Multimed Tools Appl
DOI 10.1007/s11042-013-1604-8

Head and facial gestures synthesis using PAD model for an expressive talking avatar

Jia Jia · Zhiyong Wu · Shen Zhang · Helen M. Meng · Lianhong Cai

Abstract This paper proposes to synthesize expressive head and facial gestures on talking avatar using the three dimensional pleasure-displeasure, arousal-nonarousal and dominance-submissiveness (PAD) descriptors of semantic expressivity. The PAD model is adopted to bridge the gap between text semantics and visual motion features with three dimensions of pleasure-displeasure, arousal-nonarousal, and dominance-submissiveness. Based on the correlation analysis between PAD annotations and motion patterns derived from the head and facial motion database, we propose to build an explicit mapping from PAD descriptors to facial animation parameters with linear regression and neural networks for head motion and facial expression respectively. A PAD-driven talking avatar in text-to-visual-speech system is implemented by generating expressive head motions at the prosodic word level based on the (P, A) descriptors of lexical appraisal, and facial expressions at the sentence level according to the PAD descriptors of emotional information. A series of PAD reverse evaluation and comparative perceptual experiments shows that the head and facial gestures synthesized based on PAD model can significantly enhance the visual expressivity of talking avatar.

J. Jia · S. Zhang · L. Cai
Key Laboratory of Pervasive Computing, Ministry of Education China, and Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

J. Jia
e-mail: jjia@tsinghua.edu.cn

L. Cai
e-mail: clh-dcs@tsinghua.edu.cn

Z. Wu (✉) · H. M. Meng
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK), Hong Kong, China

Z. Wu
e-mail: zywu@se.cuhk.edu.hk

H. M. Meng
e-mail: hmmeng@se.cuhk.edu.hk

Z. Wu · H. M. Meng
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

Published online: 02 August 2013

Keywords Text-to-visual-speech · Head motion · Facial expression · Talking avatar

1 Introduction

Talking avatars, with human-like appearance, synthetic speech and lip articulation, are becoming a prominent form of multimodal human computer interactions [10, 37]. However, the ideal talking avatar requires not only natural speech articulation, but also expressive head motions, emotional facial expressions, and other meaningful facial gestures. Only with the expressive nonverbal gestures, can the talking avatar communicate intelligently and expressively with human.

The rigid head motion is an important aspect to enhance the visual expressivity of talking avatar. Early analysis has found that the timing of head motion accompanying speech is typically well synchronized with prosodic structure of the text [18]. An automatic text-driven head motion synthesizer was implemented by mapping the grammatical and prosodic textual information to elementary head motion patterns based on classification and regression trees (CART) [40]. To generate head motions that are synchronized to speech, studies were devoted to model the dynamic audio-visual mappings by learning the co-occurring acoustic and visual feature patterns based on hidden Markov models (HMMs) [6, 20, 29, 34]. Audio-visual database have also been collected to support data-driven approach that selects the video segments best-matched to the pitch contour of speech and stitch them together to create synthetic head motions [9]. Head motions are further analyzed and synthesized to express emotions of several categories [5].

For facial expression, realistic and expressive animation was achieved by concatenating and modifying the best-matched captured motion data with geometry and emotion constraints [11]. A human appearance model, which provides the parameterization of human facial movements, was built and applied in retargeting the facial animation from captured videos to synthetic face [36]. Speech-driven facial expression were synthesized by learning the mapping relationship between acoustic features and facial motion patterns based on different statistical approaches, including support vector machine (SVM) [8], neural networks (NN) [22], HMMs [42]. Under the assumption that there exist a small set of basic emotions [16], many studies adopt the categorical definition of "big six" emotions (i.e., joy, sadness, surprise, fear, anger and disgust) as their scope of emotional facial expression synthesis [1, 12, 37, 45]. Further explorations were undertaken to investigate the inter-emotional difference in facial expressions belong to different emotion categories [7].

Despite the success of previous study, several issues need to be addressed for expressive talking avatar synthesis. First, the current emotion expression of talking avatar was limited in several discrete categories, and thus incapable to express the complex and varied emotions during speech. The psychological study has long ago unveiled the multi-dimensional characteristic of human emotions [33], and the dimensional description of emotions has been adopted in many recent studies of expressive speech synthesis [35, 38]. However, the use of emotion dimensions rather than emotion categories, to annotate, parameterize and directly modify the head and facial gestures for synthetic talking avatars, is still rare. Second, previous studies put much effort on speech-driven talking avatars, by investigating and reproducing the co-occurring acoustic and visual feature patterns, while ignored the use of textual information in synthesizing the head and facial gestures for talking avatars in text-to-visual-speech (TTVS) system. Although some recent study has utilize the prosodic and grammatical text information in head motion synthesis [40], it is still a big challenge for talking avatar to convey the communicative semantics (e.g. emotion, appraisal and intention)

[30] in spoken text through visual gestures. To achieve an expressive and intelligent talking avatar capable of communicating with human, the semantic information should be considered as an important aspect to synthesize its head and facial gestures.

To address the above issues, this paper seeks to incorporate the semantic-oriented expressivity derived from text in synthetic head and facial gestures on talking avatar. To bridge the semantic gap between textual information and visual motion features, we adopt Mehrabian's three-dimensional PAD model [31]. The three dimensions include pleasure-displeasure (P), arousal-nonarousal (A) and dominance-submissiveness (D). The PAD model has been applied in affective sate modeling [17], expressive speech synthesis [35], facial expression simulation [3] and expressive head and body movement generation [26]. In our previous work on expressive text-to-speech synthesis [38], a set of PAD heuristics was proposed for text semantics annotation, which will be reused in this paper.

The main contribution of our work is to build an explicit mapping between the high-level semantic expressivity (in terms of PAD descriptors) and low-level visual motion features (i.e. the head and facial animation parameters). Different with previous studies that explore the use of PAD model in discrete regions with only two dimensions [3, 26], we aims to build a general framework that is capable in covering all the PAD dimensions, explore the correlation between PAD dimensions and motion features in a quantitative way, and further apply the mapping of PAD to motion features in synthesizing the head and facial gestures for an expressive talking avatar.

The overview of our approach for PAD-driven head and facial gesture synthesis is presented in Fig. 1. First, the PAD dimensions are used to annotate the expressivity of head motion and facial expression database. Then the semantic-oriented visual expressivity is modeled at two levels: 1) the short-term expressivity of head motion is synthesized at the prosodic word level based on lexical appraisal (e.g. commendatory and derogatory); 2) the long-term expressivity of facial expression is synthesized at the sentence level based on emotional information. For each level of modeling, we build the mapping between PAD descriptors and facial animation parameters that controls the facial movement on 3D talking avatar. Finally, an expressive text-driven talking avatar is implemented by combining the visual speech articulation (viseme) with synthetic head and facial gestures based on the PAD annotation of input text.

