Emotional Chinese Talking Head System

Jianhua Tao
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science,
P.O.X 2728, Beijing, China, 100080
+86-10-82614523
jhtao@nlpr.ia.ac.cn

Tieniu Tan
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science,
P.O.X 2728, Beijing, China, 100080
+86-10-82614523
tnt@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 emotional speech synthesis and emotional facial animation are integrated and the generic model. To express emotion, both emotional speech synthesis and emotional facial animation are integrated and Chinese viseme models are also created in the paper. Finally, the emotional talking head system is created to generate the natural and vivid audio-visual output.

Keywords

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 visemes, 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.

In facial animation, a person can express an emotion state by slightly altering some facial features such as frowning. Ekman[32] brought forward a model with an organic notation of facial expressions FACS (Facial Action Coding System) and AU (action unites) to describe or control the face action to express emotions. There are also some methods based on photo-realistic[37] or musical movement [35]. The generation of facial expressions with various utterances is mostly realized by local deformations of thread model. To make efficient computing, we use FFD (free-form deformation) algorithm to simplify the interpolation process.

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, section II describes the FFD method for facial animation. Texture mapping and dynamic Chinese visemes are also analyzed and described here. In section III, 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 IV 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. FACIAL EXPRESSION

2.1 Animation

Generally, face models can be grouped into two categories named as thread model and surface model. As for thread model, it is simple and efficient to organize all the 3D points by multi planar polygons (usually triangles). Thus, different resolutions can be obtained in different detailed regions and thread model can adapt to the huge difference of texture and deformation between the obtained in different detailed regions and thread model can adapt polygons (usually triangles). Thus, different resolutions can be to different regions and other regions in the face. For the resolution above, the paper takes thread model as the basis face model. The generation of facial expressions with various utterances is mostly realized by local deformations of thread model. We use FFD (free-form deformation) algorithm to simplify the interpolation process. In this algorithm, original points in images are firstly selected to be moved to target points, meanwhile, characteristic linear muscles and characteristic contour muscles of the feature lattice move and so do characteristic linear muscles and characteristic contour muscles of the whole feature lattice. In practice, the deformation is the mapping between coordinates of image pixels, and interpolation is used to get the gray value of the coordinate.

First, we assign \( P = FFD(R) \) as the geometric model of the local deformation box, which is shown in figure 1. \( P_0 \) is the source point, orthogonal terms \( s, t \) and \( u \) define the axis \( S, T \) and axis \( U \). \( P \) is a local coordinate of a random point in the lattice and \( P \) is its global coordinate. Then we get:

\[
P = P_0 + sS + tT + uU
\]

Where, \( 0 \leq s \leq 1 \), \( 0 \leq t \leq 1 \), \( 0 \leq u \leq U \)

![Lattice deformation of parallelepiped](image)

The lattice is divided into different segments \( l, m \) and \( h \) in direction \( S, T \) and \( U \). The vertex is defined as:

\[
P_{ijk} = P_0 + \frac{i}{l}S + \frac{j}{m}T + \frac{k}{n}U
\]

Where, \( i = 0, \cdots, l \)
\( j = 0, \cdots, m \)
\( k = 0, \cdots, n \)

There is relationship between outer vertices and inner vertices shown in (3) and \( B_i^j(s) \) \( B_m^k(t) \) \( B_n^p(u) \) are Bernstein polynomials.

\[
P' = \sum_{i=0}^{l} \sum_{j=0}^{m} \sum_{k=0}^{n} P_{ijk} B_i^j(s) B_m^k(t) B_n^p(u)
\]

To make the efficient deformation, weight \( W_{ijk} \) is set for different vertices, so we get:

\[
P' = \sum_{i=0}^{l} \sum_{j=0}^{m} \sum_{k=0}^{n} W_{ijk} B_i^j(s) B_m^k(t) B_n^p(u)
\]

Facial expressions can be generated through local muscle contraction. Normally, to produce subtle changes of the objects, the number of vertex must be increased quite a lot, nevertheless, it is rather time consuming and difficult for control.

