Mutual Features Controlling of Dynamic Visemes for Emotional Talking Head System

Jianhua Tao   Le Xin
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100080
jhtao@nlpr.ia.ac.cn

Abstract

Natural Human-Computer Interface requires integration of realistic audio and visual information for perception and display. In this paper, a lifelike talking head system is proposed. The system converts text to speech with synchronized animation of mouth movements and emotion expression. The talking head is based on a generic 3D human head model. The personalized model is incorporated into the system. With texture mapping, the personalized model offers a more natural and realistic look than the generic model. To express emotion, both prosody information and Emotional Markup Language are used to generate the dynamic viseme models. Final system was proved to be able to offer the natural and vivid audio-visual output.

Key words: Emotion, Talking Head, Mutual Features

1. Introduction

Natural Human-Computer Interface requires integration of realistic audio and visual information for perception and display, Multimodal expression system makes the communication easy and friendly between human and computer. It covers several areas, such as speech processing, facial and body tracking and synthesis, etc. The face of a speaker conveys many kinds of information about both the speaker and the content of what is being said. Much of previous work has been focused on the contribution of time-varying events on the face. e.g., the behavior of the lips, jaw, cheeks, and head motion [12][14][15][16][17]. Other studies have indicated the importance of other time-varying landmarks such as the eyes and the eyebrows in providing contextual and paralinguistic cues (e.g., [1][3]). As yet, however, we know very little about what role more global structural features such as the 3D shape of the face play in visual communication, whether shape is considered statically or as it varies over time. In the paper, we apply classification method on Chinese phonemes with our 3D database of dynamic faces to determine the dynamic visems, which form the basic features for talking head.

As we know, emotion does an important role in the interaction. Basically, emotion can be expressed as *joy, sadness, anger, surprise, hate, fear* and so forth. The classification category is not defined clearly yet. So we focus on the related research about emotion in cognitive psychology domain, and five basic emotion states, happy, fear, disgust, sad and anger are selected for processing in the paper. To generate emotional talking head, there exist three basic problems, methods for facial animation, emotional speech synthesis, audio-visual synchronized with emotion.

To generate emotional speech output, most of previous work is focused on the acoustic analysis or articulatory analysis [2][5][10][24][25][26][27]. In the paper, we also do the similar analysis to get the emotion features of Chinese speech, which is traditionally necessary for speech synthesis. But for the whole emotional speech synthesis, it is also very important to get emotion state from context input. While the broad topic of emotion has been studied in psychology for decades, very little effort has been spent on attempting to detect emotion in text. Chuang[14] has developed a semantic network for emotion extraction from textual content, but there are less corpus to support the results. In the paper, we assume that the emotion reaction of an input sentence is essentially represented by its word appearance. To get emotion state, all of the words are divided into *content words* and *Emotion functional words* (EFWs). They are manually defined and used to extract emotion from the input sentence. All of the extracted emotion functional words have their corresponding connection to “basic emotion values” which are defined in the lexicon. For each input sentence, the basic emotion values are combined to give the final emotion output with emotion estimate net.

In talking head, except for the dynamic visemes, the emotional functional words are assigned as the basic trigger for emotional facial expression in the integration between emotional speech synthesis and facial animation. Final results show that the emotional talking head system created in paper could generate the vivid audio and visual output.

The paper is organized as following, in section II, the paper makes detailed description on emotional speech synthesis. The detailed acoustic analysis is made to get features of emotional Chinese speech. With the context analysis, a context based emotional speech synthesis is finally created in this section. Section III describes the integration method for emotional talking head system. With this, a bimodal emotional expression system for audio-visual synchronization is proposed based on the modeling of the realistic 3D face.