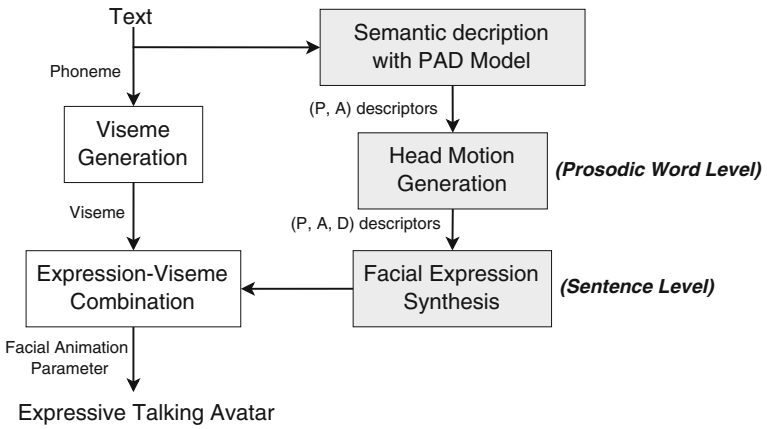


Fig. 1 Expressive talking avatar system

The rest of the paper is organized as follows. Section 2 presents the data collection and PAD annotation for head motion and facial expressions. The expressive head motion and facial expression analysis and synthesis are introduced in Sections 3 and 4 respectively. Section 5 illustrates the integration of head motion and facial expression on a text driven talking avatar, followed by a series of perceptual evaluations. Discussion on the current approach is presented in Section 6. We finally conclude this paper and give the future directions in Section 7.

2 Data collection and expressivity annotation

2.1 Head motion data

For head motion recording, a female speaker was seated in front of a teleprompter in a soundproof studio, keeping her eyes looking straight into the camera. The speaker can move her head freely and naturally while speaking. The video frame size is 720×576 pixels and the face region occupies about 300×400 pixels. Total recording time is about ten minutes with the sampling rate of 25 frames per second. Prior to recording, the speaker was asked to read through the text, which describes the attractive features of scenic spots cited from the Discover Hong Kong website of the Hong Kong Tourism Board [21]. During recording, the speaker was asked to speak expressively to convey the text semantics.

2.2 PAD annotation for head motion

Based on our previous study to map the text semantics to PAD descriptors [38], we adopt the (P, A) combination to annotate the spoken text at prosodic word level, where P value of words could be +1 (commendatory), -1 (derogatory) and 0 (neutral), and A value of words could be 1 (superlatives or with high degree), 0.5 (comparative or carrying key facts) and 0 (neutral degree). More details about the annotation heuristics can be found in [38].

The spoken text of head motion video is segmented into prosodic words by a home-grown text analyzer. There are 94 Chinese prosodic words and 460 syllables in total. Three annotators were invited to annotate the P and A values for prosodic words according to the above heuristics. If there is divergence among the three annotators, they discuss and re-annotate until at least two of the three annotators are in agreement. Finally, an agreement for 94 % of the prosodic words is achieved among the three annotators. Statistics of the annotation are shown in Table 1, and an example sentence with annotated (P, A) values is given in Table 2.

2.3 Facial expression image database

The Cohn-Kanade database [25] is adopted for facial expression analysis, which contains image sequences of 97 college students (65 % are female) performing 23 different facial displays. We choose this database because it focuses on collecting different facial

Table 1 Statistics of (P, A) annotation for prosodic words (PW)

(P, A)	(0,0)	(0,0.5)	(1,0)	(1,0.5)	(1,1)
#PW(total 94)	8	50	7	17	12
% of occurrence	8.5	53.2	7.4	18.1	12.8

Table 2 Example of (*P*, *A*) annotations for prosodic words in a sentence

Prosodic word	(P, A)	Prosodic word	(P, A)
太平山頂 (*Victoria Peak*)	(0,0.5)	山下 (*submontane*)	(0,0)
是香港 (*is Hong Kong's*)	(0,0.5)	鱗次櫛比 (*row upon row of*)	(1,1)
最受歡迎的 (*most popular*)	(1,1)	摩天高樓 *(skyscrapers)*	(0, 0.5)
名勝景點 (*scenic spot*)	(1,1)	和享譽全球 (*world famous*)	(1, 1)
登臨其間 (*climb up*)	(1,0)	維多利亞港 (*Victoria Harbor*)	(0, 0.5)
可俯瞰 (*can overlook*)	(0,0.5)	景色 (*scene*)	(0, 0)

movements rather than facial expression belongs to specific emotion categories. The PAD model is thus applied to describe the visual expressivity of facial movements. These facial movements include single and combinations of action units (AU) defined in Facial Action Coding System (FACS) [13]. Each image sequence corresponds to facial movement from neutral state to extreme state, as shown in Fig. 2. Totally 486 facial expression sequences are used in our study.

2.4 PAD annotation for facial expression

Perceptual evaluations are conducted to annotate the PAD values for facial expression images using a five-point Likert scale, scoring from −1 (*P*-negative, *A*-calm, *D*-weak) to +1 (*P*-positive, *A*-excited, *D*-strong). Five evaluators (two female and three male) were invited. To avoid personal bias on annotation, an incremental annotation strategy is applied. We randomly divide the whole dataset into five sets with about 97 samples in each. After the first set is annotated, we calculate the standard deviation of all evaluator's rating for each data sample, and the top 20 % samples with highest annotation deviation were added to the next set for re-annotation. After six-round annotation, we obtain all the evaluator's ratings for each data sample. For those samples that were annotated more than one round, we take the mean ratings of different rounds.

Finally, the inter-evaluator agreement is measured by both the standard deviation of all evaluators' ratings (*P*-std = 0.20, *A*-std = 0.32, *D*-std = 0.30) and the mean of pair-wise *Pearson correlations* between each two evaluators' ratings (*P*-corr = 0.80, *A*-corr = 0.44, *D*-corr = 0.35, $p<0.05$). Figure 3 compares the PAD annotation with the "big six" emotion labels (i.e., joy, sadness, surprise, fear, anger and disgust). The emotion labels are obtained based on the *AU to Emotion* translation rules defined in FACS manual [14].

From the distribution of the six emotion categories in PAD space, we can see that: 1) variation exists in each emotion, indicating that single emotion can result in multiple facial expressions (e.g., surprise pleasantly $P>0$ or scarily $P<0$); 2) overlap exists between emotion

Fig. 2 Facial expression image sequence in Cohn-Kanade database

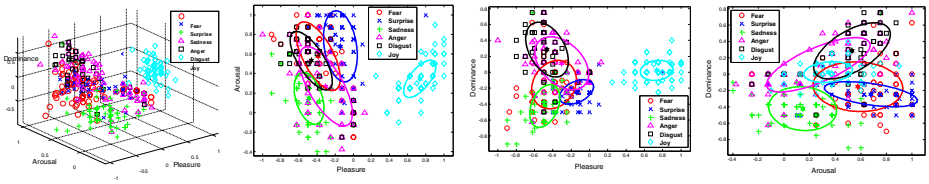


Fig. 3 PAD annotations for facial expressions of six emotion categories (with 50 % error ellipse marked for each category)

categories, indicating that different emotion categories may share similar facial expression patterns (e.g., anger and disgust); 3) blank space exists in PAD space, indicating that the basic "big six" emotion categories cannot cover all the possible facial expressions. Comparing with the emotion labels which classify facial expression into discrete categories, we believe that the PAD model is more effective and accurate in describing the facial expressions from the point of view of emotion synthesizer, and thus help us to enhance the visual expressivity of talking avatar.

3 Head motion analysis and synthesis

3.1 Head motion features extraction

Previous studies measure head motions in Euler angles of rotation around x- (*pitch*), y- (*yaw*), and z-axis (*roll*) respectively [5, 18, 34]. The three rotation angles are also defined in the MPEG-4 facial animation parameters (FAP) [32]. To extract the head rotation angles from video frames, we adopt an algorithm that estimates 3D head orientations from five facial features (four are eye-corners and the fifth is nose tip) in monocular image [23]. A face alignment toolkit is utilized to locate the above five facial features in each video frames [41]. Figure 4 presents the estimated 3D head motion from video frame sequence. The frame, in which the speaker keeps neutral state (the top-left in Fig. 4), is selected as reference frame. The 3D head motions are extracted by calculating the angle displacements of current frame to the reference frame.

