouth shapes in the same triphone

Kshirsagar[11] solves the co-articulation problem by using visual-syllable, phonemes are classified by phonetic knowledge and 900 'demi-visyllables' are used to cover various co-articulation phenomena. The limited number of visual-syllable inevitably results in some un-match in the concatenated boundary. Edge's model[12] selectes various length units by similarity of phonetic timing and context, and parameter distance is smoothed by weighted blending. Cao[13] gives a co-articulation model based on 'Anime Graph' search which could find the most smoothing trajectory of visual parameters. All of these approaches treat vowels and constants equally, and rely too much on visual parameters of adjacent phonemes.

In every Chinese syllable, there is usually an initial and a final, but some times there is no initial. In our concatenating visual speech synthesis, we use syllable video as a concatenated unit. But if we use syllable as co-articulation model element, the number of co-articulation unit would be unacceptable. This is because there is about 417 syllables in Chinese language, even for a tri-phone model, the number of tri-phone would be 4173.

Be aware of that the main duration in a syllable is covered by the final, we use the final as a selection unit and use the phoneme as a co-articulation element. Every final is affected by two to four phonemes: the preceding final, preceding initial (if exist), upcoming initial (if exist), and upcoming final. Visemes for some phoneme are easy to be affected by adjacent visemes (such as /d/, /g/, /j/), but visemes for some phoneme are nearly not affected by adjacent visemes (such as /b/, /p/, /f/). We give each viseme a hardness factor between 0 and 1. The larger the factor is, the less it is affected by others.

The whole co-articulation model is given by comparing the visual similarity of two syllables in the co-articulation environment:

$$VD_{i,j} = H_{-2} \cdot (1 - H_{-1}) \cdot D_{i-2,j-2} + H_{-1} \cdot D_{i-1,j-1} + H_{+1} \cdot D_{i+1,j+1} + H_{+2} \cdot (1 - H_{+1}) \cdot D_{i+2,j+2} \quad (4)$$

where $VD_{i,j}$ is the visual distance from any selecting syllable j to target syllable i (We force two syllable have the same final); H_t=Max($H_{t,i}$,$H_{t,j}$), t={−2,−1,+1,+2}, gives the maximum hardness factor of i,j for preceding final, preceding initial, upcoming initial and upcoming final respectively ($0 \le H_{t,i}$, $H_{t,j} \le 1$); $D_{i-2,j-2}$,$D_{i-1,j-1}$, $D_{i+1,j+1}$, $D_{i+2,j+2}$ gives the visual distance preceding final, preceding initial, upcoming initial and upcoming final in the target environment and selecting environment respectively. The distance between composite finals (such as /ia/ and /iou/) is defined as the largest distance between all possible pairs of finals.

According to Eq.(4), we can precisely compare the visual difference from one syllable to another. We can select the optimal corpus text in the analysis phase and the optimal synthesis unit in the synthesis phase.

4 Framework of Data-Driven TTVS

An effective synthesis approach is always based on analysis. So the framework of our data-driven text-to-visual speech synthesis contains two phases: analysis and synthesis, which is shown in Fig.4. In the analysis phase, we