2. Dynamic Visemes

To get the viseme models for Chinese phonemes, we classify visemes into 9 types, according to lip height, outer lip width (including upper teeth) and lip protrusion. They are listed in table 1. Here, [a] represents the first class which is related to the largest lip height; The second and fifth class typical of [o] and [u] take on the stamp of lip protrusion; The third class represented by [e] is provided with the feature of mid-vowel; The fourth class is similar to the third one but has wider lip opening; Class[d] are consistent with the wide lip opening and unconspicuous lip protrusion; The class [f] has the closer lips and has the subtle difference in the outer lip width and upper teeth revealable degree; The class [j] and class [i] behave similarly; The classes [b]/[p]/[m] are notably distinguished by their feature of close lips from other classes. Moreover, not only the static parameters of one phoneme but also its dynamic parameters for deformation and variation are used to perceive a phoneme by vision. Therefore, durations of phonemes are taken into account for their importance. Vowels and consonants should be considered respectively due to their remarkably diverse durations. It not only conforms to the perception of syllables but also accords with research routines of other languages.

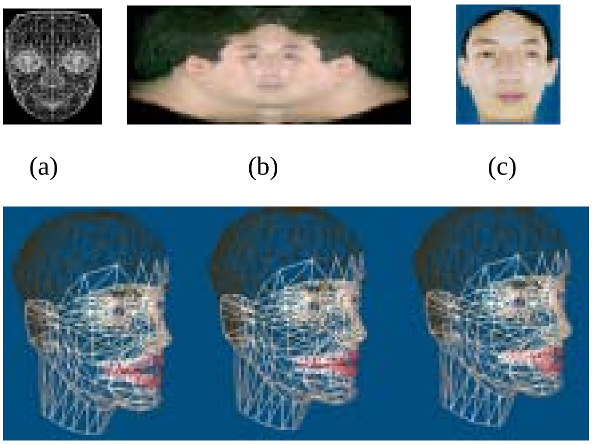

(a) (b) (c)

Figure 2, Lip movements with phonemes “ai”, “o” and “d”

Table1, The classification of Chinese visemes

viseme	**[a]**	**[o]**	**[e]**	**[i]**	**[u]**	**[d]**	**[f]**	**[j]**	**[b]**
phonemes	si an ang ao a	o ong ou	e ei en eng er	i ~i in ing	u ü	d t n l ng g k h	f	R~ j q x z c s zh ch sh	b p m

3. Emotional Speech Synthesis

3.1 Emotional features in Chinese speech

The representation of the speech correlates of emotion can proceed from a speaker model or an acoustic model. According to Cahn’s work, the parameters of the acoustic model are grouped into four categories, pitch, timing, voice quality and articulation [2]. Being the tonal structure of Chinese, the pitch information includes,

F0 mean: the average of F0 for the utterance. F0 range: the range of F0 for the utterance. Slope of top-line: the slope of the top F0 sequence of the syllables. Slope of bottom-line: the slope of the bottom F0 sequence of the syllables.

Speech rate is the major parameter for timing information. Physical state of the vocal tract was also taken as parameter. A valuable cue for the characterization of anxious speech is the amount ‘f0 jitter’ which describes the variation of f0 from one pitch period to another. Furthermore, energy is also an important parameter for emotion determine. It is measured as mean energy, or as energy at a reference point. Due to glottal pulse missing it can generate creaky voice in such fear emotions. Sometimes, breathy appears in happy or anger emotions. Another correlate for affect is articulation which is classified into normal, tense, slurring and precise. Articulation describes the changes in quality of vowels and whether the reduction of unvoiced consonants is reduced to their voiced counterparts. Table 1 lists some acoustic features related to five emotion states based on our corpus.

To analyze the correlation of speech and facial expressions further, we collect large amounts of TV signals and from them carefully select some dialogue scenes where facial features are distinct and voice is articulate (Total 2621 sentences). The head is in the centre of selected images and emotional characteristics are distinct during some time. Data of three men were analyzed. The detailed analysis is described in section 4.

Table2, Acoustic features of five emotions

	Happiness	**Anger**	**Sad**	**Fear**	**Disgust**
Speech rate	Faster, but sometimes slower	Slightly faster	Slightly slower	Much faster	Very much slower
F0 Mean	Much higher	Very much higher	Slightly lower	Very much higher	Very much lower
F0 Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Slope of top-line	Smooth, upward inflections	Abrupt, on stressed syllables	Downward inflections	Normal	Wide, downward terminal inflections
Slope of bottom-line	Smooth, upward inflections	No too much inflections	Downward inflections	Normal	downward terminal inflections
Intensity	Higher	Higher	Lower	Normal	Lower
Voice quality	Breathy, blaring	Breathy, chest tone	Resonant	Irregular voicing	Grumbled chest tone
Articulation	Normal	Tense	Slurring	Precise	Normal

