Emotional Audio-Visual Speech Synthesis Based on PAD

Jia Jia, *Member, IEEE*, Shen Zhang, Fanbo Meng, Yongxin Wang, and Lianhong Cai, *Member, IEEE*

***Abstract*—Audio-visual speech synthesis is the core function for realizing face-to-face human–computer communication. While considerable efforts have been made to enable talking with computer like people, how to integrate the emotional expressions into the audio-visual speech synthesis remains largely a problem. In this paper, we adopt the notion of Pleasure–Displeasure, Arousal–Nonarousal, and Dominance–Submissiveness (PAD) 3-D-emotional space, in which emotions can be described and quantified from three different dimensions. Based on this new definition, we propose a unified model for emotional speech conversion using Boosting-Gaussian mixture model (GMM), as well as a facial expression synthesis model. We further present an emotional audio-visual speech synthesis approach. Specifically, we take the text and the target PAD values as input, and employ the text-to-speech (TTS) engine to first generate the neutral speeches. Then the Boosting-GMM is used to convert the neutral speeches to emotional speeches, and the facial expression is synthesized simultaneously. Finally, the acoustic features of the emotional speech are used to modulate the facial expression in the audio-visual speech. We designed three objective and five subjective experiments to evaluate the performance of each model and the overall approach. Our experimental results on audio-visual emotional speech datasets show that the proposed approach can effectively and efficiently synthesize natural and expressive emotional audio-visual speeches. Analysis on the results also unveil that the mutually reinforcing relationship indeed exists between audio and video information.**

***Index Terms*—Audio-visual speech, boosting-Gaussian mixture model (GMM), emotion, facial expression, Pleasure–Displeasure, Arousal–Nonarousal, and Dominance–Submissiveness (PAD).**

I. Introduction

With the rapid development of human–computer interaction, there is little doubt that audio-visual speech synthesis is becoming a core function for developing a real-life human–computer communication interface. Emotion, the spirit of expressions, is associated with a wide variety of feelings, thoughts, and behavior. Many human feelings have to be expressed with not only language but also certain (or even accurate) emotions. Therefore, there is a clear need for methods and techniques to integrate the emotions into the audio-visual speech synthesis.

Unfortunately, many existing audio-visual speech synthesis systems [1], [2] failed to convey the affective information as in human communications, due to the lack of emotional expression in synthetic speech and talking face. The advance of research on affective computing reveals that the emotion plays an important role in rational decision making, perception, and human communication [3], which indicates the necessity to introduce emotional capability in designing human–computer speech interface.

Emotion has been studied for a long time in psychology [4]–[6]; however, the expression of human emotion only has gained special attention recently in both audio speech synthesis and talking face animation [7]–[10].

For emotional speech synthesis, many acoustic features, such as pitch variables (F0 level, range, contour and jitter), intensity, and speech rate have already been analyzed [11]. There are also some implementations in emotional speech synthesis [12]–[15]. For instance, Toda realized the conversion from neutral speech into emotional speeches using Gaussian Mixture Model (GMM) with dynamic frequency warping (DFW) [16]. Tao attempts to synthesize emotional speech with "strong," "medium," and "weak" classifications using Linear Modification Model (LMM), GMM, and Classification and Regression Tree (CART) [15]. For expressive talking face animation, most of the previous work focused on discrete emotional categories [17], [18]. In terms of the emotional dimension, Ruttkay *et al.* [19] have proposed the "Emotion Disc" and "Emotion Square" to explore the facial expression within a 2-D emotion space. Albrecht *et al.* [20] adopted the "Activation–Evaluation" emotional dimension to design an algorithm for synthesizing facial expression of pure and mixed emotions of varying intensities. Pelachaud *et al.* [21] created an embodied conversational agent by modeling the emotion with "Valence" (positive-negative) and "Timing" (past, current and future) dimensions, and generating the facial expression that conveys the nonverbal information accompanying speech. The previous works about emotional speech and facial expression synthesis are mostly focused on the simulation of discrete basic emotions, which is incapable of making the audio-visual speech as expressive as human beings.

In general, the challenges for synthesizing more natural and expressive emotional audio-visual speech are the following.

— How to describe the various and numerous human emotions as both accurately and completely as possible?

Manuscript received September 24, 2009; revised December 16, 2009; accepted May 18, 2010. Date of publication June 07, 2010; date of current version December 03, 2010. This work was supported in part by the National Natural Science Foundation of China (90820304), in part by the National Basic Research Program of China (973 Program) (No. 2006CB303101), and in part by the National High Technology Research and Development Program("863"Program) of China (2009AA011905). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yannis Stylianou.

The authors are with the Department of the Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: jiajia@mails.tsinghua.edu.cn; zhangshen05@mails.tsinghua.edu.cn; mfb03@mails.tsinghua.edu.cn; wangyongxin@mails.tsinghua.edu.cn; clh-dcs@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2052246

— Can the emotion be computed in a parametric way together with the audio/visual features?
— The incapability of expressing emotion accurately and completely only by audio or visual modality has been reported in [22], [23]. So how to enhance the emotion expression by the combination of audio speech and talking face?

To address the above challenges, we try to conduct a systematic investigation of the problem. Specifically, we adopt the PAD 3-D-emotional space [24], in which emotions are not limited to isolated categories but can be described and quantified along three independent dimensions: Pleasure–Displeasure (P), Arousal–Nonarousal (A), and Dominance–Submissiveness (D). Based on PAD emotion space description, we propose an approach on emotional audio-visual speech synthesis. Our approach primarily consists of three steps: first we take the text and the target PAD values as input, and employ text-to-speech (TTS) engine to generate neutral speeches. Then a segmental emotional speech conversion model based on PAD using Boosting-GMM is proposed. Next, we present a parametric facial expression synthesis model based on PAD. Finally, the facial expression is modulated in audio-visual speech, using acoustic features of the target emotional speech.

We design eight experiments to evaluate the performance of each model and the overall approach. The final analysis results show the effectiveness of the proposed approach for emotional audio-visual speech synthesis based on PAD, which confirms the necessity of the components in this approach. The experimental results also indicate the mutually reinforcing relationship between audio and video information.

The paper is organized as follows. Section II analyzes the acoustic features of emotional speech, and presents the segmental emotional speech conversion model using GMM and Boosting-GMM. Section III describes the facial expression synthesis model. Section IV introduces the proposed emotional audio-visual speech synthesis approach based on PAD. Section V presents and discusses the objective and subjective experiments employed to evaluate the different models and the overall approach. Finally, Section VI draws a conclusion for this paper.

II. Segmental Emotional Speech Conversion Model Based on PAD

In this section, we propose a segmental emotional speech conversion model based on PAD. We first prove that the segmental features are more effective than global features for emotion conversion by statistical experiments. The global features are extracted from the whole sentence, while the segmental features are extracted from different sentence segments. Then GMM and the Boosting-GMM are tried to predict the emotional speech acoustic features. Finally, TD-PSOLA algorithm [26] is used to convert the neutral speeches to the target emotional speeches.

A. Analysis of Emotional Speech Features

1) Acoustic Feature Extraction: We first analyze the acoustic features of speeches in neutral and four other typical emotions: anger, happiness, surprise, and sorrow. The speeches for analysis are chosen from the emotional speech dataset D_1 which will be described in detail in Section V. There are 45 utterances for each emotion, and the total 225 utterances are read by six different people, four females and two males. The average values of the following features, energy, duration, average pitch, max pitch, min pitch, and pitch range [25], are analyzed here. Energy is calculated with frame length of 20 ms and frame shift of 10 ms. To avoid the problems caused by noise, the features are calculated only in the syllables whose energy passes a certain threshold. Duration is the time from the beginning of the first syllable to the end of the last syllable, including the pauses between syllables.

To distinguish the differences between the features of neutral and other emotions, we compute the ratio (%) between the measurements of the corresponding emotional and neutral utterances. For example, the ratio of average f0 for angry emotion is computed as

$$R_{\mathrm{F0,angry}} = \frac{\sum_{i=1}^{n} F0_{i,\mathrm{angry}}}{\sum_{i=1}^{n} F0_{i,\mathrm{neutral}}} \tag{1}$$

where n indicates the number of utterances, and $F0_i$ indicates the average f0 of utterance i in a certain emotion.

To analyze acoustic features of local segments, one sentence is divided into three parts: head, body, and tail. We choose the first prosodic word of each sentence as head, the last one as tail, and the rest of the sentence as body. The prosodic word boundaries are automatically predicted by the text analysis module of TTS engine. In our TTS system, maximum entropy (ME) model is used for prosodic word boundary prediction [27]. After each sentence is divided into three parts, the segmental acoustic features of all emotions are calculated.