To analyze head motion, previous speech-driven approaches investigate the acoustic correlates with motion features at the frame level [5, 34]. However, longer range dependencies rather than frame-wise correlation, were found between acoustic and motion features in other study based on different dataset [20]. To synthesize head motion at frame-level may achieve more dynamic motion details, but the frame-wise correlation is only suitable to build audio-visual mappings and is much dependent on the dataset. For text-driven head motion synthesis, a recent study synthesizes head motion at the syllable level based on textual information [40]. In this work, since we aim to incorporate the semantic expressivity derived from text in visual gestures, the prosodic word is chosen as the basic unit for head motion analysis and synthesis.

Fig. 4 Estimated head motion angles [α, β, γ] (in degree) and corresponding video frames. The five facial features for reference frame are marked as *triangles*, and for other frames are *circles*

The prosodic word carries both prosodic structure and semantic meaning in Chinese spoken language. Early study has found that head movement are strongly correlated with the prosodic structure of the text [18]. In our data, we have observed the head moves periodically and is nearly synchronous with the segmentation of prosodic word. To validate this observation, we calculate the timing distance between neighbor extreme points in the motion trajectory, and take this distance as one motion period. To avoid the abrupt shake with small amplitude, we only select the local extreme points with head position at least one standard deviation away from the average position. There are totally 76 extreme points found and the average distance between neighbor extreme points are 27.3 video frame (i.e. 1.092 s). In our corresponding spoken text, one prosodic word contains 4.21 syllables in average, and the average syllable duration is 7.36 video frames. So the average duration of one prosodic word is 1.221 s, which is close to the average head motion period (i.e. 1.092 s).

Based on the above assumption, we investigate the head motion at the prosodic word level. Figure 5 illustrates the head *pitch* trajectory of the example sentence in Table 2. The trajectory is segmented by prosodic word boundaries (i.e. the vertical dashed lines), and the (*P*, *A*) values for each prosodic word are also marked below.

Three motion features are extracted within prosodic word:

1) *Peak/Valley point*: extreme head motion angle;
2) *Amplitude (Amp)*: distance between *peak* and *valley point*;
3) *Average position (Avg)*: mean value of head motion angle.

The *peak* and *valley* points of head *pitch* are illustrated as circles and cross marks in Fig. 5, and the *amplitude* and *average position* are also marked. Table 3 presents the statistics of motion angle at frame level, and *Amp* and *Avg* at prosodic word level. The statistics suggest head *pitch* has higher motion activity than head *yaw* and *roll* at both frame and prosodic word level, confirming the results reported in [5], and also indicates head nods are significant motion patterns found in our corpus.

3.2 Head motion correlation with textual information

To investigate the semantic correlates with head motion, we illustrate the average *Amp* motion feature and corresponding (*P*, *A*) values in Fig. 6. The *Amp* motion feature is also extracted for

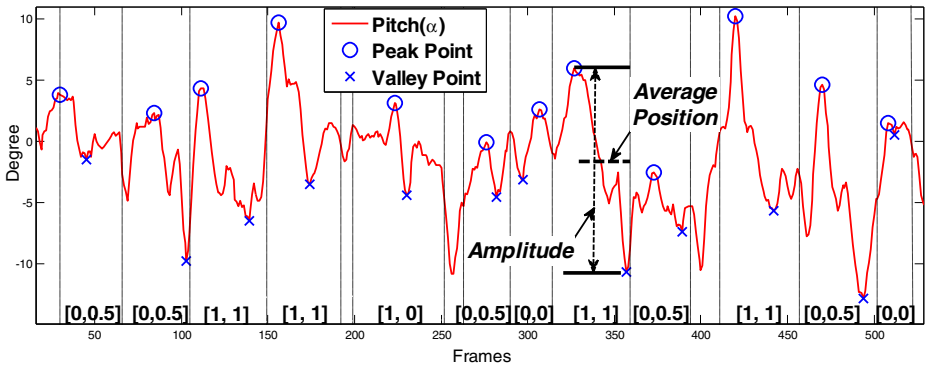


Fig. 5 Estimated head motion from video frames of example sentence in Table 2. The vertical dashed lines denote prosodic word boundaries, with [P, A] values marked below

Table 3 Statistical measurements of head rotation

Motion statistic	Pitch	Yaw	Roll
Angle Range [°]	32.09	14.61	20.03
Angle Std [°]	4.70	2.19	2.41
Amp Mean [°]	9.83	3.41	3.35
Amp Std [°]	4.34	1.72	2.57
Avg Mean [°]	−5.09	−0.47	0.57
Avg Std [°]	3.48	1.86	1.88

silence between prosodic words and annotated with (P=0, A=−0.5), since the silence takes neutral affective tendency (P=0), and lower activity than neutral speech (A=−0.5).

For head *pitch* (see Fig. 6(a)), it is observed that the average *Amplitude* increases as the A value increases. Similar trends are also observed for head *yaw* and *roll*. This agrees with common perception that speaker tends to move head with large amplitude when speak excitedly .

Since head nods are found to be powerful cues to prominence in previous study [19], we focus on the synchronization of head nods with particular syllables in prosodic word. The temporal relationship are investigated between the *peak points* (i.e. the lowest position when head pitch down) and particular syllables (i.e. stressed syllables and syllables with falling tone). The syllable tone is provided by the text analyzer of our Chinese text-to-speech engine [39]. For stressed syllables, the adjectives and adverbs with superlative meaning, such as "最(*most*)", "極(*super*)", "很(*very*)", are selected. The particular syllables are exclusively labeled, with the order of stress, falling tone and first syllable in prosodic word. Figure 7 illustrates the concurrence of *peak points* of head pitch and syllables with stress and falling tone in the example sentence in Table 2. The statistics of concurrence of *peak points* and particular syllables are shown in Table 4. This observation provides important timing cues to synchronize the head motion with syllables within prosodic word.

3.3 Head motion synthesis

Based on the correlation analysis between head motion features and textual information, we propose to model the head motion trajectory with text semantics and prosodic events under the following assumptions: 1) *Head moves nearly synchronous with the prosodic word;2)*

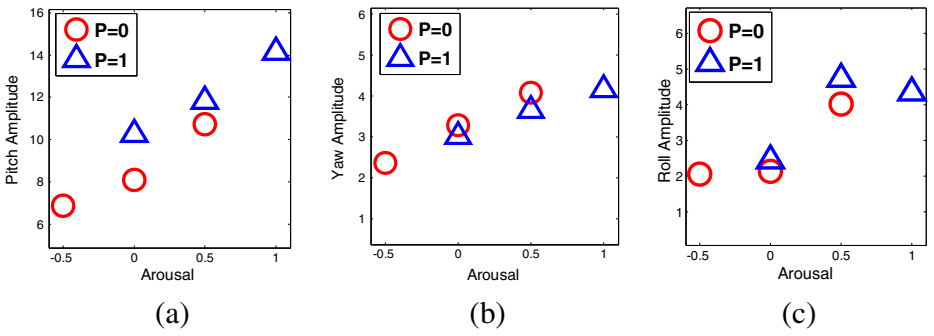


Fig. 6 Average *amplitude* of head pitch (**a**), yaw (**b**) and roll (**c**) for prosodic words with P and A values. The *circles* denote P=0 and triangles denote P=1

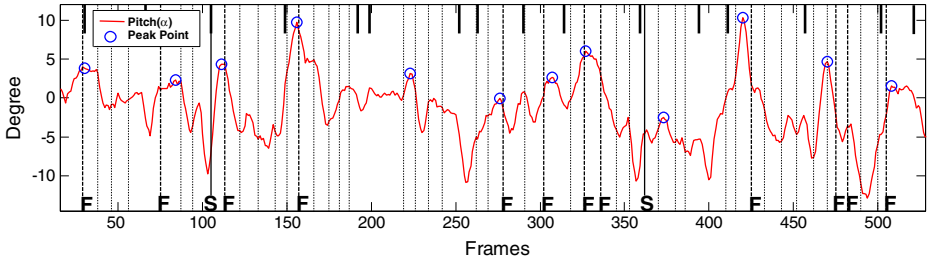


Fig. 7 Concurrence of *peak points* of head pitch and particular syllables. *Short vertical bars on top* denote prosodic word boundaries, and *vertical dashed lines* denote syllable boundaries. Stressed syllables are marked with *S* and syllables with falling tone are marked with *F*

Head usually shakes with a similar motion back. The first assumption has been validated in Section 3.1. The second assumption of head shake motion pattern during speech has also been observed in previous study [18, 20].