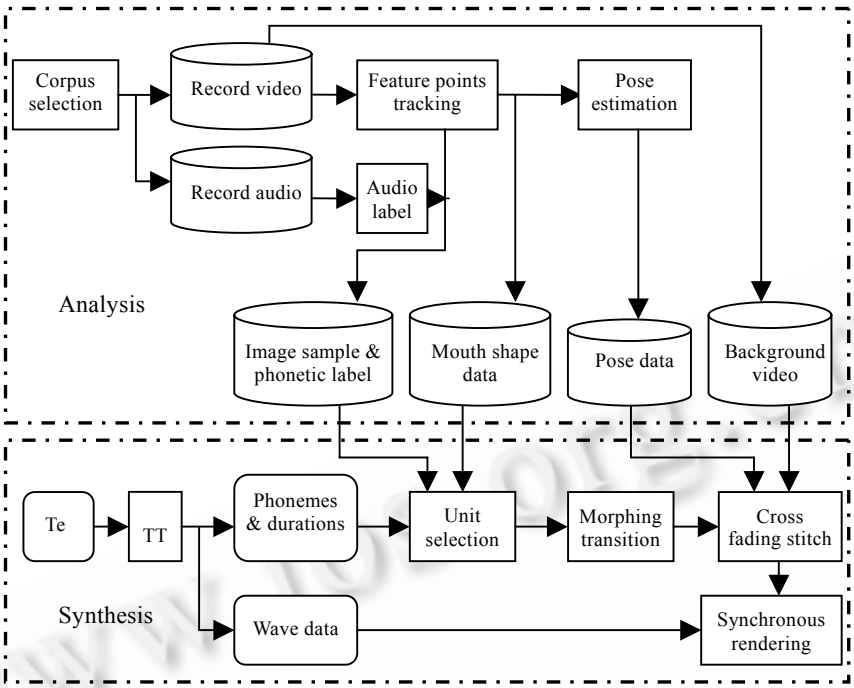


Fig.4 Text-To-Visual speech framework

need to select the proper corpus text, record video, and audio speech. After facial feature points tracking and audio labeling, we extract the mouth area images and save them in database. At the same time, we estimate the head pose and mouth shape parameters from these feature points and save them too in database. And then, we extract some video as background video. In the synthesis phase, with a given text, we synthesize the audio wave data by our Chinese TTS and get phonemes and their durations from TTS. By using the phonetic label and mouth shape data from database and our co-articulation model, we select the proper mouth image samples from image database. If the mouth shape parameters between two concatenated frames are large, we smooth the transition by the image morphing technique, and then stitch these mouth images with background video by cross fading technique according to pose data. At last, the image sequence and the wave data are rendered synchronously.

In Section 2, we have mentioned the problem of facial feature points tracking. The pose estimation algorithm we use is similar to Cosatto's in Ref.[10] that uses four eye corners and two nostril points to estimate the head pose. To keep the audio label as precisely as possible, the audio label is currently done manually in our system. So in the analysis phase we only discuss the question of corpus selection.

In the synthesis, Viterbi search is used for unit selection on the whole sentence as mentioned by Huang in Ref.[26]. The difference is that our selection of the unit is based on the co-articulation model presented in Section 3 instead of tri-phone in Ref.[26]. The cross fading stitch technique could be found in Ref.[7], and here we will only discuss the problem of morphing transition on boundary images.

4.1 Corpus selection

As mentioned in Ref.[9], in data-driven visual speech synthesis, the size of the database is directly related to the quality required for the animation. In order to make the database as small as possible under the given quality, we must select the record corpus carefully. According to the co-articulation model, we need to cover as many co-articulation environments possible. We select visual speech corpus from our TTS corpus, which contains more than 5,000 sentences and about 50,000 syllables. For every sentence under selection, we select the nearest syllable

in the current corpus according to our co-articulation model, and compute the average visual distance over all syllables. In every cycle, we add the sentence with the largest average visual distance into the current corpus until the given minimum distance is reached or maximum sentence or syllable number is reached. The select criterion is given as follows:

$$Cost_i = \frac{1}{n_i}\sum_{k=1}^{n_i} VD_{k,k'} \tag{5}$$

$$Add = \arg\max(Cost_i) \tag{6}$$

where n_i gives the syllable number in sentence i; k' gives the nearest syllable for syllable k in the current corpus, and Add gives the sentence number added to corpus in one cycle.

4.2 Morphing transition

Due to the limit of samples in image database, some times the frames to be concatenated may have obvious difference in shape and texture. This will result in some jerk in synthetic video. We smooth this kind of difference by image morphing technique.

Let A, B denote the two image frames to be concatenated, S_A, S_B denote their shapes, then they can be defined by the vector of positions of all feature points. Let $\alpha \in (0,0.5)$ denote the smooth factor, $S_{A'}$, $S_{B'}$ denote the shapes after smoothing, then we have:

$$S'_A = S_A \cdot (1-\alpha) + S_B \cdot \alpha \tag{7}$$

$$S'_B = S_A \cdot \alpha + S_B \cdot (1-\alpha) \tag{8}$$

After the shape smoothing, we need to smooth the texture between the concatenated boundary images. Let I_A, I_B denote the texture of two image frames to be concatenated, so $I_A(x,y)$ denotes the image data in point (x,y) of image A, then the smoothed image data can be calculated by:

$$I'_A(x,y) = I_A(x',y') \cdot (1-\alpha) + I_B(x'',y'') \cdot \alpha \tag{9}$$

where $(x',y') = T_{AA'} \cdot (x,y) + C_{AA'}$, $(x'',y'') = T_{AB'} \cdot (x,y) + C_{AB'}$.

$$I'_B(x,y) = I_B(x',y') \cdot (1-\alpha) + I_A(x'',y'') \cdot \alpha \tag{10}$$

where $(x',y') = T_{BB'} \cdot (x,y) + C_{BB'}$, $(x'',y'') = T_{BA'} \cdot (x,y) + C_{BA'}$.