2) Statistical Analysis of Global Acoustic Features: Fig. 1(a) shows the differences of global features between four categories of emotional speeches and neutral speeches. It is shown that the duration of sorrow speech is longer than neutral, while the duration of other emotions are shorter than neutral. The energy of sorrow speech is smaller than four other emotions. The average pitch of angry, happy, and surprise speeches are higher than neutral and sorrow. Although sorrow can be distinguished from other emotions easily using global features, it is hard to distinguish anger, happiness, and surprise.

3) Statistical Analysis of Segmental Acoustic Features: Fig. 1(b) shows the differences between acoustic features of the four non-neutral emotions in head part. We can see the energy of anger is much higher than happiness and surprise. The duration of happiness is longer than anger and surprise, and the pitch range of happiness is narrower than anger and surprise.

Fig. 1(c) shows the differences between acoustic features of the four emotions in body part. It is shown that surprise can be distinguished from anger and happiness with max pitch, min pitch, and pitch range. The duration of happiness is longer than anger and surprise.

Fig. 1(d) shows the differences between acoustic features of the four emotions in tail part. The pitch range of surprise is 2.2 times of neutral. The max pitch of surprise is higher than anger and happiness, and the duration of anger is much shorter than surprise and happiness.

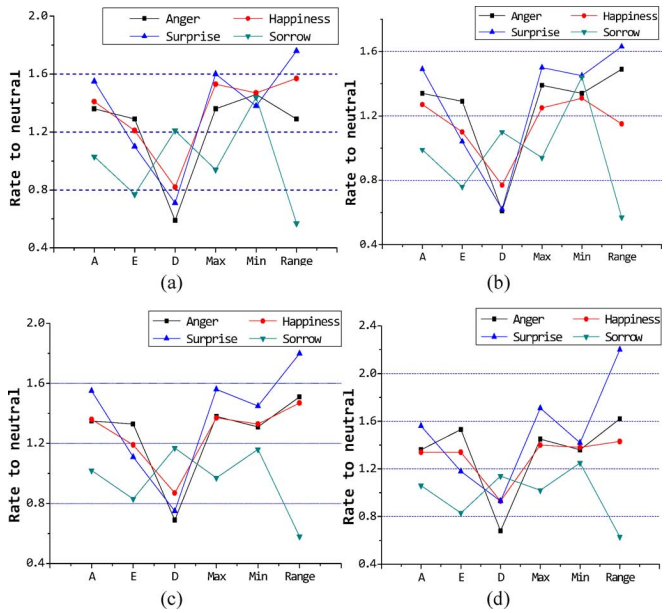


Fig. 1. Global and segmental acoustic features of different emotions A: average pitch; E: energy; D: duration; Max: max pitch; Min: min pitch; Range: pitch range. (a) The global acoustic features of different emotions. (b) The acoustic features of different emotions in head part. (c) The acoustic features of different emotions in body part. (d) The acoustic features of different emotions in tail part.

The statistics show that the anger, surprise, and happiness emotions can be easily distinguished by segmental acoustic features. For anger, the energy is the biggest at all parts of a sentence, while the duration is the shortest. For surprise, the pitch range is wider than other emotions. For happiness, the duration is a bit longer than anger and surprise in head part and body part, and in general, the tail part of a sentence contributes most to the emotion expression.

Based on the analysis, we believe that the segmental strategy using local features is better for emotion speech synthesis. So we propose a new segmental strategy based on PAD for emotion conversion in the next subsection.

B. Segmental Emotional Speech Conversion Model Based on PAD

It has been proved that segmental strategy is more effective than global strategy in Section II-A. In this subsection, we present two models to predict the acoustic features of the target emotional speeches. We use PAD values as the input of the model, and the output is the differences between the acoustic features in emotional speeches and the neutral speeches. After that, TD-PSOLA algorithm [26] is used to convert the neutral speeches to the target emotional speeches.

1) Regression Model Based on GMM With Joint Density Estimation: The acoustic features are related to each other. For example, pitch often rises when energy strengthens, and an emphasized syllable can be realized not only with a higher pitch but also a longer duration. So it is inappropriate to predict the different acoustic features separately as in most traditional ways. In 1998, Kain proposed a regression algorithm using GMM whose parameters are trained with joint density estimation [28]. GMM has the advantage of considering the relationship between the outputs. Kain used GMM in voice conversion and got a good result on a small training set. Tao has built a mapping model from neutral features to basic emotion features using this algorithm [15].

TABLE I
SAMPLE COMPOSITION

Input Variables			Output Variables			
P	A	D	Maximum Pitch Difference	Pitch Range Difference	Duration Difference	Energy Difference

We adopt the algorithm proposed by Kain to build the model which predicts the acoustic feature differences between emotional speeches and neutral speeches based on PAD. Three different regression models are established for head part, body part, and tail part, respectively. The model input is the PAD values represented by a 3-D vector, and the output is the feature differences represented by 4-D vector. As shown in Table I, each training sample consists of the PAD vector (as input variables) and four acoustic feature differences (as output variables). We define the difference as the ratio of the target emotional acoustic feature to that of the original neutral one. For example, the duration difference is defined as $R_{\mathrm{Duration}} = Duration_{Emotional}/Duration_{Neutral}$.

2) New Regression Model Based on Boosting-GMM: To further improve prediction accuracy, ensemble learning method is considered. Ada-Boost is a classic ensemble learning algorithm, which is proposed by Freund and Schapire in 1995 [29], [30]. Its basic idea is learning from failure. We use the idea of Ada-Boost with the algorithm proposed by Kain to build a new ensemble learning algorithm, Boosting-GMM.

Boosting-GMM algorithm contains several weak prediction models. One of them is the basic prediction model and the others are assistant prediction models. The basic model is the regression model for predicting feature differences between emotional and neutral speeches using GMM, which is built in the last part. According to the prediction errors of GMM, we resample the corpus. The number of the samples that have large prediction error is increased, while the number of other samples is decreased. Then assistant prediction models are built based on new corpus. For each acoustic feature, we build an assistant model. This phase can be explained as follows.

Let k represent the size of training set $\boldsymbol{S}$, m represent the number of acoustic features, and M represent the basic model. The algorithm for constructing the ith assistant prediction model based on the prediction error of the ith acoustic feature is as follows.

— calculate the prediction error on the ith acoustic feature of the basic model M, which are $e_{i1}, e_{i2}, e_{i3}, \ldots, e_{ik}$

$$e_{ij} = pre_{ij} - t_{ij} \tag{2}$$

where pre_{ij} is the predicted value of the ith feature of the jth sample from the basic prediction model, and t_{ij} is the real value of the ith feature of the jth sample.

— resample the training set to get a new training set $\boldsymbol{S_i}$, the probability of each sample to be chosen is $p_{i1}, p_{i2}, p_{i3}, \ldots, p_{ik}$, which is calculated as

$$p_{ij} = \frac{e_{ij}}{\sum_{j=1}^{k} e_{ij}}. \tag{3}$$

— build the ith assistant prediction model M_i on the new training set $\boldsymbol{S_i}$.

After that, based on the training set and its corresponding prediction errors, error estimation GMM models are built for each assistant prediction model, which predict estimation errors from PAD. Boosting-GMM is built with the basic prediction model, assistant prediction models and their corresponding error estimation models, using the ensemble learning method. At the predicting phase of the Boosting-GMM, each output feature value is the prediction result of the weak prediction model whose error estimation for that feature is the smallest.

3) Emotional Speech Conversion Using TD-PSOLA: In the last two parts, two regression models based on GMM and Boosting-GMM are built. The model inputs are PAD values, and the outputs are the differences between the target emotional acoustic features and the original neutral ones. In order to convert the neutral speech to the target emotional speech, we use TD-PSOLA algorithm [26] to modify the pitch and duration. TD-PSOLA modifies pitch by adjusting the locations of peaks, and modifies duration by adding or removing periods. It requires three inputs: the original waveform, the original pitch sequence and the target pitch sequence. The output is the target waveform. The original neutral pitch sequence is provided by TTS engine. The target emotional pitch sequence and the target waveform are obtained through *Step1* and *Step2* shown as follows. Finally, the energy of the target waveform is scaled in *Step 3*.