A sinusoidal function is proposed to generate head motion sequence within prosodic word based on the above assumptions. We choose the sinusoidal function because 1) it is simple and suitable to simulate the periodical and reciprocating movement; 2) it offers the flexibility to modulate head motion with the control of parameters. Moreover, with the sinusoidal function, we can easily modulate the high-level motion features (e.g., amplitude, peak points) through semantic annotations in terms of (P, A) values and particular syllables, rather than predict and interpolate the head positions for each frames.

$$y_t = \frac{1}{2} Amp \cdot \sin\left[\frac{2\pi}{T}\left(t - t_{peak}\right) + \frac{\pi}{2}\right] + Avg \quad , \quad t = 1, 2, .., T \tag{1}$$

The sinusoidal function is shown in (1), where y_t is the head rotation angles (*pitch*, *yaw* or *roll*) at t frame in prosodic word which consists of T frames. The peak point is placed on the frame of t_{peak}. The *Amp* is estimated as a quadratic function of the semantic annotations of (P, A) values, as shown in (2), where C_1 to C_6 are function coefficients estimated by linear least-square regression.

$$Amp = P\left(C_1 A^2 + C_2 A + C_3\right) + (1 - P)\left(C_4 A^2 + C_5 A + C_6\right) \tag{2}$$

The *Avg* is randomly drawn from a normal distribution with corresponding statistical mean and standard deviation shown in Table 3. The period of the sinusoidal function is set to be T to ensure there is one peak-valley pair within one prosodic word. For temporal synchronization, exclusive intra-prosodic-word rules are proposed to adjust t_{peak} for placing the peak point:

1) Place the peak point on the stressed syllable (if any);
2) Place the peak point on the syllable with falling tone (if any);
3) Place the peak point on the first syllable in prosodic word.

Table 4 Concurrence of peak points and syllables of: stress (SYL_S), falling tone (SYL_F), first syllable in prosodic word (SYL_P)

Syllable type	SYL_S	SYL_F	SYL_P	Others
# of Peak Points	15	26	22	31
% of occurrence	16.0	27.7	23.4	33.0

Since the head motions in three Euler angles are nearly independent with each other (the correlation coefficient between *pitch*, *yaw* and *roll* is lower than 0.3), we generate head motion around three axis respectively based on the above sinusoidal model. The synthetic head motion is further integrated in a text-driven talking avatar system, with details in Section 5.

4 Facial expression analysis and synthesis

Facial expression is another important aspect to enhance the visual expressivity of talking avatar. The facial expression conveys rich nonverbal information, such as emotion, intention and attitude. In this section, we extend our previous work [43] to explore the use of facial expression to convey communicative semantics in terms of PAD descriptors.

4.1 Partial expression parameters

For parameterized facial expression synthesis, the MPEG-4 facial animation parameter (FAP) has been adopted for animation purpose [42]. However, the FAP focuses on the animation of single facial point (e.g. mouth corner raise), rather than motion patterns of facial features (e.g. smiling mouth). It is complicated and not intuitive to define every FAP parameter for synthetic facial expression. To describe expressive facial motion patterns, we propose a set of partial expression parameters (PEP) based on FAP, as shown in Table 5.

The value of each PEP dimension ranges from −1 to +1, corresponding to continuous facial movement from one extreme state to the opposite state, such as from mouth bent-down to mouth bent-up in Fig. 8, and the zero value represents neutral state. Each PEP dimension is associated with a subset of FAPs (see Table 5) to controls the animation of facial points. The PEP-FAP mapping is thus learned to reduce the complexity of facial expression synthesis by FAP in the following section.

4.2 Learning mapping between PEP and FAP

To learn the PEP-FAP mapping, a FAP dataset is created by extracting FAPs from facial images in Cohn-Kanade database [25]. The FAPs are calculated based on the manual annotation of 59 facial points provided by LAIV laboratory [28]. These facial points (see Fig. 9(a)–(c)) are a subset of MPEG-4 facial definition points (FDP, see Fig. 9(d)), which are

Table 5 Definition of partial expression parameter

PEP#	Partial expression parameter	FAP subset (key-FAP is in parentheses)
1	Raise eyebrow	31, 32, 33, (34), 35, 36
2	Squeeze eyebrow	(37), 38
3	Open eye	19, (20), 21, 22
4	Look left/right	(23), 24
5	Look up/down	(25), 26
6	Mouth open (upper lip)	4, 8, (9), 51, 55, 56
7	Mouth open (bottom lip)	3, 5, 10, 11, 52, 57,(58)
8	Mouth bent	(12), 13, 59, 60
9	Mouth stretch	(6), 7, 53, 54

Fig. 8 Partial expression for mouth-bent up/down

the geometric reference for FAP measurement. We extract the 37 FAPs in Table 5 by measuring the facial point displacement between the expressive image (Fig. 9(b) and (c)) and the neutral image (Fig. 9(a)). The FAPs related with eyeball movement (FAP23~FAP26) are extracted approximately by measuring the horizontal and vertical displacement of pupils.

In previous work we propose to define the key-FAP for each PEP, where the key-FAP has the highest correlation with respect to other FAP and thus is used to present the motion patterns described by PEP. The key-FAP for each PEP dimension is also shown in Table 5. The PEP value is obtained by linear scaled the key-FAP into [−1, +1], and the other non-key FAP is linear interpolated by the key-FAP as shown in (3). The interpolation coefficient α_i is estimated by minimizing mean square error as shown in (4), where N is the number of facial expression samples, $FAP^n{}_i$ is the value of FAP_i extracted from the nth sample. More details can be found in previous work [43].

$$\mathrm{FAP}_i \approx \mathrm{FAP}_i^{'} = \alpha_i \bullet \ \mathrm{FAP}_{\mathrm{key}}^{'} \quad (i \neq \mathrm{key}) \tag{3}$$

$$\alpha_i = \left(\sum_{n=1}^{N} FAP_i^n \cdot FAP_{key}^n\right) \Big/ \sum_{n=1}^{N} \left(FAP_{key}^n\right)^2 , \quad (i \neq key) \tag{4}$$

4.3 Learning PAD-PEP mapping model

Before building the mapping from PAD to PEP, the *Canonical Correlation Analysis* (CCA) is conducted to validate the correlation between the PAD annotations (see Section 2.4) and PEP values (see Section 4.2). The first-order canonical correlation between PAD and PEP is 0.81, indicating that facial expressions are strongly linked with semantic expressivity. Following this, we propose both a multivariate linear estimator and a neural network based non-linear estimator to predict the PEP values from the PAD descriptors.