3.2 Emotion state prediction from context input

To get emotional speech synthesis output, it is very important to know where and how emotion behaves after the text inputting. There are two ways for emotion generation. One is from semantic information, human likes to presenting the emotion according to what he (or she) wants to say. The other is from the environment and psychology. Sometimes, the content is not very important, if he (or she) has a strong feeling to express the emotion while they are in different situations. To extract the appropriate emotion state from context information, each input sentence could be considered as the combination of content word and emotion functional words (EFWs), though most of them only contain content words. EFW is a kind of word which could be linked to the special emotion states or has some influence on them. In emotion detection, they supply the basic emotion values or connection. With the semantic relation, the emotion value will be propagated in the whole sentence.

3.2.1 Emotion Keywords

The most important words in EFWs are emotion keywords, which provide the basic emotion value of the input sentence. Normally, it is not easy to classify the words to different emotion state, emotion hides in human's experience long before the history of language. There are lots of ambiguities, most of them occur in anger and sadness. For example, the word "unhappy" may indicate angry or sad, according to the different personality and situation. To get more accurate description, we should assign the weight for each of them. For example, we assigned the weight as 0.5 (for angry) and 0.5 (for sad) for the word "unhappy". It means "unhappy" has equal possibilities for emotion state angry and sad. The final results could be the combination with these values. In Chinese, there are 390 emotional keywords in total. Most of them are noun, verb and adjective. To reduce the complexity, we only choose 6 types for the basic emotion states, *joy, sadness, anger, surprise, hate* and *fear*.

3.2.2 Modifier words

Except for the emotional keywords, the emotion state could also be influenced by some words which behave as modifier in the sentence, such as "very, so, too much, not, etc.". Normally, emotion keywords represent the major emotion reaction of the sentences related to a certain topic. The modifier words are normally used to enhance the mood. The effect of "intense" mood could be obviously represented by emphasizing the emotion keywords. For example, *I'm so angry*. The phrase "*so angry*" denotes the key emotion state of the sentence, and is extremely emphasized.

3.2.3 Metaphor words

The other part of EFWs seems to have no direct action on the emotion states but do have the latent influence on them. They denote the attitude and moral character to make positive and negative influence on emotional keywords. For example, "asperity" is more like related to exaggerating and negative emotion, "*anger*" and "*hate*", but "*kindness*" always concerns the gentle and positive emotion, "*joy*" or "*neutral*". The metaphor words can be divided into two types, one is for spontaneous expressing, such as "*anxious, deferential, ardent, fierce,* etc.", the others denote personal character, such as "*chipper, arrogant,* etc". In our work, the whole amount of metaphor words is 440. Most of them are adjective, among which 201 are related to positive feeling and the others are for negative expression.

3.2.4 EFW Lexicon

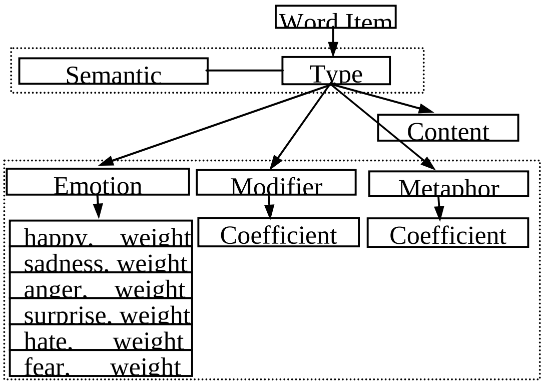


Figure 3, Lexicon Structure

The basic emotion value of each EFW is manually defined based on a Chinese lexicon, which includes 462 words, but there are 65620 words with semantic tagging in the lexicon in total. In order to eliminate the error due to subjective judgment, all words are firstly tagged by three people individually and then crossly validated by the other two people. For each word, if the results tagged by different people are close, the average of these values will be set as the basic emotion value of the word. If the three people can not reach a common consensus, an additional person will be asked to tag the word and the result will be taken into consideration. Experientially, only few words need additional suggestion. The whole lexicon is organized in figure 3.