Step 1) We denoted the rendition speech for syllable i with $\mathbf{S}_i(n)$, where n is the pitch point index. $\mathbf{S}_i(n)$ begins at b_i and ends at e_i. $\mathbf{P}_i(n)$ is a vector representing the pitch sequence of the ith syllable of a neutral speech. Vector $\mathbf{D}_i(n)$ represents the corresponding time location of $\mathbf{P}_i(n)$. $P_{\mathrm{Min},i}$ and $P_{\mathrm{Max},i}$ represent the minimum and the maximum pitch values. Let R_{Max}, R_{Range}, R_{Duration} and R_{Energy} represent the difference (ratio) of the maximum pitch, pitch range, duration and energy between emotional speech and neutral speech, which are provided by GMM or Boosting GMM. Then the target emotional pitch sequence $\tilde{\mathbf{P}}_i(n)$ and the corresponding time location $\tilde{\mathbf{D}}_i(n)$ are obtained as

$$\tilde{P}_{\mathrm{Min},i} = (P_{\mathrm{Max},i} \times R_{\mathrm{Max}}) - (P_{\mathrm{Max},i} - P_{\mathrm{Min},i}) \times R_{\mathrm{Range}} \tag{4}$$

$$\tilde{\mathbf{P}}_i(n) = \tilde{P}_{\mathrm{Min},i} + (\mathbf{P}_i(n) - P_{\mathrm{Min},i}) \times R_{\mathrm{Range}}, n\epsilon[b_i, e_i] \tag{5}$$

$$\tilde{\mathbf{D}}_i(n) = \mathbf{D}_i(n) \times R_{\mathrm{Duration}}, n\epsilon[b_i, e_i]. \tag{6}$$

Step 2) TD-PSOLA is applied to modify the pitch sequence and duration. $\mathbf{P}_i(n)$ and $\mathbf{D}_i(n)$ are provided by our TTS engine. The target $\tilde{\mathbf{P}}_i(n)$ and $\tilde{\mathbf{D}}_i(n)$ are obtained by *step 1*. Then the target emotional speech waveform is modified as follows:

$$\tilde{\mathbf{S}}_i(\tilde{n}) = \mathrm{f}\left(\mathbf{S}_i(n), \mathbf{P}_i(n), \mathbf{D}_i(n), \tilde{\mathbf{P}}_i(n), \tilde{\mathbf{D}}_i(n)\right), \quad n\epsilon[b_i, e_i], \tilde{n}\epsilon\left[\tilde{b}_i, \tilde{e}_i\right] \tag{7}$$

where f(.) represents the synthesis function of TD-PSOLA, which are presented in [26], and $\tilde{\mathbf{S}}_i(\tilde{n})$ represents new waveforms, begins at $\tilde{b}_i$ and ends at $\tilde{e}_i$.

Step 3) Energy of $\tilde{\mathbf{S}}_i(\tilde{n})$ is adjusted by scaling with R_{Energy}. Then $\tilde{\mathbf{S}}_i(\tilde{n})$ is smoothed by a Hamming window $\mathbf{W}_i(\tilde{n})$ in preparation for syllable waveform segment concatenation. Finally, the entire emotional speech $\hat{\mathbf{S}}(\tilde{n})$ is generated by concatenating all the syllable waveforms

$$\hat{\mathbf{S}}(\tilde{n}) = \left\{\hat{\mathbf{S}}_1(\tilde{n}), \ldots, \hat{\mathbf{S}}_i(\tilde{n}), \ldots, \hat{\mathbf{S}}_N(\tilde{n})\right\}$$

$$\text{where } \hat{\mathbf{S}}_i(\tilde{n}) = \tilde{\mathbf{S}}_i(\tilde{n}) R_{\mathrm{energy}} \mathbf{W}_i(\tilde{n}), \tilde{n}\epsilon\left[\tilde{b}_i, \tilde{e}_i\right]$$

$$\mathbf{W}_i(\tilde{n}) = 0.54 - 0.46\cos\left(\frac{2\pi\left(\tilde{n} - \tilde{b}_i\right)}{\tilde{e}_i - \tilde{b}_i}\right), \quad \tilde{n}\epsilon\left[\tilde{b}_i, \tilde{e}_i\right]. \tag{8}$$

III. Facial Expression Synthesis Model Based on PAD

Previous research on facial expression mainly focused on the basic emotion categories [5]. Other studies have used two emotion dimensions to synthesize facial expression by rules [18]. In this section, PAD emotional parameters are applied to describe emotional facial state, and a three-layered framework for parametric facial expression synthesis is proposed.

A. PAD-PEP-FAP Framework

To synthesize facial expression for emotional state in PAD space, a layered parametric framework is proposed. The PAD is taken as high-level description of emotional state. A set of partial expression parameters (PEP) are proposed as middle-level description to measure the dominant expressive movement in face regions, and the MPEG-4 facial animation parameters (FAPs) [31], [32] are used as low-level descrption to animate a 3-D talking avatar[1]. The definition of PEP is determined based on the previous work on FAPs manipulation [2], [33]. The PEP aims to describe facial expression by the motion patterns of specific facial organs rather than single facial feature points. FAPs related to PEP are defined by the FAP subgroup in MPEG-4 standards [31], [32]. The details of PEP definition and related FAPs are shown in Table II.

B. PAD-PEP-FAP Mapping Model

The images from JAFFE expression database [34] are taken as the training set to get the PAD-PEP-FAP mapping model. For each facial expression sample, PAD values are annotated using the 12-item questionnaire [35] by three annotators. Before

[1]The talking avatar "*JingJing*" is jointly developed by Tsinghua University and Chinese University of Hong Kong

TABLE II
DEFINITION OF PEP PARAMETERS AND RELATED FAP GROUPS

PEP code Left/Right	Partial Expression Description PEP ∈[0,-1]	PEP ∈ [0,1]	Related FAP Group (F*i* refers to the *i*th FAP)
1.1(L/R)	Lower Eyebrow	Raise Eyebrow	（F33）,F31,F35,F34,F32,F36
1.2(L/R)	Relax Eyebrow	Squeeze Eyebrow	（F37）,F38
1.3(L/R)	the shape of "\ /"	the shape of "/ \"	（F31）,F35,F33,F32,F36,F34
2.1(L/R)	Close eye-lid	Open eye-lid	（F19）,F21,F20,F22
2.2(L/R)	(Eyeball) right	(Eyeball) left	（F23）,F24
2.3(L/R)	(Eyeball) up	(Eyeball) down	（F25）,F26
3.1	Close mouth	Open mouth	（F5）,F10,F11,F52,F57,F58, F4,F8,F9,F51,F55,F56
3.2	Mouth-corner bent down	Mouth-corner bent up	（F12）,F13,F59,F60
3.3	Mouth sipped	Mouth protruded	（F16）,F17
3.4	Mouth stretch in	Mouth stretch out	（F6）,F7,F53,F54
4.1	Jaw move up	Jaw lower down	（F3）
4.2	Jaw move right	Jaw move left	（F15）

TABLE III
12-ITEM PAD QUESTIONNAIRE FOR EXPRESSION ANNOTATION AND EVALUATION

#	Question	-4	-3	-2	-1	0	1	2	3	4	Question
Q1	Angry										Activated

Q2: Wide awake – Sleepy; Q3: Controlled – Controlling;
Q4: Friendly – Scornful; Q5: Calm – Excited;
Q6: Dominant – Submissive; Q7: Cruel – Joyful;
Q8: Interested- Relaxed; Q9: Guided – Autonomous;
Q10: Excited – Enraged; Q11: Relaxed – Hopeful; Q12: Influential - Influenced

annotation, the annotators are trained by psychological experts on how to use the PAD questionnaire. During annotation, the annotators are required to describe the facial expression with the 12 pairs of emotional words, each of which are just like two ends of a scale, shown in Table III.

The annotators should choose one of them that better describes the facial expression with a 9 level score varying from -4 to $+4$. The original P, A and D values are then calculated from this questionnaire based on

$$\begin{aligned} P &= \frac{(\mathrm{Q1} - \mathrm{Q4} + \mathrm{Q7} - \mathrm{Q10})}{16} \\ A &= \frac{(-\mathrm{Q2} + \mathrm{Q5} - \mathrm{Q8} + \mathrm{Q11})}{16} \\ D &= \frac{(\mathrm{Q3} - \mathrm{Q6} + \mathrm{Q9} - \mathrm{Q12})}{16}. \end{aligned} \tag{9}$$

The facial expression database with PAD and PEP manually annotations is used to train a PAD-PEP mapping model. As the experimental result reveals, the second order polynomial function [(10)] achieve the best fitting result [33]

$$\boldsymbol{PEP} = \alpha \boldsymbol{E}^2 + \beta \boldsymbol{E} + \delta \tag{10}$$

where $\boldsymbol{PEP}$ is the PEP configuration of static facial expression, $\boldsymbol{E}$ is the corresponding PAD annotation, and $\boldsymbol{E}^2$ is a vector in which each element is the square value of its counterpart in $\boldsymbol{E}$, i.e., $[P^2, A^2, D^2]$. α and β are the corresponding coefficient matrix, δ is a constant offset vector. We adopt the k-fold cross validation training process ($k = 10$) to obtain the function coefficients using the least square errors estimation method. The function with the average fitting performance among all ten iterations is chosen as the final result.