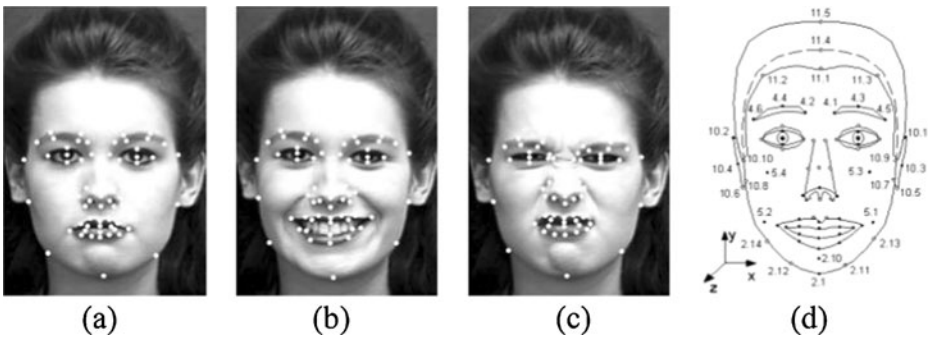


Fig. 9 Facial points annotation and MPEG-4 FDP definition

Linear predictor The *affine minimum mean square error estimator* (AMMSE) is adopted to build the linear estimator that predicate the PEP values from PAD descriptors by a transformation matrix $\boldsymbol{T}$ and a translation vector $\boldsymbol{V}$ as defined in (5)–(7).

$$\widehat{\mathbf{E}}_{\mathrm{PEP}} = \mathbf{T}\cdot\mathbf{E}_{\mathrm{PAD}} + \quad \mathbf{V} \tag{5}$$

$$\mathbf{T} = \mathbf{K}_{\mathrm{FE}}\cdot\mathbf{K}_{\mathrm{E}}^{-1} \tag{6}$$

$$\mathbf{V} = -\mathbf{K}_{\mathrm{FE}}\cdot\mathbf{K}_{\mathrm{E}}^{-1}\cdot\mathbf{U}_{\mathrm{E}} + \mathbf{U}_{\mathrm{F}} \tag{7}$$

where $\boldsymbol{E}_{\mathrm{PAD}}$ is the input PAD vector $[P, A, D]^{\mathrm{t}}$, $\hat{\boldsymbol{E}}_{\mathrm{PEP}}$ is the estimated PEP vector $[P_1, P_2,..P_9]^{\mathrm{t}}$. $\boldsymbol{K}_{\mathrm{FE}}$ is the cross covariance matrix between PEP parameters and PAD descriptors, $\boldsymbol{K}_{\mathrm{E}}$ and $\boldsymbol{U}_{\mathrm{E}}$ are the covariance and mean vector of PAD descriptors, and $\boldsymbol{U}_{\mathrm{F}}$ is the mean of PEP parameters. This mapping model is an optimum linear estimator that minimizes the mean square error. The advantage of linear prediction lies in its simplicity in model design and training. The AMMSE has also been used to build audio-visual mappings between acoustic features and facial gestures [7]. We take this linear predictor as a benchmark to be compared with the nonlinear predictor.

Nonlinear predictor We adopt *artificial neural networks* (ANN) to build the nonlinear mapping model from PAD to PEP. Since each dimension of PEP is loosely correlated with each other (average correlation coefficient is 0.31, $p<0.05$), we build nine neural networks to estimate each PEP dimensions (PEP_j) respectively. The feed-forward network composed of one nonlinear hidden layer and one linear output layer is designed and illustrated in Fig. 10. Three input units are designed for P, A and D respectively, and the hidden layer consists of H neurons with tanh(x) as the transfer function. The neural network estimates each PEP dimension (PEP_j) as a function of PAD descriptors ($\boldsymbol{E}_{\mathrm{PAD}}$) shown in (8)–(11). The hidden layer has the weight matrix $\boldsymbol{W}^1{}_j$, and the bias vector $\boldsymbol{b}^1{}_j$. The output layer is defined by the weight vector $\boldsymbol{w}^2{}_j$, and the bias factor $b^2{}_j$.

$$\mathrm{PEP}_j = \mathbf{w}_j^2\cdot\tanh\left(\mathbf{W}_j^1\cdot\mathbf{E}_{\mathrm{PAD}} + \mathbf{b}_j^1\right) + b_j^2 \quad , \quad j = 1, 2, ..., 9 \tag{8}$$

$$\mathbf{W}_j^1 = \begin{bmatrix} w_{1P_j}^1 & w_{1A_j}^1 & w_{1D_j}^1 \\ \vdots & \vdots & \vdots \\ w_{HP_j}^1 & w_{HA_j}^1 & w_{HD_j}^1 \end{bmatrix} \tag{9}$$

$$\mathbf{b}_j^1 = \left[b_{1j}^1, b_{2j}^1, ..., b_{Hj}^1\right]^t \tag{10}$$

$$\mathbf{w}_j^2 = \left[w_{1j}^2, w_{2j}^2, ..., w_{Hj}^2\right] \tag{11}$$

The Levenberg-Marquardt optimization algorithm is carried out to train the networks by minimizing the mean square error between the estimated and real values. The number of

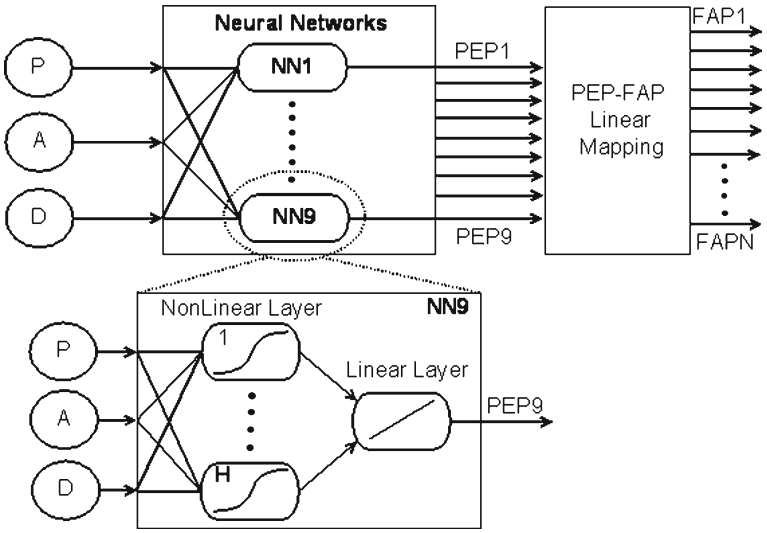


Fig. 10 Neural networks for modeling the nonlinear correlation between PAD descriptor and PEP parameters

hidden units (*H*) is experimentally determined. Once the predicted PEP values are obtained according to the PAD descriptors, either by linear predictor or neural networks, corresponding FAPs can be predicted by PEP-FAP linear translation and is used to animate 3-D face model to display synthetic facial expressions.

4.4 Experiments and evaluation

Experiments are carried out to train and evaluate the performance of PAD-PEP and PEP-FAP mapping models. The 486 facial expression samples in Cohn-Kanade database are randomly partitioned into training set (80 % samples) and test set (20 % samples). The normalized mean square error (NMSE) and the *Pearson's correlation coefficient* (ρ) between the predicted value ($\hat{y}$) and real value (y) is used to measure the prediction accuracy, see (12):

$$\mathrm{NMSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{\sigma_y}\right)^2, \quad \rho = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(y_i - \mu_y\right)\left(\hat{y}_i - \mu_{\hat{y}}\right)}{\sigma_y \bullet \sigma_{\hat{y}}} \tag{12}$$

where σ_y and $\sigma_{\hat{y}}$ are the standard deviations of y and $\hat{y}$ respectively, and μ_y and $\mu_{\hat{y}}$ are the mean values of y and $\hat{y}$ respectively. The K-fold cross validation (K=10) is adopted to train the PAD-PEP linear model (AMMSE) and the model with intermediate performance (in terms of NMSE) is taken as the final linear predictor. For the neural networks (ANN), we have tried different numbers of hidden unit (H from 1 to 20) and the results of experiments reveal that the best performance is achieved when H=8.