To make object deformation, the neighbored vertices have the strong corelation due to the musical connection among them. For any points in line \( P_1 \sim P_2 \), such as \( P \). They will be transformed while the lines changing. If \( P_1 \sim P_2 \) changed to \( P'_1 \sim P'_2 \), then we get new point with linear interpolation,

\[
P' = \frac{P'_2 - P'_1}{P_2 - P_1} (P - P_1)
\]

With linear interpolation (5), all of the points in each line could be calculated to obtain the coordinate mapping during deformation. Similar methods can be also used to process the skin and muscle movement.

2.2 Texture Mapping

In generating the realistic face, the reality of the synthetic face may be reduced due to the distortion of profile textures of the face model. We utilizes the multi-orientation texture mapping algorithm to solve this problem. The whole orientation of each Bézier surface composing the face model is determined and then texture information of face images of corresponding orientation is mapped. When the angle between the whole orientations of surface and frontal face is lower than 45 degree, the texture information of frontal face is used, otherwise, two texture information of profile face are used. Given neural image texture information of frontal and profile face is mapped to the given facial expression model with regard to the corresponding grid topological structure. Furthermore, the high quality realistic 3D face is generated.
3D facial texture, realistic 3D face and its transformation in the smile state are listed in figure 2. In order to solve the collision relation of all the parts of the face such as ears, teeth, tong and hair and achieve a better visual effect, the fluid flow method [11] brought forward by Thalmann is adopted in the paper and a good result is obtained using some handwork interactions.

2.3 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 inconspicuous 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.

| phone | si | an | a | o | ang | ou | e | ei | eng | i | i | ing | u | ü | n | ng | g | k | h | d | t | n | l | ng | f | f | r | q | x | z | c | s | zh | ch | sh |
| mes   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| viseme|    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [a]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [o]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [u]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [e]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [i]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [u]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [d]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [f]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [r]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [q]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [x]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [z]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [c]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [s]   |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [zh]  |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [ch]  |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| [sh]  |    |    |   |   |     |    |   |   |     |   |   |     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

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.

<table>
<thead>
<tr>
<th>speech rate</th>
<th>Happiness</th>
<th>Anger</th>
<th>Sad</th>
<th>Fear</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 Mean</td>
<td>Faster, but sometimes slower</td>
<td>Slightly higher</td>
<td>Slightly lower</td>
<td>Much faster</td>
<td>Very much slower</td>
</tr>
<tr>
<td>F0 Mean</td>
<td>Much higher</td>
<td>Very much higher</td>
<td>Slightly lower</td>
<td>Very much higher</td>
<td>Very much lower</td>
</tr>
</tbody>
</table>
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 sentences related to a certain topic. The modifier words are emotion keywords represent the major emotion reaction of the sentence, such as "very, so, too much, not, etc.". Normally, be influenced by some words which behave as modifier in the environment.

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.

- **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.

- **Metaphor words**

![Figure 4, Lexicon Structure](image)

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 4.

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 \( \alpha \).

\[
\hat{E}_t = \mathcal{D}(\hat{E}_{t-1}) = \hat{E}_{t-1} \exp(-\alpha \times t^3) + \mathcal{C}_t
\]

(6)

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

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

If the node is emotional keyword, \( \hat{E}(t) = (\omega_{t,joy}, \omega_{t,sadness}, \omega_{t,anger}, \omega_{t,apprise}, \omega_{t,hate}, \omega_{t,fear}) \) is the initial emotion weights defined in the lexicon.

If the node is modifier or metaphor word, all of the \( \delta(t) \) should be multiplied by coefficient \( \beta \).

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”, 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.

\[
M_t = \arg \max_{s} \left( \sum_{s} e_{s,t} \right)
\]

(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, happiness 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.

| 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     |
The whole system is composed as figure 5. 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.

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.

6. REFERENCES


[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