The PEP is translated to FAP by a linear interpolation function based on the previous study on FAPs correlation analysis [36], which aims to interpolate the unknown FAP values from decided FAPs. Based on FAPs correlation property, we defined a linear PEP-FAP translation function shown in (11). For the ith PEP in region $R(P_i^R)$, we defined a key-FAP (F_k) which has the highest correlations with other FAPs in regions R as reported in [36]. The FAP with parenthesis in Table II is the key FAP of each subset. The value of F_k is linearly determined by P_i^R directly with a bound of $F_k^{\max}$. The value of non-key FAP (F_j) is linearly interpolated by the key-FAP (F_k) with a coefficient α_k^j. The $F_k^{\max}$ and α_k^j are both experimentally determined

$$\begin{cases} F_k = P_i^R \bullet F_k^{\max} & (P_i^R \in [-1, +1]) \\ F_j = \alpha_k^j \bullet F_k & (\alpha_k^j \in [-1, +1],\ \ k \neq j) \end{cases}. \tag{11}$$

IV. EMOTIONAL AUDIO-VISUAL SPEECH SYNTHESIS APPROACH BASED ON PAD

In this section, we will introduce the overall architecture of our emotional audio-visual speech synthesis approach. The Boosting-GMM and the segmental strategy are chosen in emotional speech conversion model.

A. Overall Approach

As in typical TTS synthesis applications, we use text as input of this approach as well as the target PAD. The emotional audio-visual speeches are synthesized by the following steps.

Step 1) Synthesize the neutral speech using TTS engine. TTS engine can also provide the information of Pin Yin, pitch sequence, syllable duration, and the prosodic word/phrase boundaries.

Step 2) Extract the neutral acoustic features of each segment, including pitch contour information and the short-time energy of syllables.

Step 3) Predict the emotional acoustic features for different segments using Boosting-GMM according to the target PAD values.

Step 4) Modify the acoustic features of the neutral speech using TD-PSOLA algorithm to obtain the emotional speech.

Step 5) Synthesize facial expression using PAD-PEP-FAP model. The acoustic features of the emotional speech predicted in *Step 3* are used to modulate the facial expression in continuous speech, which will be discussed in the next subsection.

Step 6) Generate the viseme based on a previous work on Chinese dynamic viseme proposed by our lab [37]. A total of 20 Chinese static viseme categories are defined. A weight blending dynamic viseme model [37] is proposed to describe the viseme dynamics in coarticulation context, and speed or pause duration change in spontaneous speaking. Based on MPEG-4 facial animation framework, we extract the FAPs

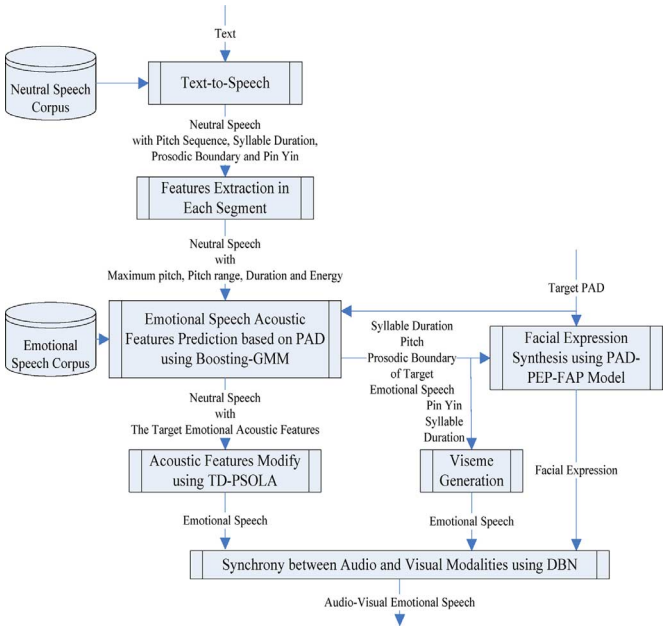


Fig. 2. Flowchart of the emotional audio-visual speech synthesis approach.

of lip movement from videos for each of the Chinese phoneme. The static viseme set is determined by constructing the visual confusion tree [37] based on a measure of normalized visual distances of each phoneme, and the number of viseme classes is determined by this confusion tree.

Step 7) Solve the synchrony problem between audio and visual modalities with a dynamic Bayes network (DBN). We proposed a DBN-based audio-visual correlative model (AVCM) [38], [39], where the loose timing synchronicity between audio and video streams is restricted by word boundaries. The model inputs are the predicted emotional acoustic features and the FAPs related to the viseme. The previously-trained DBN based AVCM is applied to calculate the probability score, which is used as a measure of synthesis error. The downhill simplex method is used to adjust the FAPs until we got the smallest synthesis error. After the synchronization between emotional speech and its viseme, the FAPs related to viseme and facial expression in mouth region are combined, which will be discussed in Section IV-C of this section.

The overall architecture of our approach is shown in Fig. 2.

B. Facial Expression Modulation in Audio-Visual Speech

To synthesize the dynamic facial expression in continuous speech, the speech acoustic features (e.g., pitch) are taken as important cues to modulate the facial expression (i.e., PEP) on sentence level. In our emotional audio-visual speech database, it is found that speakers tend to keep a certain steady facial expression while speaking with a particular emotion, and when such emotions get aroused, the typical expression will be consequently intensified, sometimes even to an exaggerated extent.

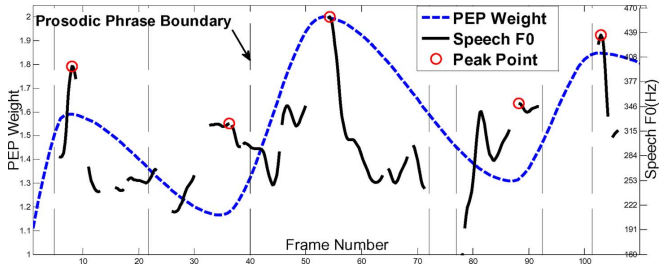


Fig. 3. PEP interpolation weight over a sentence based on F0 peak.

Based on such observation, a two-step facial expression modulation process is proposed. First, the PAD-PEP mapping function described in Section III is utilized to obtain the static facial expression (i.e., PEP) based on the annotated PAD values of each input sentence, then the static facial expression is taken as the dominant facial state over each sentence. Second, the speech pitch (f0) is extracted and applied to locate the "*peak*" time (i.e., local maximum of f0) where the facial expression will be intensified in most cases. The number of "*peaks*" in a sentence is limited to one per prosodic phrase, and each prosodic phrase usually contains 4–8 syllables in Chinese.

The PEP value for intensified expression at peak time is then calculated as (12), while the dynamic facial expression over the whole sentence is obtained by interpolating the PEP values at peak time using the Piecewise Cubic Hermite Interpolating method [40]

$$\boldsymbol{PEP}^{\mathbf{peak}} = \alpha^{\boldsymbol{c}} \bullet \boldsymbol{PEP}^{\mathbf{main}} \quad \left(\alpha = \frac{\boldsymbol{F0}^{\mathbf{peak}}}{\boldsymbol{F0}^{\mathbf{average}}}\right) \tag{12}$$

where $\boldsymbol{PEP}^{\mathbf{main}}$ is the PEP value obtained by PAD-PEP mapping function, α *is* the ratio between peak and average value of speech pitch (F0), c is a constant which controls the interpolated curve of PEP.

It should be noticed that the interpolation only takes place for PEP of non-mouth region. The mouth movement is mainly controlled by the viseme parameters. We selected 24 FAPs (F3–F14, F16, F17, F51–F60) related to the mouth movement to describe and quantify the Chinese viseme [37]. For speaking animation, the viseme is mapped to its corresponding FAP value, and then it controls the face model to perform articulation movement. The 3-D facial animation is in accordance with the MPEG-4 facial animation and implemented based on the XFace toolkit [41]. The interpolation between expression and viseme is discussed in the next subsection. Fig. 3 illustrates the interpolation weight of PEP 1.2 (eyebrow squeeze shown in Table II) for a sentence with anger emotion. Fig. 4 presents the synthetic facial expression series and its counterpart in video data.

C. Audio-Visual Emotional Speech Generation

Since the mouth movement plays an important role in both speaking and expression, we should consider the effect of viseme and expression equally when animating the mouth. The final step for generating expressive visual speech is to properly combine the viseme and expressive facial movement. The viseme is synthesized based on the previous research on Chinese dynamic viseme [37] by our lab, and we take a linear

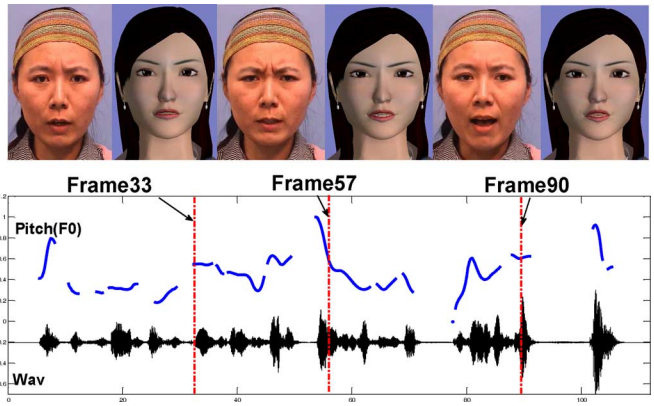


Fig. 4. Synthetic emotional audio-visual speech and its counterpart in real video for a sentence of anger (P $= -0.69$, A $= 0.51$, D $= 0.11$).

