Dynamic Mapping Method based Speech Driven Face Animation System

Panrong Yin and Jianhua Tao

National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences, Beijing
{pryin, jhtao}@nlpr.ia.ac.cn

Abstract. In the paper, we design and develop a speech driven face animation system based on the dynamic mapping method. The face animation is synthesized by the unit concatenating, and synchronous with the real speech. The units are selected according to the cost functions which correspond to voice spectrum distance between training and target units. Visual distance between two adjacent training units is also used to get better mapping results. Finally, the Viterbi method is used to find out the best face animation sequence. The experimental results show that synthesized lip movement has a good and natural quality.

1 Introduction

Visual speech synthesis has been developed for improving human-machine interface such as virtual announcer, email reader, mobile messenger reader and so on. It also helps hearing-impaired people to communicate with others or those in noisy environment.

According to driven sources, visual speech synthesis can be categorized into text driven face animation and speech driven face animation. Since TTVS [8],[9] depends too much on languages content analysis and can not offer co-articulation information, more and more researchers pay attention to synthesis from real speech [1],[2],[3],[4].

A speech driven face animation system is usually established in four steps as defined in "picture my voice"[1]: label a audio-visual database; give representation of both the auditory and visual speech; find a method to describe the relationship between two representations; synthesize the visible speech given the auditory speech.

A labeled audio-visual database is the base of the whole system. But there are little labeled databases of visual speech and none universal one, so many researchers have to record their own database for training visual speech. Some recorded a stream of dynamic images of interested regions in videos [6],[8],[9],[13], another recorded static facial visemes by 3D laser scans, the other obtained dynamic 3D coordinates of marked-up face through motion capture system [2],[14] or 2D to 3D image reconstruction method [3],[4].

Good feature representations have an important impact on the system performance. For audio features, some researchers used text-dependant language

units to analyze their corresponding static visemes; but in order to reduce manual intervention, many researchers turned to extract acoustic features of speech as audio features. For visual features, many researchers used images [6],[8],[13], visemes[9], FAPs[14], PCs[3],[4], 3D coordinates[14], 3D distance measurements [2],[3], optical flows[13].

The key component of visual speech synthesis is audio-visual mapping. Many methods have been applied in this area: TDNN[1], MLPs[4], KNN[3], HMM[2], GMM, VQ[2], Rule-based[9],[13],[14]. At present, there are mainly two approaches for synthesis: through speech recognition and not through speech recognition. The first approach divides speech signal into language units such as phonemes, syllables, words, then maps the units directly to the lip shapes and concatenates them. Yamamoto E.[2] recognized phonemes through training HMM, and mapped them directly to corresponding lip shapes, through smoothing algorithm, the lip movement sequence is obtained. The second approach analyzes the bimodal data through statistical learning model, and finds out the mapping between the continuous acoustic features and facial control parameters, so as to drive the face animation directly by a novel speech. "Picture my voice"[1] trained an ANN to learn the mapping from LPCs to face animation parameters, they used current frame, 5 backward and 5 forward time steps as the input to model the context.

After getting the FAP streams from audio-visual mapping method, there are two ways to synthesize talking head: model-based animation [3],[4],[11],[14] and image-based animation[6],[8],[9],[13].

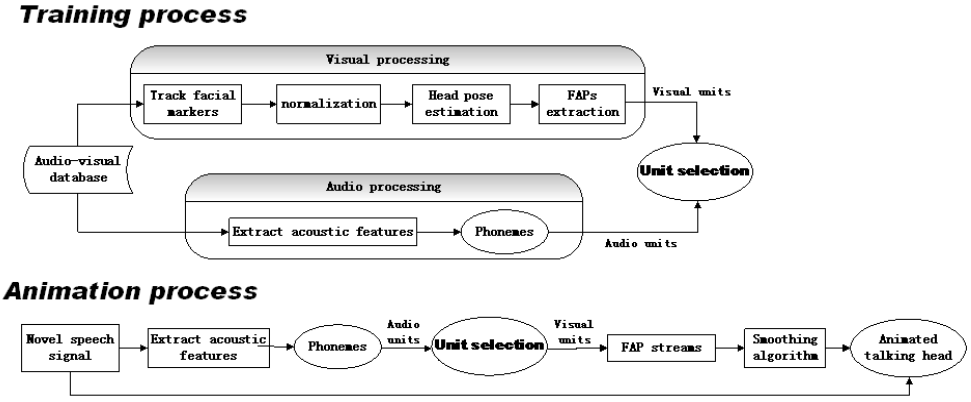


Fig. 1. Speech driven face animation system framework

Recently, a new approach for audio-visual mapping has arisen [8],[10], which is inspired from speech synthesis [7]. This method means to construct new data stream by concatenating stored data units in training database. It has advantage of that the synthesis result appears very natural and realistic. Our approach con-