3.3 Emotion Estimation

Normally, the emotion state could be decided by emotion keywords and other labeling information. However, in some situation, the ambiguity may appear. It mainly results from the multivocal of emotion keywords. The emotion reaction is first deduced from initial emotion assign according to word classification above, and also from the combination of the semantic relations. In order to get the context sensitive model, we propose a unified architecture based on Emotion eStimation Net (ESiN) that seamlessly integrates context-dependent probabilistic hierarchical sub-lexical modeling.

ESiN is composed of nodes and routes. Each node denotes a word which has three attributes: emotion states, the corresponding weights and semantic tagging. The route of ESiN represents the propagation of the emotion. It contains three attributes: direction, transmission probability P_{ij} (denote the probability from one emotion state i to another state j) and propagation decreasing coefficient α.

$$\vec{E}_t = D(\vec{E}_{t-1}) = \vec{\delta}(t) + \vec{E}_{t-1}\exp(-\alpha \times t^2) + \vec{C}_t \qquad (6)$$

Here, $\vec{E}_t = (e_{t,joy}, e_{t,sadness}, e_{t,anger}, e_{t,\sup rise}, e_{t,hate}, e_{t,fear})$ contains the emotion values of all emotion states. $\vec{\delta}(t)$ contains all emotion values generated in node t. E_t represents the emotion vector in node t.

If current node is content word, $\vec{\delta}(t) = P(o_t \mid o_{t-1})\vec{\delta}(t-1)$ get from the semantic relation between two nodes.

If the node is emotional keyword, $\vec{\delta}(t) = (\omega_{t,joy}, \omega_{t,sadness}, \omega_{t,anger}, \omega_{t,\sup rise}, \omega_{t,hate}, \omega_{t,fear})$ is the initial emotion weights defined in the lexicon.

If the node is modifier or metaphor word, all of the $\vec{\delta}(t)$ should be multiplied by coefficient β_t.

Without new simulation, the vector E_t will damp to zero through some words.

To make more detailed description, we give a sample as following,

"*Mr.Wang is too introversive to speak out, though he feels very pleasure while he hears the news.* "

Detection of emotion in text by ESiN is then followed by the following steps.

- **First step: EFWs Detection**

In the first step the text is tagged with a POS tagger. The tagger learns sentence structures for a language as a set of transition rules. These rules are then applied to the text to label each word as a noun, verb, etc. Once words are labeled, they are checked for EFWs and assigned an emotional rating.

From above sample, the EFWs are emotion keyword "pleasure", modifier word "too, very", and metaphor word "introversive".

- **Second Step: Weight Assign and Link Construction**

The second step is to assign the weight for emotional keywords and construct the link among EFWs. Here, "pleasure" is tagged as emotion "joy", From the lexicon, the initial weight is 0.9.

There are two modifier words in the sample, but only "very" linked to emotion keyword "pleasure". "very" is tagged as positive to modify the preliminary valence score. metaphor word "introversive" could decrease preliminary valence score.

- **Final Step: Propagation, Collection and Decision**

In propagation, the emotion keyword is considered as the propagating source. The scores are then summed across all sentences and finally run through a fuzzy-logic process to determine an overall score for the correspondence. In ESiN, valence is determined from a proprietary list of emotionally charged words, abbreviations, and emoticons. Administrators of the system are free to add new emotion words, or change the values associated with existing words.

The aim of the emotion trigger is to integrate the non-zero emotion vector according to the emotion state history by path searching.