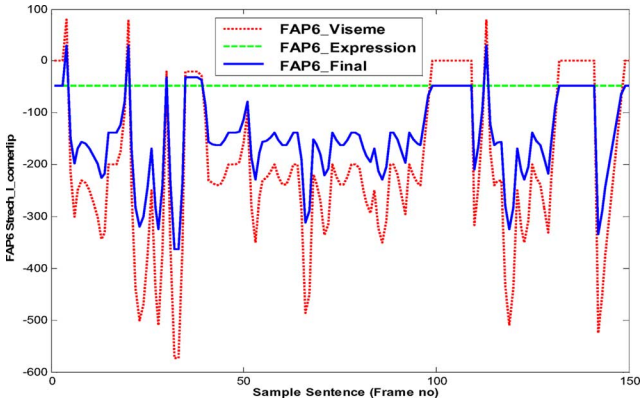


Fig. 5. Viseme and expression mergence (F6).

weighted function to merge the animation parameter of viseme and facial expression in mouth region

$$\boldsymbol{FAP}_{\text{final}}^{\text{mouth}} = \alpha \boldsymbol{FAP}_{\text{viseme}}^{\text{mouth}} + (1-\alpha)\boldsymbol{FAP}_{\text{expr}}^{\text{mouth}}. \quad (13)$$

The coefficient α can be manually defined to determine which is the dominant factor for animating mouth, the viseme or the expression. In our experiment, we take $\alpha = 0.8$ to obtain the final animation parameters for mouth. Fig. 5 shows the curves of No.6 FAP (F6, Strech_left_cornerlip) by combining viseme and expression, where the dashed line denotes the value of F6 in expression, the dotted line refers to F6 in viseme, and the solid line is the merged result of F6.

V. Experiments and Discussions

To test our proposed approach, we conduct extensive experiments on two emotional speech datasets. The experimental results validate the effectiveness of our approach. We design a full-scale evaluation method which includes three objective evaluations and five subjective evaluations. The objective evaluations are used to evaluate the accuracy of the segmental emotional speech conversion model based on PAD using Boosting-GMM, while the subjective evaluations are used to evaluate the effectiveness of our emotional audio-visual speech synthesis approach.

For subjective evaluations, we invite ten participants. All of them are Ph.D. or Masters candidates at Tsinghua University.

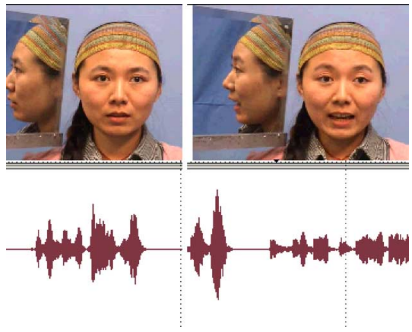


Fig. 6. Data sample of "Surprised" on D_2.

Next, we first introduce our emotional speech datasets and the quality of the datasets. Then we evaluate the emotional speech conversion models, by comparing the performance of GMM and Boosting-GMM, also the segmental conversion strategy and the global conversion strategy. After that, a subjective experiment is conducted to evaluate the PAD-PEP-FAP facial expression synthesis model. Finally, we give the evaluation results of our emotional audio-visual speech synthesis approach, including a mean opinion score (MOS) experiment, and two other experiments which show the mutually reinforcing relationship between audio and video information.

A. Datasets

1) Dataset Setup: We have established two emotional speech datasets. D_1 has 495 emotional speech sentences. The details of the D_1 are shown as follows:

— 45 different Chinese texts, without emotion tendency;
— 11 emotions: Neutral, Relax, Submissiveness, Surprise, Happiness, Disgust, Contempt, Fear, Sorrow, Anxiety, and Anger;
— read by six different people, four females and two males.

D_2 has 132 emotional audio-visual speeches. A data sample of D_2 is shown in Fig. 6. The details of the audio-visual speech corpus D_2 are shown as follows:

— 2 different Chinese texts, without emotion tendency;
— 11 emotions: Neutral, Relax, Submissiveness, Surprise, Happiness, Disgust, Contempt, Fear, Sorrow, Anxiety, and Anger;
— read by six different people, four females and two males.

We invite three human labelers to annotate each emotional speech and audio-visual speech with PAD values and emotion category, respectively. Each of the labelers is required to finish a 12-item questionnaire [35] for each speech to obtain the PAD, and then the PAD values were normalized to $[-1, 1]$. Both of these datasets can be used for training the models, and also as the standard references in subjective evaluations.

2) Quality of the Datasets: Because the datasets are provided by three human labelers, we wish to evaluate the quality of these datasets. In this subsection, we analyze the consistency between the three labelers in the datasets. For each emotion category, Table IV shows the mean value and the standard deviation of labeled PAD on D_1, and Table V shows the statistical results on D_2.

In Table IV, it seems that the PAD values of some emotion categories have large standard deviations. This is because

TABLE IV
MEAN VALUE AND STANDARD DEVIATION OF PAD ON D_1

Emotion	Mean Value			Standard Deviation		
Category	P	A	D	P	A	D
Neutral	0.0031	-0.32	-0.12	0.051	0.23	0.21
Relax	-0.0085	-0.61	0.074	0.18	0.32	0.44
Submissiveness	0.23	-0.22	-0.26	0.18	0.41	0.31
Surprise	0.11	0.57	0.20	0.41	0.20	0.42
Happiness	0.53	0.63	0.40	0.49	0.23	0.35
Disgust	-0.36	-0.25	0.50	0.10	0.48	0.27
Contempt	-0.41	0.28	0.46	0.18	0.21	0.22
Fear	-0.14	0.54	-0.20	0.28	0.17	0.67
Sorrow	-0.14	-0.22	-0.28	0.16	0.24	0.56
Anxiety	-0.29	0.53	0.17	0.39	0.13	0.64
Anger	-0.80	0.61	0.90	0.12	0.13	0.14

TABLE V
MEAN VALUE AND STANDARD DEVIATION OF PAD ON D_2

Emotion	Mean Value			Standard Deviation		
Category	P	A	D	P	A	D
Neutral	-0.24	-0.46	-0.19	0.02	0.15	0.12
Relax	-0.02	-0.61	-0.29	0.11	0.08	0.10
Submissiveness	0.04	-0.55	-0.72	0.08	0.10	0.17
Surprise	-0.06	0.46	0.42	0.18	0.06	0.16
Happiness	0.48	0.25	0.29	0.04	0.08	0.18
Disgust	-0.46	0.08	0.06	0.13	0.12	0.14
Contempt	-0.59	0.34	0.15	0.09	0.14	0.15
Fear	-0.36	0.18	-0.80	0.12	0.09	0.09
Sorrow	-0.50	-0.57	-0.73	0.17	0.14	0.07
Anxiety	-0.52	0.23	-0.43	0.13	0.07	0.19
Anger	-0.75	0.36	0.45	0.05	0.06	0.17

people may have perception confusion about some emotions by only listening to the audio. For example, it is hard to distinguish fear and anxiety only by audio, so the labelers may give PAD values to these two kinds of emotional speeches in a large range, and for D_2, the three labelers have a general consistency in annotating the emotional audio-visual speeches. Therefore, we used the mean PAD values of the three labelers as the PAD of a speech.

B. Evaluation of Segmental Emotional Speech Conversion Model Based on PAD Using Boosting-GMM

1) Comparison Between GMM and Boosting-GMM: Most previous works focus on the emotional speech conversion from neutral to certain emotion categories or degrees [12]–[16]. Our feature prediction model is used in continuous emotion space. So it is hard to compare the accuracy of our model with the previous models. Therefore, in this experiment, we compare three emotional acoustic feature prediction models based on Boosting-GMM, GMM, and Polynomial Regression Model (PRM). The experiment is conducted on D_1.

For both GMM and Boosting-GMM, we build three different regression models for head part, body part, and tail part respectively as described in Section II-B. For PRM, 12 models are built, each of which is used to predict one of the acoustic features (maximum pitch, pitch range, duration, and energy) in one segment (head part, body part, or tail part). The input of each PRM model is the PAD values, and the output is the difference of the corresponding acoustic feature between emotional speech and neutral speech.