Table 6 compares the performance of PAD-PEP mapping by linear and nonlinear predictors, and also show the performance of PEP-FAP mapping. From the result, we can see that the nonlinear model (ANN) provides better prediction accuracy than linear model (AMMSE) for PAD-PEP mapping. Furthermore, the low prediction errors in PEP- FAP mapping support our layered approach to synthesize facial expression by mapping PAD to PEP and then translate to FAP.

Table 6 Evaluation results of PAD-PEP and PEP-FAP estimation

PEP#	PAD-PEP (AMMSE)		PAD-PEP (ANN)		PEP-FAP
	NMSE	Corr. (ρ)	NMSE	Corr. (ρ)	NMSE
1	0.57	0.70	0.25	0.89	0.13
2	0.92	0.29	0.90	0.31	0.33
3	0.67	0.63	0.42	0.77	0.19
4	0.97	0.17	1.04	0.11	0.22
5	0.74	0.59	0.55	0.71	0.15
6	0.76	0.52	0.70	0.56	0.21
7	0.64	0.60	0.48	0.74	0.10
8	0.53	0.69	0.45	0.75	0.06
9	0.76	0.49	0.37	0.79	0.62

5 Synthesis and evaluation

In this section, we illustrate the synthetic result of PAD-driven head motion and facial expressions on an expressive talking avatar. By adopting the heuristics to parameterize the PAD descriptors according to the semantic content of spoken text in our previous work on Chinese text-to-speech synthesis [38], we incorporate the semantic expressivity in terms of PAD values in synthetic head motions and facial expressions for an expressive Chinese talking avatar. A series of evaluation is further conducted to validate the effectiveness of our approach.

5.1 Head motion

The head motion is synthesized in three Euler angles respectively, based on the sinusoidal model proposed in Section 3. The synthetic head motion angles are translated to FAP to animate the talking avatar, namely the head pitch (FAP48), head yaw (FAP49) and head roll (FAP50). Figure 11 presents the snapshots of synthetic head motion on 3D talking avatar with the corresponding (P, A) values sourced from the example sentence in Table 2.

5.2 Facial expression

The facial expression is synthesized based on the ANN-based PAD-PEP-FAP mapping model. Figure 12 illustrates the synthetic facial expressions on talking avatar with the PAD values for 14 emotional adjectives. These PAD values are assessed by hundreds of Chinese undergraduates using a 12-item PAD scales [27, 31]. A PAD reverse evaluation experiment is conducted to validate the effectiveness of our approach in facial expression synthesis. 15 research students are invited to rating their perception of the synthetic facial expressions using the 12-item PAD

Fig. 11 Synthetic head motions with (P, A) values marked below

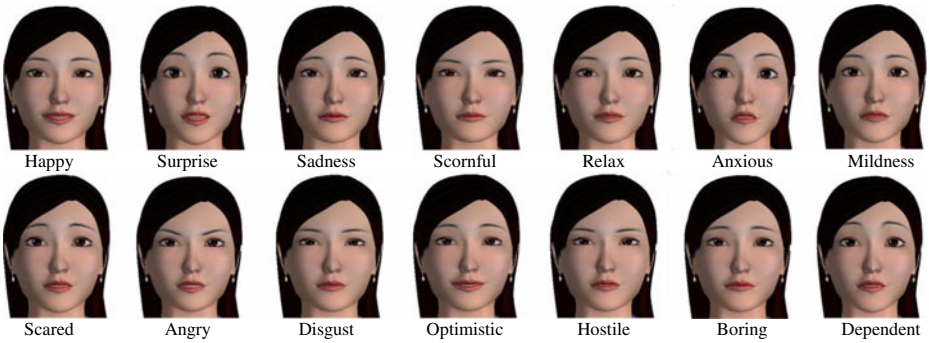


Fig. 12 Synthetic facial expression based on PAD values of emotional words

scales shown in Table 7. With the 12-item scales, each facial expression is rated with 12 bipolar adjectives on a −4 to 4 Likert scale, and the *P*, *A* and *D* values will be calculated from each group of 4 bipolar adjectives as shown in (13) [27, 31]. Previous psychological research has revealed this method can avoid subjective cognitive differences rather than directly annotating the PAD values. More details can be found in [27].

$$\begin{aligned} P &= (Q1{-}Q4 + Q7{-}Q10)/16 \\ A &= (-Q2 + Q5{-}Q8 + Q11)/16 \\ D &= (Q3{-}Q6 + Q9{-}Q12)/16 \end{aligned} \tag{13}$$

Table 8 compares the input PAD descriptors of the emotional adjectives and the evaluated PAD values of the synthetic facial expressions. The Pearson correlation coefficient between the input and evaluated PAD values is 0.873 (P), 0.569 (A) and 0.624(D) respectively. This result shows that the evaluated PAD is nearly consistent with the input PAD, indicating that the synthetic facial expression based on our approach can transfer the emotional information effectively.

5.3 Talking avatar integration

For expressive talking avatar, the lip movement is controlled by both speech gestures (viseme) and facial expressions (e.g. smile). Under the assumption that the speech and facial expressions contribute weighted equally to the lip movement, we adopt the co-articulation strategy in [37], by linear combining the viseme and facial expressions as follows:

$$\mathrm{FAP}_f = \delta \cdot \mathrm{FAP}_v + (1{-}\delta) \cdot \mathrm{FAP}_e \tag{14}$$

Table 7 12-Item PAD scales for facial expression evaluation

#	Adjective	−4	−3	−2	−1	0	1	2	3	4	Adjective
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	

Table 8 Comparison of input PAD and evaluated PAD for synthetic facial expressions

Word	Input PAD			Evaluated PAD		
	P	*A*	*D*	*P*	*A*	*D*
Happy	0.55	0.24	0.28	0.42	0.12	0.1
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
Scornful	−0.32	0.06	0.2	−0.25	0.18	0.08
Relax	0.44	−0.13	0.21	0.05	−0.26	−0.14
Anxious	−0.19	0.06	−0.13	−0.12	0.14	−0.22
Mildness	0.31	−0.16	0.08	0.14	−0.05	0.04
Scared	−0.19	0.26	−0.13	0.01	−0.04	−0.25
Angry	−0.4	0.22	0.12	−0.56	0.51	0.42
Disgust	−0.36	0.08	0.13	−0.56	0.15	0.44
Optimistic	0.5	0.21	0.35	0.24	0.16	0.05
Hostile	−0.42	0.2	0.22	−0.32	0.16	0.14
Boring	−0.11	−0.25	−0.17	−0.14	−0.02	−0.34
Dependent	0.08	−0.16	−0.3	0.13	0.13	−0.16

The weight coefficient δ is manually defined. To avoid jerky mouth movement (such as a wide-opened mouth *of happy* may conflict with a closed syllable in speech), the combination is conducted only when the viseme and facial expression has the same direction. If FAP_v and FAP_e move in the opposite direction, the mouth movement (FAP_f) is absolutely controlled by viseme (FAP_v) to guaranty the speech articulation is correct.