$$M_t = \arg\max_n \left(\sum_{i=0}^{t} e_{n,i}\right) \qquad (7)$$

4. Audio-Visual Integration with Emotion Expression

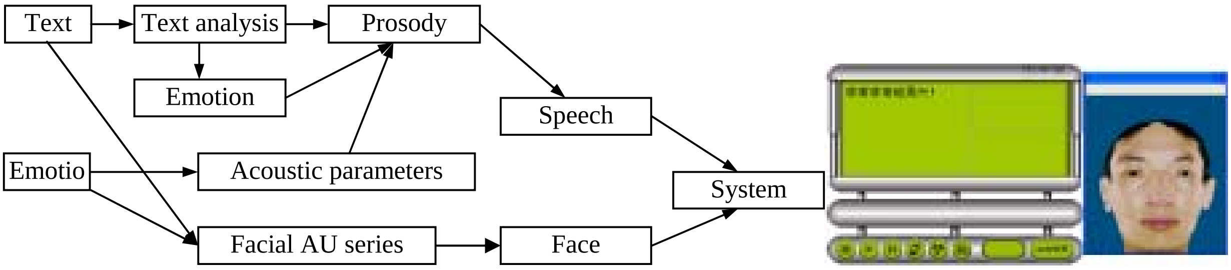
Based on above results, we integrate the emotional speech synthesis and facial animation to generate the talking head system. The motion of lip shape and eye region is related to AU in FACS, such as happy is related to AU12 and AU25 with faster speech rate. In many cases, happyness and sadness cannot be distinguished very effectively merely using facial features, while their corresponding acoustic features can work well. Generally, emotional face features and lips changes during speaking in happy, fear and sadness present certain stability, i.e., they maintain the basic emotional states during the time. Facial and acoustic features of anger are relatively complex and may be inconsistent at different emotion focus, in which, the prosody features could be also affected.

The whole system is composed as figure5. Except for the dynamic visemes, the emotional functional words are also assigned as the basic trigger for emotional facial expression during the procedure of integration with emotional speech synthesis and facial animation. Final results show that the emotional talking head created in paper could generate the vivid audio and visual output.

In the system, the dynamic AU series generation is the core issue. Both dynamic visemes and emotion expression are converted to AU series for final and detailed processing. Emotion expression is highly related to emotional focus information with huge F0 and intensity changing. As mentioned above, Emotional functional words plays an important role in emotional focus detection, while predicting speech acoustic control parameters and AU series parameters under different emotional states can also be processed through the statistics analysis of data.

Table 3: Variations of video and acoustic parameters in four emotional states

Emotion State	**Happy**	**Fear**	**Anger**	**Sad**
Lip shape	AU12+25	AU25	AU17+23+24	AU12+25
Eye region	AU4	Normal	AU1+4	AU4
Speech rate	Faster, but sometimes slower	Much faster	Slightly faster	Slightly slower
F0 mean	Much higher	Very much higher	Very much higher	Slightly slower
F0 range	Much wider	Much wider	Much wider	Slightly narrower



(a) System framework

(b) System interface

Figure 4, Emotional Taking Head System

5. Conclusion

Bimodal emotional expression based on voice and face is a relatively new research region of multimodal human-machine interaction and it emphasizes particularly on complex facial expressions analysis and processing under various voice overloads. The paper preliminarily analyzes and summarizes the synchronous characteristics of voice and face from continuous video streams and proposes a bimodal audio-video emotional expression prototype system combined with realistic 3D facial emotional expression. Since the facial feature variation is still dynamic and complex even in the same emotional state during speaking and large scale database is required for more detailed dynamic modeling approaches, related research work is expected in the future.

References

1. Black, M. and Yacoob, Y. Recognizing facial expressions in image sequences using local parameterized models of image motion. International Journal of Computer Vision, 1997, 25, 23-49.

2. J. Cahn, “Generating Expression in Synthesized Speech,” Master’s thesis, MIT, 1989.

3. Essa, I.A. and Pentland, A. A vision system for observing and extracting facial action parameters. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’94).

4. Ekman, P. & Friesen, W.V. Facial action coding system. Palo Alto: Consulting Psychologist Press, 1978

5. Katz, G.S., Cohn, J.F., & Moore, C.A. A Combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant directed speech. Child Development,1996, 67, 205-217.

6. Murray, I.R. & Arnott, J.L. Toward the simulation of emotion in synthetic speech: A review of the literature on human emotion. Journal of the Acoustical Society of America, 1993(2), 1097-1108.

7. Otsuka, T. & Ohya, J. Spotting segments displaying facial expression from image sequences using HMM. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, 442-447.