$T_{AA'}$, $T_{AB'}$, $T_{BB'}$ and $T_{BA'}$ denote the mapping matrixes from S_A to $S_{A'}$ and $S_{B'}$, S_B to $S_{B'}$ and $S_{A'}$ for every triangle of the two images. These mapping matrixes can be calculated by solving six linear equations on the six vertexes of the two corresponding triangles.

Figure 5 gives the original two image sequences to be concatenated (up row) and the smooth results (low row). The transition between two sequences becomes more like each other after smoothing. The larger the smooth factor, the smoother the result, but the resulted images take more co-articulation effect, and look like none of the original

A B

Fig.5 Transition morphing (up: origin; low: smoothed; α=0.3)

images. In practice, the factor is usually between 0.2 and 0.3.

5 Experimental Results

In our experiments, we estimate 8 groups of the viseme data from 3 persons, and the visual confusion trees are built from those viseme data by the approach described in section 2.2. Figure 1 shows one of the facial feature point tracking results, and Fig.2 gives the final visual confusion trees of the Chinese phoneme of intial and final.

After building those visual confusion trees, visual distance between visemes could be found from the tree, and the co-articulation model could be built. The hardness of each viseme can be desired by phonetics knowledge, but it is more reasonable to acquire by experiment. One approach is to estimate the magnitude of the dominance function in Cohen's co-articulation model[2].

Figure 6 shows the selected 'e' viseme for sentence 'wo de shu (my book)' and its neighbor viseme. Up panel is selected by our co-articulation model, and low panel is selected by the common tri-phone model. Clearly, the viseme selected by our model is quite similar to the previous viseme, so it is more suitable for visual speech synthesis.

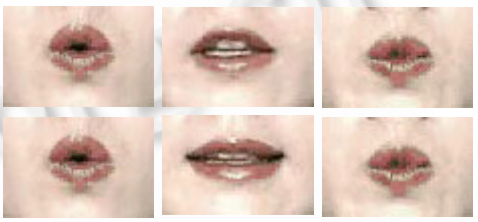


Fig.6 The selected 'e' viseme for sentence 'wo de shu (my book)'. up: our co-articulation model; low: common tri-phone model

Based on our co-articulation model, we select 100 sentences (576 syllables) form 5,000 sentences by the approach mentioned in subsection 4.1. These sentences are used to build the data-driven TTVS.

First, all of these sentences are recorded by a woman participant. Because the current synthesis TTVS is a 2D talking head, only the frontal view video is tracked in the experiment. Then all viseme images are extracted from the videos and the image database is built. In the synthesis phase, image samples are selected from the database according to the visual distance computed by our co-articulation model.

With 10 test sentences (total 383 frames), we compute the average relative mouth height and width differences between the synthesis images and the origin images (using natural height and width as reference). Table 1 shows the experimental results of our co-articulation model and the tri-phone model. Clearly, our model is better than the common tri-phone model.

Table 1 Relative mouth height and width differences between synthesis images and origin images

Co-articulation model	Height (%)	Width (%)
Common tri-phone model	8.15	5.22
Our model	4.87	3.89

Figure 7 shows the video of Chinese sentence 'ni hao (hello)'synthesis by our system. Another synthesis result could be found on http://media.cs.tsinghua.edu.cn/~wangzm.

6 Conclusions

In this paper, we present a Chinese text-to-visual speech system based on data-driven approach. First, we develop a method to build a visual confusion tree by FAPs estimated from facial feature point tracking. Based on the visual confusion tree, we propose a co-articulation model for data-driven visual speech synthesis. Our model can deal with the co-articulation effect better than the tri-phone model, in which the influence of preceding and upcoming phones on current phone is determined by a hardness factor.