5.4 Perceptual evaluation

In this section, three experiments are conducted to evaluate the visual expressivity of talking avatar with (I) only head motion; (II) only facial expression; (III) integrated head and facial gestures respectively.

We select 20 sentences from the descriptive genre of Hong Kong tourist information corpus [38]. The prosodic words are annotated with (*P*, *A*) values as described in Section 2.2, and the *D* value denotes the communicative goal of the sentence (i.e. *D*=1 for confirmation or imperatives, *D*=0.5 for explanation or declaratives, *D*=−0.5 for suggestion, thanks or request, *D*=−1 for apology, imploring or interrogative, *D*=0 for neutral) [38]. The prosodic word with the largest absolute (*P*, *A*) value are considered as semantic prominence in the sentence, and is combined with the *D* value to form the target (*P*, *A*, *D*) descriptors.

The head motion sequence is generated for each prosodic word, with smoothing techniques applied to obtain natural transition from current prosodic word to the next. A five-frame (0.2 s) moving average window is applied at each prosodic word boundary or boundary between prosodic word and silence. The facial expression is synthesized according to the target (*P*, *A*, *D*) descriptors of the sentence, and is then linearly combined with the speech gestures (viseme) if needed.

To evaluate the naturalness and expressiveness of synthetic avatar, a five-point Likert scale is proposed as follows:

5 Expressive: natural and expressive gestures like human;
4 Natural: appropriate gestures to message content;

3 Acceptable: neutral gestures with some expressivity;
2 Unnatural: robotic gestures with little expressivity
1 Erratic: inappropriate and weird gestures

The mean opinion score (MOS) is calculated based on the rating scores from 15 research students, who have rich experiences in perceptual evaluation. In each experiment, the synthetic avatars are presented as short animation videos and shown to the subjects in random order. To minimize the learning effect that may affect the evaluation results, we conduct the following three experiments with one-month interval as presented in Table 9.

The evaluation result (MOS) in each experiment is shown in Table 10. It can be seen that the head motion and facial expression generated by our approach gains much higher MOS than neutral gestures, and the combination of head motion and facial expression will largely enhance the visual expressivity of talking avatar. The one-way analysis of variance (ANOVA) test is performed to validate evaluation result [2]. The ANOVA F-value is shown in Table 10, indicating a significant difference between each group of the synthetic results. This result demonstrates the feasibility and effectiveness of our approach to enhance the visual expressivity of talking avatar by incorporating the text semantic expressivity (in terms of PAD descriptors) in head and facial gestures of talking avatar.

6 Discussion

In this paper, the PAD model is introduced to bridge the gap between high-level expressivity and low-level motion features with parameter mapping model, and how to effectively obtain the PAD values becomes a practical problem. Since the PAD model is originally designed for human emotion evaluation, it is the most popular and recommend way to obtain PAD values through human ratings [4, 5, 31]. Previous engineering approaches have gain convincible PAD values from human annotation [3] and predefined rules [26, 38].

In our work, human annotations are utilized to obtain the PAD values for training data, and heuristics rules were adopted to derive the PAD values from the input text in specific scope. Despite the limited variability covered in PAD space, we propose to build a general mapping model between PAD descriptors and motion features, which can be easily extended to accommodate higher variability in the PAD continuum. Some preliminary study has shown that it is

Table 9 Experimental setup of perceptual evaluation

Groups	Exp1 (Head motion)	Exp2 (Facial expression)	Exp3 (Talking avatar)
I	Visual speech without head motion (i.e. static head)	Visual speech with neutral facial expression	Visual speech only with prosodic head motion
II	Visual speech with random head motion[a]	Visual speech with facial expression synthesized by linear predictor (AMMSE)	Visual speech only with synthetic facial expression[b]
III	Visual speech with prosodic head motion (i.e. synthesized by our approach)	Visual speech with facial expression synthesized by non-linear predictor(ANN)	Visual speech with both prosodic head motion and facial expression

([a] the random head motion is also synthesized by the sinusoidal function for each prosodic word, but with random generated amplitude and average position within the statistical range in Table 3, and random "*peak*" position within the prosodic word)

([b] the facial expression in Exp3 is synthesized by the ANN-based predictor)

Table 10 MOS result of perceptual evaluation

Experiment	Group I	Group II	Group III	$F[2, 57]$ ($p<0.01$)
Head Motion	2.23	2.57	3.87	197.3
Facial Expression	1.84	3.18	4.22	122.6
Talking Avatar	2.62	3.75	4.36	38.7

possible to predict the PAD values automatically from different media, such as the backward mapping facial displaying emotions to PAD values [3], and the PAD-PEP-FAP framework can also be reversed to predict the PAD values from facial animations.

Our approach is applied in an expressive talking avatar for Chinese text-to-visual-speech synthesis. However the approach itself is designed to be language-independent thanks to the PAD model. The PAD model essentially shades the modeling of head and facial gestures from the high-level text semantics, so that we can focus on mapping the PAD descriptors to visual motion features. Toward the PAD parameterization for input text, we adopt the heuristics that are proposed in the PAD based expressive text-to-speech synthesis [24, 38]. To extend our approach to talking avatar in other languages, similar PAD parameterizations need to be devised according to the specific language. For head motion, the basic unit for trajectory generation and prosodic features for temporal synchronization may need to be modified according to the dataset in specific language. The facial expression is considered to be universal recognized non-verbal gestures [15] and thus be language-independent. However, the PAD rating of human perception was previously reported to be culture related [27], so it is better to collect PAD annotation for training data from subjects with the same culture background as the synthetic talking avatar.

One limitation of this work is that we analyzed the head motion of a single subject with a small dataset. A larger dataset is under construction by recording the speech and gestures from different text genres and more subjects, which will help to validate our approach and to investigate the personal styles in head motion. For the facial expression, we choose a relative larger and widely-used database rather than building ourselves. This is because we want to cover more variations of expressivity in facial expression from more data and more subjects. Note that, this also leads to the inconsistency between the training data for head motion and facial expression. The current approach should not be affected by such inconsistency because the head motion and facial expression are treated as two independent gestures and their motion models are also trained respectively in this paper. However, a larger dataset with both head and facial gestures will be our future direction.

Another noteworthy point is that the parameter-based animation of talking avatar has enabled us to build the explicit mapping between PAD descriptors and animation parameters. Other data-driven approach [9, 11] by building large head/facial motion library may achieve more realistic animation with dynamic details, but at a cost of the flexibility and efficiency of precise control over different face model, which is a critical requirement for real-time application like text-to-speech synthesis. Moreover, the synthetic result based on data-driven approach is much dependent on the quality and style of the captured motion data. To improve the realism and animation details of synthetic avatar, we may consider combining the advantage of both parameter and performance based approach as proposed in a recent study [36].

7 Conclusion and future work

In this paper, we develop the techniques of expressive talking avatar synthesis by incorporating semantic expressivity in head motion and facial expressions. The PAD descriptors of text

semantics are adopted to depict the expressivity in head and facial gestures, and a layered framework is built to explicitly map the PAD descriptors to motion features. At the prosodic word level, a sinusoidal function is proposed to generate head motion sequences with the (P, A) values modulating the motion parameters. At the sentence level, a nonlinear mapping based on neural network is adopted to predict the facial expressions from PAD descriptors. The PAD reverse evaluation validates the effectiveness of our approach in transferring the expressivity from text to visual gestures. The perceptual evaluation of the PAD-driven expressive talking avatar indicates that the visual expressivity of synthetic avatars is significantly enhanced by incorporating the text semantics in head and facial gestures.