TABLE VI
ACCURACY OF PRM, GMM, AND BOOSTING-GMM

	%	Max Pitch	Pitch Range	Energy	Duration
Head	PRM	79.3	67.9	93.2	63.7
	GMM	83.3	71.2	95.3	67.4
	Boosting-GMM	83.4	74.1	95.3	76.3
Body	PRM	82.6	68.2	92.5	87.5
	GMM	84.8	71.1	94.9	91.9
	Boosting-GMM	87.0	72.5	94.4	91.5
Tail	PRM	79.2	68.7	91.2	83.3
	GMM	84.7	72.1	92.6	85.3
	Boosting-GMM	84.9	74.9	92.5	85.4

The experiment is conducted by the method of 10-folder cross validation. The experiment corpus is evenly divided into ten parts. All the feature prediction models run ten times. Each time, nine of the ten parts are used as training set, and the other one is used as the test set. The prediction accuracy A is computed as

$$A = 100 \times \left(1 - -\frac{1}{10}\sum_{j=1}^{10}\frac{1}{s_j}\sum_{i=1}^{s_j}\left|\frac{t_{ij} - t'_{ij}}{t_{ij}}\right|\right) \tag{14}$$

where s_j is the size of the jth testing set, t_{ij} is the real feature difference between emotion and neutral speeches in the ith sample from the jth testing set, and t'_{ij} is the corresponding predicted feature difference. Table VI shows the average feature prediction accuracy of PRM, GMM, and Boosting-GMM for each of the three segments.

The results show that both of the GMM and Boosting-GMM achieve acceptable average feature prediction accuracy. It indicates that establishing a mapping model between PAD values and acoustic features is feasible. Both GMM and Boosting-GMM achieve higher prediction accuracy than PRM. This result demonstrates the advantage of GMM and Boosting-GMM, which considers the relationship between the predicted acoustic features. From Table VI, we can also see that for duration in head part, the prediction accuracy of GMM is considerably low. It increases 8.9% using Boosting-GMM instead of GMM, and the prediction accuracy of pitch range in head and tail parts increase 2.9% and 2.8%, respectively. The prediction accuracy is improved because Boosting-GMM implements a re-sampling process, which increases the proportion of the samples with large prediction errors in training set. On the other hand, the features with higher prediction accuracy by GMM have made less improvement. That is because there are few samples with large prediction errors for these features, such as duration and energy. Considering that for most features, Boosting-GMM performs better than GMM, we choose Boosting-GMM as the feature prediction model in our approach.

2) Comparison Between Segmental and Global Conversion Strategy: The emotional speeches converted using global features and segmental features are compared in this subsection. A subjective evaluation and an objective evaluation are conducted. Both the global features and the segmental features are predicted by the Boosting-GMM.

In subjective evaluation, we select 16 neutral speeches and 16 emotional speeches of the same text from D_1. The emotional speeches are from four typical emotions: anger, happi-

TABLE VII
RESULTS OF THE PREFERENCE TEST

	Picking Global	Picking Segmental
Anger	12.5%	87.5%
Happiness	22.5%	77.5%
Surprise	15.0%	85.0%
Sorrow	35.0%	65.0%

ness, surprise, and sorrow. They are used as the criteria data. For each emotional speech, we record its PAD values, and each neutral speech is converted to the emotional speech according to the corresponding PAD values by two strategies, respectively. We design a preference test in which two converted emotional speeches with the same PAD values but by different strategies are played with random order. Ten participants are invited to pick the one which expresses a more similar emotion to the criteria speech. There are 16 pairs of converted emotional speeches in total, and Table VII shows the average percentage of the participants who picked the sentences converted by a certain strategy for different emotion categories.

From the results we can find that 87.5% participants feel segmental strategy is better than global for anger emotion expression, and 85% participants feel segmental is better for surprise emotion expression. For happiness and sorrow expression, more than 65% participants think the segmental strategy is better. The results show that the segmental strategy is more effective in emotion expression than global strategy significantly.

To get more statistical comparison results between global and segmental strategies, we conduct an objective experiment to show the average pitch prediction accuracy of the converted speeches by two strategies respectively. The prediction accuracy is computed as (15), and we use the average value of the prediction accuracy of maximum pitch and pitch range as our final result:

$$\text{Accuracy}_{\text{pre}} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}\frac{\left|T_{i,k} - T'_{i,k}\right|}{T_{i,k}} \qquad (15)$$

where n represents the speech number of an emotion category, m represents the number of the segments (1 for global and 3 for segmental), $T_{i,k}$ represents the real pitch value of the emotional speech, while $T'_{i,k}$ represents the predicted pitch value. 180 emotional utterances in D_1 are selected from 4 emotions (Anger, Happiness, Surprise, and Sorrow). There are 45 utterances for each emotion. The results are shown in Table VIII.

Fig. 7 gives an example of the emotional speeches converted by different strategies. The original neutral sentence is "bai3 tuo1 gu2 di3 tai4 ji2," which is a meaningless Chinese sentence without emotion tendency. The target emotion is "surprise." Comparing the two converted emotional speeches, we can see that the acoustic features of the speech converted by segmental strategy are more similar to the target surprise speech. Especially, the speech duration is shorter, the modified pitch range of the last syllable is larger, and the modified pitch range of the first syllable is smaller.

From Table VIII and Fig. 7, we can see that the prediction accuracy of the segmental strategy is generally higher than those of the global strategy. Especially for the emotion of surprise, it achieves about 8% improvement in the segmental strategy. We believe this is because the acoustic features of surprise differ a lot in different segments. For example, the pitch range of surprise is 2.2 times of neutral in the tail part, 1.6 times in the head part and 1.8 times in the body part. Therefore, we can conclude that the segmental strategy is more in line with the human habit of emotional expression.

TABLE VIII
AVERAGE PREDICTION ACCURACY

Emotion	Strategy	Average Prediction Accuracy
Anger	Global	84.90%
	Segmental	86.90%
Happiness	Global	80.00%
	Segmental	84.30%
Surprise	Global	81.60%
	Segmental	89.80%
Sorrow	Global	84.50%
	Segmental	87.60%
Average	Global	82.75%
	Segmental	87.15%

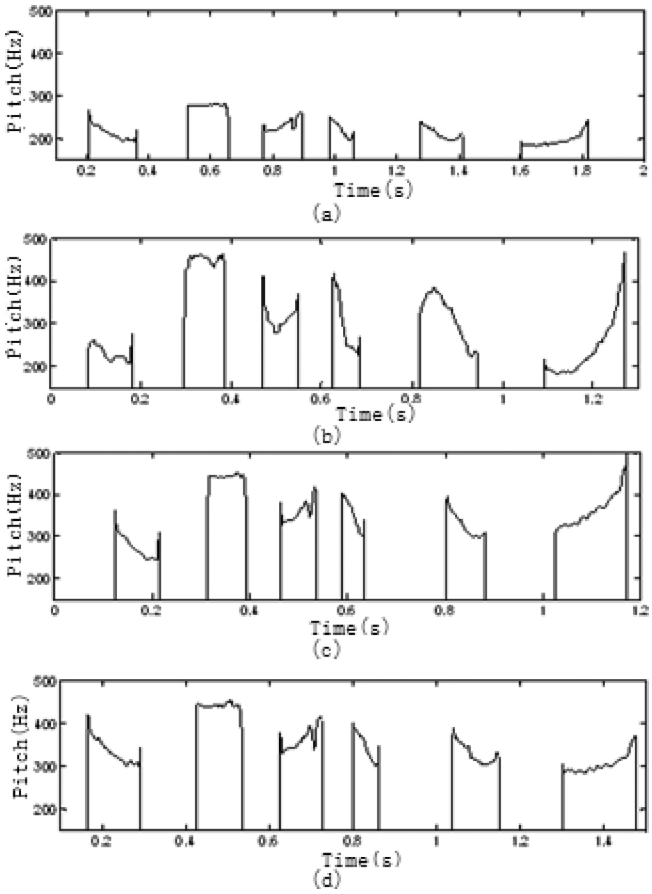


Fig. 7. Example of emotional speeches converted by different strategies. (a) Original neutral speech. (b) Target "surprise" speech. (c) "Surprise" emotional speech converted by segmental strategy. (d) "Surprise" emotional speech converted by global strategy.

C. Evaluation of PAD-PEP-FAP Model

In this subsection, an objective evaluation and a subjective evaluation are conducted to evaluate the performance of the PAD-PEP-FAP facial expression synthesis model.

TABLE IX
PAD EVALUATION ON SYNTHETIC EXPRESSION

	Input PAD			Evaluated PAD		
	P	A	D	P	A	D
Happy	0.55	0.24	0.28	0.42	0.12	0.10
Surprise	0.34	0.34	0.04	0.36	0.45	-0.05
Sad	-0.18	0.03	-0.14	-0.01	-0.26	-0.27
Angry	-0.40	0.22	0.12	-0.17	0.02	-0.08
Disgust	-0.36	0.08	0.13	-0.56	0.15	0.44

The objective evaluation is designed to evaluate the PAD-PEP mapping function. A 3-D facial expression database is created by animating the talking avatar with the PEP parameters extracted from static human facial image. The images from JAFFE expression database [34] are taken as the archetype expressions. The expression database consists of 213 expression samples, which has ten subjects and seven expression categories. PEP parameters are first extracted by measuring the movement of related feature points [by (11)], which are used as the criteria data. Then the PAD-PEP mapping function are used to predict the PEP value based on the annotated PAD values. The correlation between original PEP and its predicted value by PAD is computed getting an average of 0.70.