siders phonemes as units instead of frames, because frame-by-frame mapping is difficult to take account for phoneme context and may lose some co-articulation information. In our system, model-based animation model is chosen as talking head model, though it appears less reality than image-based model, it requires less computation cost, less training database and can be easily replanted to embedded system such as PDA, mobile phone. In this paper, 3D coordinates of FDPs which are compatible with MPEG-4 standard [5] can be recorded by Motion Capture System, they reflect real facial movements. Since animated talking head is required to be more lifelike and individual, FAPs are extracted as visual features so as to apply our synthesized animation control parameters to different 3D mesh models.

Our talking head system is composed of two processes: training process and animation process (Fig. 1). In the rest of the paper, section 2 introduces the database setup and audio-visual feature extraction, section 3 focuses on the unit selection realization, section 4 gives the experimental result of synthesized talking face, section 5 is the conclusion and future work description.

2 Data Acquisition and Analysis

2.1 Data Acquisition

Since FAPs in MPEG-4 facial animation standard can denote face movement which is speaker independent, a commercially available motion capture system (Motion Analysis) is used to track the FDPs on speaker's face as shown in Fig. 2 and 8 cameras with 75 Hz sampling frequency are employed. According to FDPs defined in MPEG-4, 50 points are selected to encode the face shape. The output of the motion capture system are 3D trajectories of all the 50 markers. During experiment, the rigid head movement is not avoided, so 5 markers on the head are used to compensate the global head movement to get the non-rigid facial deformations.

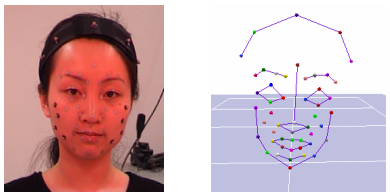


Fig. 2. Placement of markers on neutral face and capture data sample

2.2 FAPs Extraction

In order to extract FAPs correctly, the 3D trajectories of face markers have to be normalized into an upright position in positive XYZ space firstly. And then

the least-square based fitting method of two 3D point sets is used to extract the rigid head movements[12], in our case, the two point sets correspond to the 5 points on the head before and after head movement. Their relationship can be described by:

$$q_i = Rp_i + t, \quad i = 1, \ldots, 5. \tag{1}$$

Once rotation R and translation t matrixes are estimated, the global head movements can be eliminated by back projecting Equation.1 to 3D trajectories of rest points.

There are two most popular visual representation standard, one is the Facial Action Coding System[15] developed by Ekman and Friesen, the other is MPEG-4 SNHC [5]. FACS defined 44 AUs to denote the independent facial muscles' action, more than 7000 AU combinations have been observed. 68 FAPs which are defined in MPEG-4 SNHC also can represent the basic facial actions, including two high-level parameters, viseme and expression, and 66 low-level parameters. They are divided into 10 groups according to effect regions. Since AUs can not offer quantitative description for face animation, our approach is to convert 3D FDPs movements into low-level FAPs. FAPs are computed and normalized in terms of FAPUs which are the distances between the main facial feature points. So they can be applied to different face models in a consistent way. In the paper, 15 FAPs (Table. 1) involved in group 2 and group 8 are extracted to represent the lip movement.

Table 1. FAPs for visual representation in our system

Group	FAP name	Group	FAP name
2	Open_jaw	8	Stretch_l_cornerlip_o
2	Trust_jaw	8	Lower_t_lip_lm_o
2	Shift_jaw	8	Lower_t_lip_rm_o
2	Push_b_lip	8	Raise_b_lip_lm_o
2	Push_t_lip	8	Raise_b_lip_rm_o
8	Lower_t_midlip_o	8	Raise_l_cornerlip_o
8	Raise_b_midlip_o	8	Raise_r_cornerlip_o
8	Stretch_r_cornerlip_o		