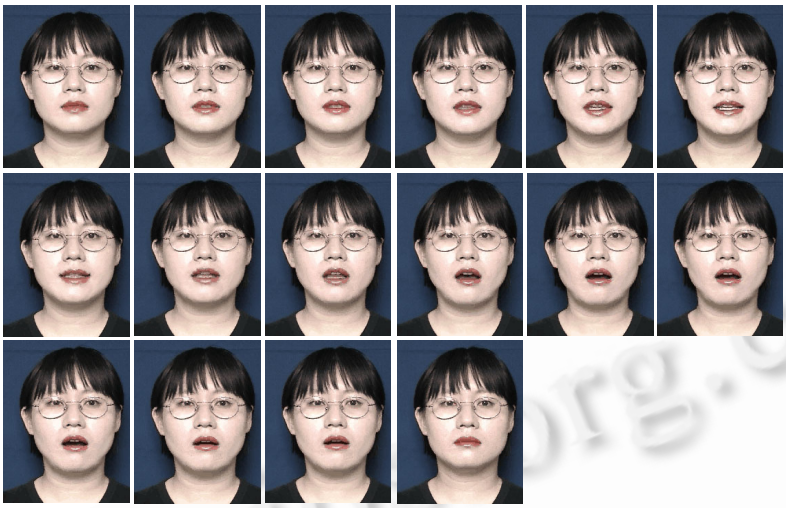


Fig.7 Synthesis bitmap sequences of sentence ‘ni hao (hello)’ by our system

In text-to-visual speech synthesis, we focus on two key problems: corpus selection and transition morphing. By selecting the corpus based on our co-articulation model, we minimize the unit selection match distance under the given corpus size. By triangle-based image morphing, we realize a smooth transition between different video segments.

Up to now, the data-driven approach has the most realistic visual speech synthesis result. It can be used to generate a virtual newscaster for movie dubbing or to help the hearing-impaired children learn language. In order to make the synthesis visual speech more like a real human speaker, we have to simulate visual prosody of human speech, which includes the patterns of head movements, gestures and body movements. Graf [27] has studied the relation between speech and head movement. But to achieve a good visual prosody prediction, further study on the prosodic structure of the text should be considered, which is a really tough question in text-to-speech system.

References:

[1] Macleod A, Summerfield AQ. Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology, 1987,21(2):131−141.

[2] Cohen MM, Massaro DW. Modeling coarticulation in synthetic visual speech. In: Thalmann NM, Thalmann D, eds. Models Techniques in Computer Animation. Tokyo: Springer-Verlag, 1993. 139−156.

[3] Waters K, Levergood T. DECface: An automatic lip-synchronization algorithm for synthetic faces. CRL Technical Report 93/4, Digital Equipment Corporation Cambridge Research Laboratory, 1993.

[4] Le Goff B, Benoit C. A text-to-audiovisual-speech synthesizer for French. In: Proc. of the 4th Int’l Conf. on Spoken Language Processing (Ⅳ). Philadelphia, 1996. 2163−2166.

[5] Masuko T, Kobayashi T, Tamura M, Masubuchi J, Tokuda K. Text-to-Visual speech synthesis based on parameter generation from HMM. In: Proc. of the 1998 IEEE Int’l Conf. on Acoustics, Speech and Signal Processing (Ⅵ). 1998. 3745−3748.

[6] Petajan E. Approaches to visual speech processing based on the MPEG-4 face animation standard. In: 2000 IEEE Int’l Conf. on Multimedia and Expo (ICME 2000). 2000. 575−578.

[7] Bregler C, Covell. M, Slaney M. Video rewrite: Driving visual speech with audio. In: Proc. of the ACM (Association for Computing Machinery) SIGGRAPH Conf. on Computer Graphics. 1997. 353−360.

[8] Ezzat T, Poggio T. Visual speech synthesis by morphing visemes. International Journal of Computer Vision, 2000,38(1):45−57.

[9] Cosatto E, Potamianos G, Graf HP. Audio-Visual unit selection for the synthesis of photo-realistic talking-heads. In: IEEE Int’l Conf. on Multimedia and Expo (Ⅱ). New York, 2000. 619−622.

[10] Cosatto E, Graf HP. Photo-Realistic talking-heads from image samples. IEEE Trans. on Multimedia, 2000,2(3):152−163.