The subjective evaluation is designed to evaluate the PAD-PEP-FAP mapping model. The PAD values of five typical emotion categories from D_2 are taken as the input of the mapping model, and the output FAP parameters are used to animate the talking avatar to obtain the synthetic expressions. Our participants are invited to finish the 12-item PAD questionnaire [35] to evaluate the synthetic expressions. The average PAD values of emotional categories and synthetic expressions are compared in Table IX, with the correlation of 0.89(P), 0.68(A), and 0.70(D), respectively.

The results are consistent with the reliability and validity of the Chinese version abbreviated PAD emotion scales reported in [35]. Both the objective and subjective experimental results indicate that the PAD model can be used to describe the emotion state as well as facial expression, and the proposed PAD-PEP-FAP mapping model is effective for PAD-driven facial expression synthesis.

D. Evaluation of the Emotional Audio-Visual Speech Synthesis Approach Based on PAD

To demonstrate the effectiveness of our proposed approach to emotional audio-visual speech synthesis, three subjective experiments are conducted in this section. In these experiments, we use the Boosting-GMM and the segmental strategy for emotional speech conversion.

1) Subjective Evaluation on the Overall Approach: At this part, a MOS evaluation is performed on our audio-visual speech synthesis approach. We first randomly select 50 audio-visual utterances from D_2, each of which is annotated with PAD values. Using the PAD values and the corresponding texts as inputs, we synthesize the 50 audio-visual utterances on a 3-D talking avatar with our approach. We invite ten participants to compare the synthetic audio-visual speeches to the recorded speeches from D_2 with the same PAD values and texts. Each synthetic audio-visual speech is scored from 1 to 5 according to its similarity to the corresponding video:

5—the audio-visual speech and the video have the same emotion tendency and degree;
4—the audio-visual speech and the video have the same emotion tendency, and the similar degree;
3—the audio-visual speech and the video have the same emotion tendency, but different degree;
2—the audio-visual speech and the video have the similar emotion tendency;
1—the audio-visual speech and the video have the different emotion tendency.

TABLE X
RESULTS OF THE PAIRED COMPARISON TEST

Emotion Category	Numbers of Scores	Average Score	Standard Deviation
Neutral	4×10	4.03	0.83
Relax	4×10	3.15	1.05
Submissiveness	4×10	2.78	1.12
Surprise	6×10	3.43	0.83
Happiness	6×10	3.58	0.78
Disgust	5×10	3.62	0.85
Contempt	3×10	3.40	0.77
Fear	5×10	3.00	0.92
Sorrow	3×10	3.27	1.14
Anxiety	5×10	3.26	0,77
Anger	5×10	3.76	0.74
Average	45	3.41	0.90

We use the average score to describe the similarity between the emotional audio-visual speeches and the corresponding videos. In Table X, the second column shows the numbers of scores in each emotion category, and the third column gives the average scores. The standard deviations are shown in the fourth column. The average standard deviation is around 1, which is acceptable in a five-mark evaluation. The MOS average score is 3.4, which indicates the proposed approach based on PAD can synthesize a natural and expressive emotional audio-visual speech.

Some typical emotions, such as anger, disgust, and happiness, achieve higher scores than the other emotions. Fig. 8 suggests that the higher scores the emotion categories achieve, the smaller their stand deviations are. We believe that it is because people have more consistent understandings of the typical emotions.

2) Effectiveness of Video for Enhancing Emotion Expression: In this experiment, we demonstrate that using facial expression can help express emotions more accurately than only using speech.

We first synthesize 121 audio-visual speeches on 3-D talking avatar using 121 PAD values of 11 emotions from the corpus D_2. Each participant is asked to annotate each synthetic speech with PAD twice. At the first time, they annotate each speech with PAD by only listening to the audio, and the second time they annotate each speech by watching the video and listening to the audio at the same time.

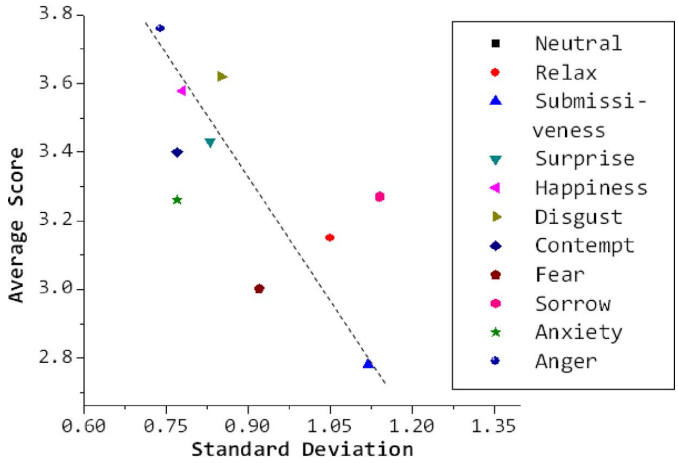


Fig. 8. The correlation between the average score and the standard deviation of each emotion category.

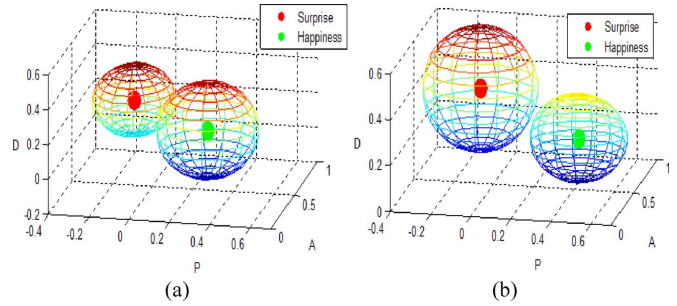


Fig. 9. Overlap ratio of surprise and happiness. (a) Audio. (b) Audio – Visual.

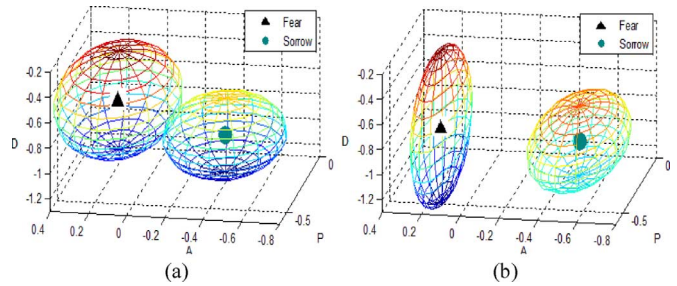


Fig. 10. Overlap ratio of emotions of fear and sorrow. (a) Audio. (b) Audio – Visual.

After that, we draw an ellipsoid to describe the annotated PAD values in the PAD 3-D space for each emotion, with the average PAD values as the centers and the standard deviations as the radii. The overlap ratio of the ellipsoids of each emotion pair is examined.

Fig. 9 shows the ellipsoids of surprise and happiness, and Fig. 10 shows the ellipsoids of fear and sorrow. The results indicate that it is not always easy to tell the accurate emotion only by listening to speeches, such as surprise and happiness, fear, and sorrow, but with the help of proper facial expressions, it is much easier to identify the real emotions. That means the audio-visual speech can achieve a better emotion expressivity.

3) Improving Video Effect Using Audio: In order to demonstrate that the acoustic features of the emotional speeches are also very important to the facial emotional expressions, we conduct a paired comparison experiment on D_2. We first synthesize the audio-visual speeches on 3-D talking avatar using the PAD values of 11 emotions with our proposed approach, and then for the same PAD values, we also synthesize the audio-visual speeches on 3-D talking avatar without using the acoustic features. The listeners determine which audio-visual speech has a higher similarity with the reference video. The final score for an approach is the number of times it is considered the better one in the pair comparisons.

TABLE XI
RESULTS OF THE PAIRED COMPARISON TEST

	Average Score	Confidence Interval (95%)
Our Approach	0.88	[0.72, 1.04]
$A_{without}$	-0.88	[-1.04, -0.72]

The evaluation scale is shown as follows:

— if A is much more similar than B, then A gets 2 points, B gets -2 points;

— if A is a little similar than B, then A gets 1 point, B gets -1 points;

— if A has equal similarity with B, then both A and B get 0 points.

The experimental results are shown in Table XI.

The average score of our approach is 0.88, which indicates that the audio-visual speeches synthesized by our approach have higher similarity with the reference videos. The confidence interval (95%) is [0.72, 1.04]. The results suggest that generating expression using the acoustic features of emotional speech is more in line with the people's habits of expression.