8. MPEG Video, Information technology – Coding of audio-visual objects – Part 5: Reference software, Amendment 1: Reference software extensions, ISO/IEC JTC 1/SC 29/ WG 11/N3309, March, 2000

9. Jianhua Tao, Emotion Control of Chinese Speech Synthesis in Natural Environment, Eurospeech2003, Genever, 2003,9

10. Schröder M, Emotional Speech Synthesis: A Review, Eurospeech2001

11. Hadap and Thalmann ,Fluid flow and vector field,EGCAS 2000

12. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature, 1976, 264(5588): 746~748

13. 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).

14. Rabi G, Si Wei Lu. Energy minimization for extracting mouth curves in a facial image. In: Proceedings, Intelligent Information Systems, IIS '97, 1997, 381~385

15. Chiou G I, Jenq-Neng Hwang. Lipreading from color video. IEEE Transactions on Image Processing, 1997, 6: 1192~1195

16. Lievin M, Delmas P, Coulon P Y, et al. Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme. In: IEEE International Conference on Multimedia Computing and Systems, 1999. 1:691~696

17. Delmas P, Coulon PY, Fristot V. Automatic snakes for robust lip boundaries extraction. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. 6: 3069~3072

18. Yang J, Xiao J, Ritter M. Automatic selection of visemes for image-based visual speech synthesis. In: IEEE International Conference on Multimedia and Expo, 2000, 2: 1081~1084

19. Bothe H H, Frauke R. Visual speech and coarticulation effects. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993. 5: 634~637

20. Breen A P, Bowers E, Welsh W. An investigation into the generation of mouth shapes for a talking head. In: Fourth International Conference on Spoken Language (ICSLP 96), 1996. 4: 2159~2162

21. Cohen M M, Massaro D W. Modeling coarticulation in synthetic visual speech. In: Models techniques in computer animation, Tokyo Springer-Verlag, 1993, 139~156

22. Morishima S, Aizawa K, Harashima H. An intelligent facial image coding driven by speech and phoneme. In: International Conference on Acoustics, Speech, and Signal Processing, 1989. 3: 1795~1798

23. Ezzat T, Geiger G, Poggio T. Trainable Videorealistic Speech Animation. In: Proceedings of ACM SIGGRAPH 2002, San Antonio, Texas, 2002

24. Keikichi Hirose, Nobuaki Minematsu, etc, “Analytical and perceptual study on the role of acoustic features in realizing emotional speech”, ICSLP2000

25. Schröder M1, Cowie R2, Douglas-Cowie E2, “Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis”, Eurospeech2001

26. Kazuhito Koike, Hirotaka Suzuki, Hiroaki SAITO, “Prosodic Parameters in Emotional Speech”, ICSLP98

27. J.M. Montero, J. Gutiérrez-Arriola, etc, “Emotional speech synthesis: from speech database to tts”, ICSLP98

28. Murray, I. R.; Arnott, J.L.: 1992, 'Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion'. Journal of the acoustic society of America, *93*, 1097-1108.

29. Williams, C.E.; Stevens, K.N.: 1981, 'Vocal correlates of emotional states'. In: J.K.Darby (eds.): Speech evaluation in psychiatry, New York, Grune & Stratton, pp. 221-240.

30. Ignasi Iriondo, etc, Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques, ISCA Workshop on Speech and Emotion, Belfast 2000

31. Akemi Iida, Nick Campbell,etc, A Speech Synthesis System with Emotion for Assisting Communication, ISCA Workshop on Speech and Emotion, Belfast 2000

32. Ze-Jing Chuang and Chung-Hsien Wu, “Emotion recognition from textual input using an emotional semantic network”, ICSLP2002, Denver

33. Parke F. Computer generated animation of faces. In: Proceedings of the ACM annual conference, Boston, USA, 1972, 1: 451~457

34. Parke F. Computer graphic models for the human face. In: The IEEE Computer Society's Third International Computer Software and Applications Conference, 1979, 724~727

35. Magnenat-Thalmann N, Primeau E, Thalmann D. Abstract muscle action procedures for human face animation. Visual Computer, 1988, 3(5): 290-297

36. Ezzat T, Poggio T. Video Realistic Talking Faces: A Morphing Approach. In: Proceedings of the Audiovisual Speech Processing Workshop, Rhodes, Greece, 1997

37. Graf H P, Cosatto E, Ezzat T. Face analysis for the synthesis of photo-realistic talking heads. In: 4th International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000. 189~192