So the output of motion capture system at t step:

$$d = \{x_1, y_1, z_1, \ldots, x_n, y_n, z_n\}^T \in R^{3n}, \quad n = 45.$$

can be converted into:

$$\bar{v}^t = \{FAP_1^t, FAP_2^t, \ldots, FAP_{15}^t\}^T.$$

So the visual representation for a data sample which contains m frames is:

$$V = \{\bar{v}^1, \bar{v}^2, \ldots, \bar{v}^m\}^T.$$

2.3 Acoustic Feature Vectorization

There are different parameters for representation of acoustic speech, such as LPC, LSF, MFCC, formant and so on. MFCC gives an alternative representation to speech spectra, which contains some audition information. So in our approach, MFCCs in each audio frame are chosen as the audio feature. The speech signal, sampled at 16 kHz, is blocked into frames of 27 ms. In each frame, we computed 16 coefficients. So each t frame of speech can be represented as:

$$\bar{a}^t = \{c_1^t, c_2^t, \ldots, c_{16}^t\}^T.$$

These vectors are then grouped into a matrix in a data sample with m frames:

$$A = \{\bar{a}^1, \bar{a}^2, \ldots, \bar{a}^m\}^T.$$

3 Audio-visual Dynamic Mapping

In order to synthesize a vivid and realistic talking head, an audio-visual dynamic mapping method which is inspired from the concatenative speech synthesis[7] is employed in our system. The difference is that they select speech units to synthesize a novel concatenative speech, but we select audio-visual units according to a novel speech input to synthesize continuous FAP streams to drive a talking head.

This method is based on two cost functions (Equation. 2):

$$COST = \alpha C^a + (1 - \alpha)C^v. \tag{2}$$

where C^a is the voice spectrum distance, C^v is the visual distance between two adjacent units, and the weight α balances the effect of the two cost functions.

It is also much like HMM, the synthesized unit flows are the hidden states, phoneme pieces of speech input are the observable states. Difference stays at the cost functions instead of probabilities[7].

In the unit selection procedures, according to the target phoneme unit of novel speech, we list the candidates which have smaller acoustic distances with them. Because the acoustic parameters MFCCs are related to people's vocal tract, so smaller acoustic spectrum distance reflects smaller visual difference.

The audio content distance measures how close the candidate unit compared with the target unit, it determines whether the most appropriate audio-visual unit have been selected. The voice spectrum distance also accounts for the context of target unit. The context covers the former and latter two phonemes for vowels and the former and latter phoneme for consonants. So the voice spectrum distance is defined by Equation. 3:

$$C_j^a = \sum_{t=1}^{n} \sum^{m} w_m a(t_{t+m}, u_{t+m}), \quad m = [-1, 0, +1] or [-2, -1, 0, +1, +2]. \tag{3}$$

The weights w_m are determined by the method used in[11], the difference is that they computed the linear blending weights in terms of a phoneme's time duration, but we take three or five adjacent phonemes into account. $a(t_{t+m}, u_{t+m})$ is the Eucilidian distance of the acoustic feature of two phonemes. For the sake of reducing the complexity of Viterbi search, we set a limit of candidate number for selection.

Not only we should found out the correct speech signal unit, but also the smooth synthesized face animation should be considered. In training process, we just label the phoneme positions in each sentence without segmenting them, so it will enable the unit selection to find out the smoothest unit following the previous one. The concatenation cost measures how closely the mouth shapes of the adjacent units match. So the FAPs of the last frame of former phoneme unit are compared with that of the first frame of current phoneme unit (Equation. 4).

$$C^v = \sum_{t=2}^{n} v(u_{t-1}, u_t). \tag{4}$$

Where $v(u_{t-1}, u_t)$ is the Eucilidian distance of the adjacent visual features of two phonemes.

Once the two cost functions (audio content distance and visual distance) are computed, the graph for unit selection is constructed. Our approach finally aims to find the best path in this graph which generates minimum COST. Just like HMM, Viterbi is a valid search method for this application.