VI. CONCLUSION

In this paper, we have described how emotional audio-visual speeches are synthesized in continuous emotion space. The notion of PAD 3-D-emotional space is used, in which emotion can be described and quantified from three different dimensions. Based on this new definition, we present a novel emotional audio-visual speech synthesis approach. Specifically, the Boosting-GMM is used to convert the neutral speeches to emotional speeches, and the facial expression is synthesized simultaneously, while the acoustic features of the emotional speech are used to modulate the facial expression in the audio-visual speech. The experimental results show that the proposed approach can effectively and efficiently synthesize natural and expressive emotional audio-visual speeches. Analysis on the results also unveil that the mutually reinforcement relationship indeed exists between audio and video information.

ACKNOWLEDGMENT

The authors would like to thank Prof. H. Meng, for her direction of the PAD-PEP-FAP facial expression synthesis model, and they would also like to thank the students of the Institute of Psychology at CAS and the HCSI Lab at Tsinghua University for their cooperation with the dataset setup and experiments.

REFERENCES

[1] E. Cosatto, J. Ostermann, H. P. Graf, and J. Schroeter, "Lifelike talking faces for interactive services authors," *Proc. IEEE, Special Iss. Human–Computer Multimodal Interface*, vol. 91, no. 9, pp. 1406–1429, Sep. 2003.

[2] Z. Y. Wu, S. Zhang, L. H. Cai, and H. M. Meng, "Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 1802–1805.

[3] R. W. Picard, "Affective Computing," MIT Media Lab., Perceptual Comput. Section, Mass. Inst. Technol., Cambridge, MA, Tech. Rep., 1995.

[4] C. Darwin, *The Expression of the Emotions in Man and Animals*. London, U.K.: John Murray, 1872.

[5] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 384–392, Apr. 1993.

[6] J. A. Russell, J. Bachorowski, and J. Fernández-Dols, "Facial and vocal expressions of emotion," *Annu. Rev. Psychol.*, vol. 2002, pp. 329–349, 2003.

[7] M. Schröder, "Emotional speech synthesis: A review," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, vol. 1, pp. 561–564.

[8] E. Bevacqua and C. Pelachaud, "Expressive audio-visual speech," *Comput. Animat. Virtual Worlds*, vol. 15, no. 3–4, pp. 297–304, 2004.

[9] M. Fabri and D. J. Moore, "The use of emotionally expressive avatars in collaborative virtual environments," in *Proc. Symp. Empathic Interact. With Synth. Charact.*, 2005.

[10] H. Tang, Y. Fu, J. Tu, M. Hasegawa-Johnson, and T. S. Huang, "Humanoid audiovisual avatar with emotive text-to-speech synthesis," *IEEE Trans. Multimedia*, vol. 10, pp. 969–981, Oct. 2008.

[11] N. Audibert, D. Vincent, V. Auberg, and O. Rosec, "Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions," in *Proc. Speech Prosody*, 2006.

[12] M. Bulut, S. Lee, and S. Narayanan, "A statistical approach for modeling prosody features using pos tags for emotional speech synthesis," in *Proc. Int. Conf. Acoust. Speech, Signal, Process.*, Honolulu, HI, Apr. 2007, pp. 1237–1240.

[13] J. Cabral and L. Oliveira, "Emovoice: A system to generate emotion in speech," in *Proc. Interspeech*, Pittsburgh, PA, 2006.

[14] N. Campbell, "Synthesis units for conversational speech using phrasal segments," in *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 2004.

[15] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[16] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikamo, "GMM-based voice conversion applied to emotional speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, Nov. 1999.

[17] Y. Du and X. Lin, "Emotional facial expression model building," *Pattern Recognition Lett.*, vol. 24, no. 16, pp. 2923–2934, 2003.

[18] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 10, pp. 1021–1038, 2002.

[19] Z. Ruttkay, H. Noot, and P. Hagen, "Emotion disc and emotion squares: Tools to explore the facial expression space," *Comput. Graphics Forum*, vol. 22, no. 1, pp. 49–53, 2003.

[20] I. Albrecht, M. Schröder, J. Haber, and H. P. Seidel, "Mixed feelings: expression of non-basic emotions in a muscle-based talking head," *Virtual Reality*, vol. 8, no. 4, pp. 201–212, 2005.

[21] C. Pelachaud and M. Bilvi, "Computational model of believable conversational agents," in *Communication in Multiagent Systems*, M. P. Huget, Ed. New York: Springer-Verlag, 2003, vol. 2650, Lecture Notes in Computer Science, pp. 300–317.

[22] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Trans. Speech, Audio,. Speech, Lang. Process.*, vol. 15, no. 8, pp. 2331–2347, Nov. 2007.

[23] Z. Zeng, P. Maja, G. I. Roisman, and S. Thomas, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[24] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychol.: Development., Learn., Personal., Soc.*, vol. 14, pp. 261–292, 1996.

[25] W. Xiong, D. Cui, F. Meng, and L. Cai, "Analysis and conversion of emotional speech based on the prosodic features," in *Proc. 8th Phon. Conf. China (PCC)*, 2008.

[26] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–467, Dec. 1990.

[27] X. Zhang, J. Xu, and L. Cai, "Prosodic boundary prediction based on maximum entropy model with error-driven modification," in *Proc. ISCSLP*, Singapore, 2006.

[28] A. Kain and M. W. Macon, "Spectral voice conversions of text-to-speech synthesis," in *Proc. Int. Conf. Acoust, Speech, Signal Process.*, 1998, pp. 285–288.

[29] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997.

[30] Y. Freund and R. E. Schapire, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, pp. 771–780, 1999.

[31] *International Standard, Information Technology Coding of Audio-Visual Objects. Part 2: Visual; Amendment 1: Visual Extensions., International Organization for Standardization Std*, ISO/IEC 14496-2: 1999/Amd. 1: 2000(E).

[32] I. S. Pand'zić and R. Forchheimer, *MPEG-4 Facial Animation – The Standard, Implementations and Applications*. New York: Wiley, 2002.

[33] S. Zhang, Z. Wu, and L. Cai, "Region-based facial expression synthesis on a three-dimensional avatar," in *Proc. China Conf. Human Comput. Interact.*, 2006.

[34] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. 3rd IEEE Conf. Face Gesture Recogn.*, 1998, pp. 200–205.

[35] X. Li, H. Zhou, S. Song, T. Ran, and X. Fu, "The reliability and validity of the chinese version of abbreviated pad emotion scales," in *Proc. Int. Conf. Affective Comput. Intell. Interact.*, 2005.

[36] F. Lavagetto and R. Pockaj, "An efficient use of MPEG-4 FAP interpolation for facial animation at 70 bits/frame," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 1085–1097, Nov. 2001.

[37] Z. Wang, L. Cai, and H. Ai, "A dynamic viseme model for personalizing a talking head," in *Proc. 6th Int. Conf. Signal Process.*, Beijing, China, Aug. 2002.

[38] Z. Wu, L. Cai, and H. M. Meng, "Multi-level fusion of audio and visual features for speaker identification," in *Proc. Int. Conf. Biometrics (ICB2006)*, 2006, pp. 493–499.

[39] Z. Wu, L. Cai, and H. M. Meng, "DBN based audio-visual correlative model for audio-visual speech synthesis," in *Proc. NCMMSC2005* (in Chinese), Beijing, China, 2005, pp. 334–337.

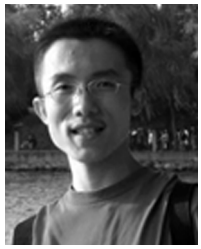
[40] "MATLAB Function Reference," 1-D Data Interpolation MathWorks.

[41] "Xface: MPEG-4 based open source toolkit for 3d facial animation," in *Proc. AVI04, Working Conf. Adv. Visual Interfaces*, Gallipoli, Italy, May 25–28, 2004.

Jia Jia (M'10) received the Ph.D. degree from Tsinghua University, Beijing, China, in 2008.

She is currently an Assistant Professor with the Department of Computer Science and Technology, Tsinghua University. Her current research interests include affective computing and computational speech perception.

Dr. Jia a member of the IEEE Signal Processing Society and the Multimedia Committee of Chinese Graphics and Image Society. She has been awarded Scientific Progress Prizes from the Ministry of Education, China.



Shen Zhang received the B.E. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2005. He is currently pursuing the Ph.D. degree at Tsinghua University.

His main research interests include expressive talking avatar synthesis.

Fanbo Meng received the B.E. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2007. He is currently pursuing the Ph.D. degree at Tsinghua University.

His main research interests include emotional speech conversion and expressive speech synthesis.

Yongxin Wang received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in 2005. He is currently pursuing the Ph.D. degree at Tsinghua University.

His research interests include prosody modeling, analysis, and conversion.

Lianhong Cai (M'09) received the B.E. degree from Tsinghua University, Beijing, China, in 1970.

She is currently a Professor with the Department of Computer Science and Technology, Tsinghua University. She directs the Human–Computer Speech Interaction Laboratory.

Prof. Cai has been awarded Scientific Progress Prizes and the Invention Prizes from the Ministry of Mechanism and Electronics and the Ministry of Education, China.