[11] Kshirsagar S, Magnenat-Thalmann N. Visyllable based speech animation. In: Brunet B, Fellner D, ed. Eurographics 2003. Oxford: Balckwell, 2003. 631−639.

[12] Edge J, Sanchez M, Maddock S. Reusing motion capture data to animate visual speech. In: Symp. on Language, Speech and Gesture for Expressive Characters, part of the AISB 2004 Convention: Motion, Emotion and Cognition, University of Leeds, 2004.

[13] Cao Y, Faloutsos P, Kohler E, Pighin F. Real-Time speech motion synthesis from recorded motions. In: Jean-Dominique Gascuel, ed. ACM SIGGRAPH/Eurographics Symp. on Computer Animation 2004. Grenoble, 2004.

[14] Yang DN, Guo F, Wen CY. Media transformation from Chinese text to mouth image. ACTA Electronica Sinica, 1996, 24(1):122−125 (in Chinese with English abstract).

[15] Yan J. Text-Driven lip motion synthesis system. Computers Engineering and Design, 1998,19(1):31−34 (in Chinese with English abstract).

[16] Gao W, Chen XL, Yan J, Song YB, Yin BC. Synthesis of facial behavior for virtual human. Chinese Journal of Computers, 1998, 21(8):694−703 (in Chinese with English abstract).

[17] Dong LF, Wang X, Chen YY. Implementation and application of virtual face. Mini-Micro System, 2002,23(1):90−92 (in Chinese with English abstract).

[18] Wang ZM, Cai LH, Tao JH, Wu ZY. Study of text to visual speech in Chinese. Mini-Micro Systems, 2002,23(4):474−477 (in Chinese with English abstract).

[19] International Standard, Information Technology-Coding of Audio-Visual Objects-Part 2: Visual; Admendment 1: Visual extensions, ISO/IEC 14496-2: 1999/Amd.1:2000(E).

[20] Wang ZM, Cai LH. Study of Chinese viseme. Applied Acoustics, 2002,21(3):29−34 (in Chinese with English abstract).

[21] Chen T. Audiovisual speech processing. IEEE Signal Processing Magazine, 2001,18(1):9−21.

[22] Hong PY, Wen Z, Huang TS. Real-Time speech-driven face animation with expressions using neural networks. IEEE Trans. on Neural Networks, 2002,13(4):916−927.

[23] Kim JW, Song M, Kim IJ, Kwon YM, Kim HG, Ahn SC. Automatic FDP/FAP generation from an image sequence. In: Hasler M, ed, Proc. of the 2000 IEEE Int'l Symp. on Circuits and Systems. 2000. 40−43.

[24] Sarris N, Grammalidis N, Strintzis MG. FAP extraction using three-dimensional motion estimation. IEEE Trans. on Circuits and Systems for Video Technology, 2002,12(10):865−876.

[25] Jan PH, Van Santen, Richard WS, Olive JP, Hirschberg J. Progress in Speech Synthesis. New York: Springer-Verlag, 1997.

[26] Huang FJ, Graf HP, Cosatto E. Triphone-Based unit selection for concatenative visual speech synthesis. In: Taylor FJ, ed. Proc. of the Int'l Conf. on Acoustics Speech and Signal Processing, ICASSP2002. 2002.

[27] Graf HP, Cosatto E, Strom V, Huang FJ. Visual prosody: Facial movements accompanying speech. In: Proc. of the Int'l Conf. on Automatic Face and Gesture Recognition (FG02). 2002.

附中文参考文献:

[14] 杨丹宁,郭峰,文成义.由文本至口形的媒体变换技术的研究.电子学报,1996,24(1):122−125.

[15] 晏洁.文本驱动的唇动合成系统.计算机工程与设计,1998,19(1):31−34.

[16] 高文,陈熙霖,晏洁,宋益波,尹宝才.虚拟人面部行为的合成.计算机学报,1998,21(8):694−703.

[17] 董兰芳,王洵,陈意云.真实感虚拟人脸的实现和应用.小型微型计算机系统,2002,23(1):90−92.

[18] 王志明,蔡莲红,吴志勇,陶建华.汉语文本-可视语音转换的研究.小型微型计算机系统,2002,23(4):47−477.

[20] 王志明,蔡莲红.汉语语音视位的研究.应用声学,2002,21(3):29−34.