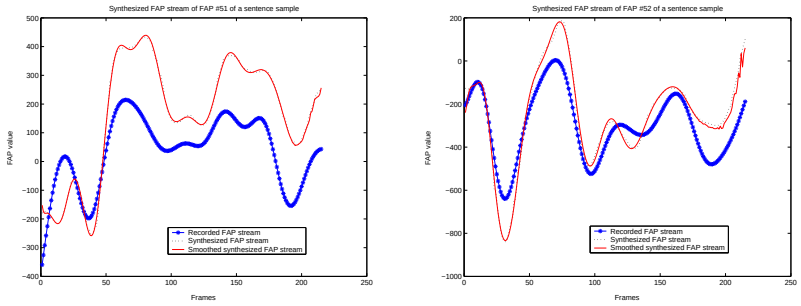
4 Experimental Results

In the paper, a collection of 286 sentences are recorded. For each, four times are used as training data, one time as validation data and another 9 sentences as test data. Fig. 3 indicates the selected synthesized FAP stream results. In order to evaluate the synthesized result, both quantitative evaluation and qualitative observation results are shown. Since the synthesized FAP streams are concatenated by units in training database, so it is impossible to be absolutely same with the recorded FAP stream. Correlation coefficients (Equation. 5) are used to represent the deviation similarity between them, because it is a measure of how well trends in the predicted values follow trends in past actual values. Table. 2 shows the average correlation coefficients for each FAPs on the whole database.

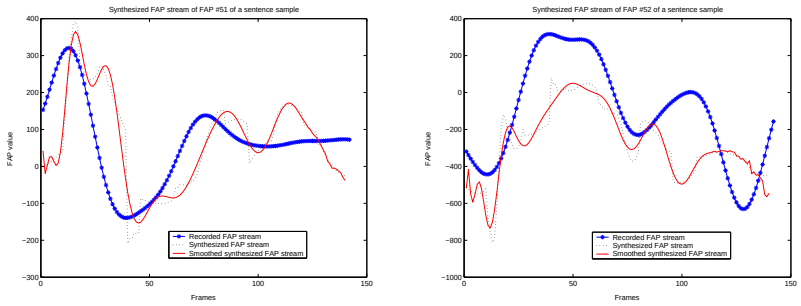
$$CC = \frac{1}{T} \sum_{t=1}^{T} \frac{(\hat{f}(t) - \hat{\mu})(f(t) - \mu)}{\hat{\sigma}\sigma}. \tag{5}$$

where $f(t)$ is the recorded FAP stream, $\hat{f}(t)$ is the synthesis, T is the total number of frames in the database, and μ and σ are corresponding mean and standard deviation.

From Table 2, the synthesized results show a good performance of our method. Although the test sentences show smaller correlation coefficients, they still can



(a) validation



(b) test

Fig. 3. Selected synthesized FAP streams from (a) validation and (b) test sentence

Table 2. The average correlation coefficient for each FAPs on the whole database

Correlation Coefficients		Validation	Test
FAP No.	#3	0.855	0.622
	#14	0.674	0.656
	#15	0.553	0.761
	#16	0.767	0.677
	#17	0.716	0.788
	#51	0.820	0.732
	#52	0.812	0.741
	#53	0.817	0.722
	#54	0.832	0.636
	#55	0.709	0.684
	#56	0.768	0.736
	#57	0.749	0.714
	#58	0.787	0.720
	#59	0.661	0.640
	#60	0.569	0.730

synthesize most of FAPs. It is also noticed that the CC of FAP #15 and #60 of validation data are even smaller than those of test data, it may be mainly because of the speaker's habit of that she speaks sentences with randomly shifting mouth corner.

At last, a MPEG-4 facial animation engine is used to access our synthesized FAP streams qualitatively. The animation model displays at a frame rate of 75 fps. Fig. 4 shows some frames of synthesized talking head.

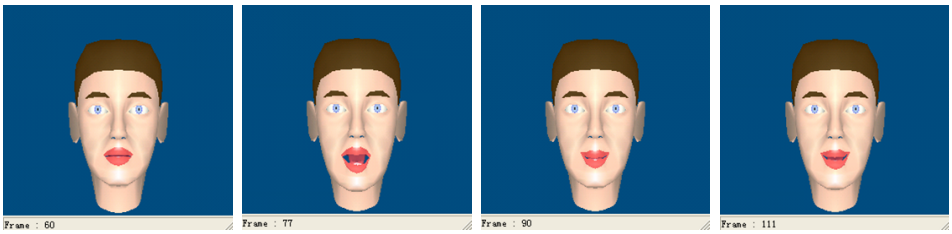


Fig. 4. Some frames of synthesized talking head

While many other researchers used images, optical flows, principal components. etc as stored units, in this paper, FAPs are extracted. It has advantage of small storage, being able to be applied to non-specific model, and also to drive both 3D mesh model and 2D images. The unit selection dynamic mapping method we used is a simplification of HMM, but it appears very good and natural, because the real talker's face movements are treated as the candidate units. Without parameter adjust, under-fitting, over-fitting problems which are involved in learning method like ANN, it is a simple, direct and effective way to synthesize the talking head.

5 Conclusion and Future work

In the present work, we have presented a speech driven face animation system which is based on the MPEG-4 animation model. The system employs FAPs which are extracted from 3D coordinates of FDPs to represent the visual feature. The acoustic features of real speaker's speech are used to directly drive animation model. It makes the driven source independent from text content and languages. The unit selection method is applied for audio-visual dynamic mapping. Both audio content distance and the visual distance between the adjacent units are taken into account. Finally, we give the quantitative and qualitative evaluation for the synthesized concatenative FAP streams.

For a big training database, our method appears lower speed to find out the corresponding audio-visual units, so we now investigate some statistic methods to cluster the phonemes in training database. As one speaker's speech is tested for the present work, different person' novel speech will be inputs in real application,

so in the future work, we can apply some methods of voice conversion to extend our system to different voices.

References

1. Dominic W. Massaro, Jonas Beskow, Michael M. Cohen, Christopher L. Fry, Tony Rodriguez: Picture My Voice: Audio to Visual Speech. Synthesis using Artificial Neural Networks. Proceedings of AVSP'99, pp.133-138. Santa Cruz, CA.(1999).
2. Yamamoto E., Nakamura, S., Shikano, K: Lip movement synthesis from speech based on Hidden Markov Models. Speech Communication, Vol. 26,(1998)105-115.
3. R. Gutierrez-Osuna, P. K. Kakumanu, A. Esposito, O. N. Garcia, A. Bojorquez, J. L. Castillo, I. Rudomin: Speech-Driven Facial Animation With Realistic Dynamics. IEEE Trans on Multimedia, Vol.7, No.1,(2005)
4. Pengyu Hong, Zhen Wen, Thomas S. Huang: Real-time speech-driven face animation with expressions using neural networks. IEEE Trans on Neural Networks, Vol.13, No.4,(2002)
5. A. Murat Tekalp, JoK rn Ostermann: Face and 2-D mesh animation in MPEG-4, Signal Processing: Image Communication Vol.15,(2000)387-421
6. Bregler C., Covell M., Slaney M.: Video Rewrite: Driving Visual Speech with Audio. ACM SIGGRAPH, (1997)
7. Hunt A., Black A.: Unit selection in a concatenative speech synthesis system using a large speech database. ICASSP, vol.1,(1996)373-376
8. Cosatto E, Potamianos G, Graf H P: Audio-visual unit selection for the synthesis of photo-realistic talking-heads. IEEE International Conference on Multimedia and Expo, ICME Vol.2,(2000)619-622
9. T. Ezzat, T. Poggio: MikeTalk: A Talking Facial Display Based on Morphing Visemes. Proc. Computer Animation Conference, Philadelphia, USA, (1998)
10. Ram R. Rao, Tsuhan Chen, Russell M. Mersereau: Audio-to-Visual Conversion for Multi-media Communication. IEEE Trans on Industrial Electronics, Vol.45, No.1,(1998)
11. Jian-Qing Wang, Ka-Ho Wong, Pheng-Ann Pheng, Meng H.M., Tien-Tsin Wong: A real-time Cantonese text-to-audiovisual speech synthesizer. ICASSP '04, Vol.1,(2004)
12. K.S. Arun, T.S. Huang, S.D.Blostein: Least-square fitting of two 3-D point sets. IEEE Trans on Pattern Analysis and Machine Intelligence, vol.9, no.5,(1987)698-700
13. Ashish Verma, L. Venkata Subramaniam, Nitendra Rajput, Chalapathy Neti, Tanveer A. Faruquie: Animating Expressive Faces Across Languages. IEEE Trans on Multimedia, Vol.6, No.6,(2004)
14. S. Kshirsagar, T. Molet, and N.M. Thalmann: Principal Components of expressive speech animation. Proc. Of Computer Graphics International,(2001)
15. P. Ekman and W. V. Friesen: Facial Action Coding System. Palo Alto, Calif: Consulting Psychologists Press,(1978)