This study obtains preliminary but promising experimental results for expressive talking avatar synthesis. However, there are many opportunities to improve our approach. First, the simple linear combination of visemes and facial expressions does not consider their complicated and dynamic interactions under different scenarios (e.g., different visemes and different expressions during the real speech). Much more efforts are needed on this open problem. Second, the comparative evaluation between the proposed method and other existing methods is not conducted in this paper. For one thing, the coupling between non-verbal gestures (e.g. head motion, facial expression) and text/speech is not so strong as the phonemes and visemes. For the same input text/speech, there can be many different but appropriate head motions and facial expressions. Moreover, different studies usually collect their own dataset to train their specific models. It is hard to compare different methods directly without both a standard dataset and evaluation standards. However, with the help of some versatile perceptual evaluation tools like 12-item PAD scales [27] and self assessment manikin [4], it would be much easy to conduct the comparative evaluations.

Notwithstanding the above limitations, this study does suggest the feasibility of conveying semantic expressivity in head and facial gestures, and thus enhance the intelligibility of talking avatars.

Acknowledgment This work is supported by the National Basic Research Program of China (2011CB302201). This work is also partially supported by 973 program (2012CB316401), the research funds from the Hong Kong SAR Government's Research Grants Council (N-CUHK414/09, N-CUHK417/04), the National Natural Science Foundation of China (60805008).

The basic idea of this paper appeared in our conference versions [43, 44]. In this version, we extend our approach to be combined with both head and facial gesture, carry out detailed analysis, and present more performance results. The authors would like to thank Professor Haizhou Ai for providing the face alignment toolkit [41], and the Microsoft Research Asia-Tsinghua Univertity Joint Laboratory for its funding.

References

1. Albrecht I, Schröder M, Haber J, Seidel H-P (2005) Mixed feelings:expression of non-basic emotions in a muscle-based talking head. Virtual Real 8:201–212
2. ANOVA TEST. Available: http://en.wikipedia.org/wiki/Analysis_of_variance
3. Boukricha H, Wachsmuth I, Hofstatter A, Grammer K (2009) Pleasure-arousal-dominance driven facial expression simulation. In 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009). Amsterdam, Netherlands, IEEE Computer Society, pp. 21–26
4. Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. Behav Ther Exp Psychiatry 25:49–59
5. Busso C, Deng Z, Grimm M, Neumann U, Narayanan S (2007) Rigid head motion in expressive speech animation: analysis and synthesis. IEEE Trans Audio, Speech, Lang Process 15:1075–1086
6. Busso C, Deng Z, Neumann U, Narayanan S (2005) Natural head motion synthesis driven by acoustic prosodic features. Comput Animat Virtual Worlds 16:283–290
7. Busso C, Narayanan S (2007) Interrelation between speech and facial gestures in emotional utterances: a single subject study. IEEE Trans Audio, Speech, Lang Process 15:2331–2347

8. Cao Y, Tien WC, Faloutsos P, Pighin F (2005) Expressive speech-driven facial animation. ACM Trans Graph 24:1283–1302
9. Chuang E, Bregler C (2005) Mood swings: expressive speech animation. ACM Trans Graph 24:331–347
10. Cosatto E, Ostermann J, Graf HP, Schroeter J (2003) Lifelike talking faces for interactive services. Proc IEEE 91:1406–1429
11. Deng Z, Neumann U (2006) eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls. In Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation. Vienna, Austria, Eurographics Association, pp. 251–260
12. Deng Z, Neumann U, Lewis JP, Kim T-Y, Bulut M, Narayanan S (2006) Expressive facial animation synthesis by learning speech coarticulation and expression spaces. IEEE Trans Vis Comput Graph 12:1523–1534
13. Ekman P, Friesen W (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto
14. Ekman P, Friesen WV, Hager JC (2002) FACS investigator's guide. A human face, Salt Lake
15. Ekman P, Friesen WV, O'Sullivan M, Chan A, Diacoyanni-Tarlatzis I, Heider K, Krause R, LeCompte WA, Pitcairn T, Ricci-Bitti PE, Scherer K, Tomita M, Tzavaras A (1987) Universals and cultural differences in facial expressions of emotion. J Pers Soc Psychol 53:712–717
16. Ekman P, Oster H (1979) Facial expressions of emotion. Annu Rev Psychol 30:527–554
17. Gebhard P (2005) ALMA: a layered model of affect. In 4th international joint conference on Autonomous agents and multiagent systems (AAMAS '05). Association for Computing Machinery, New York, pp 29–36
18. Graf HP, Cosatto E, Strom V, Huang FJ (2002) Visual prosody: facial movements accompanying speech. In 5th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR 2002), pp. 381–386
19. Granström B, House D (2005) Audiovisual representation of prosody in expressive speech communication. Speech Commun 46:473–484
20. Hofer G, Shimodaira H (2007) Automatic head motion prediction from speech data. In Interspeech 2007. Antwerp, Belgium
21. Hong Kong Tourism Board (2006, 2006-04-20) Discover Hong Kong (2006 ed.). Available: http://www.discoverhongkong.com
22. Hong P, Wen Z, Huang TS (2002) Real-time speech-driven face animation with expressions using neural networks. IEEE Trans Neural Netw 13:916–927
23. Horprasert T, Yacoob Y, Davis LS (1996) Computing 3-D head orientation from a monocular image sequence. In International Conference on Automatic Face and Gesture Recognition (IEEE COMPUTER SOC), pp. 242–247
24. Jia J, Zhang S, Meng F, Wang Y, Cai L (2011) Emotional audio-visual speech synthesis based on PAD. IEEE Trans Audio, Speech, Lang Process 19(3):570–582
25. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In 4th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR2000), pp. 46–53
26. Lance B, Marsella S (2007) Emotionally expressive head and body movement during gaze shifts. In 7th Internationl Conference on Intelligent Virtual Agents (IVA'07) (Springer), pp. 72–85
27. Li X, Zhou H, Song S, Ran T, Fu X (2005) The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales. In: Affective Computing and Intelligent Interaction (ACII2005), pp. 513–518.
28. Lipori G (2005) Manual annotations of facial fiducial points on the Cohn Kanade database
29. Mana N, Pianesi F (2006) HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads. In: 8th International Conference on Multimodal Interfaces (ICMI06), Banff, AB, Canada, pp. 380–387
30. Mehrabian A (1972) Nonverbal communication, 2nd edn. Aldine-Atherton, Chicago
31. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14:261–292
32. Motion Pictures Expert, G (1999) ISO/IEC 14496–2. international standard, information technology-coding of audio-visual objects. Part 2: visual; amendment 1: visual extensions
33. Russell J, Mehrabian A (1977) Evidence for a three-factor theory of emotions. J Res Personal 11:273–294
34. Sargin ME, Yemez Y, Erzin E, Tekalp AM (2008) Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. TPAMI 30:1330–1345
35. Schröder M (2006) Expressing degree of activation in synthetic speech. IEEE Trans Audio, Speech, Lang Process 14:1128–1136

36. Stoiber N, Seguier R, Breton G (2010) Facial animation retargeting and control based on a human appearance space. Comput Animat Virtual Worlds 21:39–54
37. Tang H, Fu Y, Tu J, Hasegawa-Johnson M, Huang TS (2008) Humanoid audio-visual avatar with emotive text-to-speech synthesis. IEEE Trans Multimedia 10:969–981
38. Wu Z, Meng HM, Yang H, Cai L (2009) Modeling the expressivity of input text semantics for Chinese text-to-speech synthesis in a spoken dialog system. IEEE Trans Audio, Speech, Lang Process 17:1567–1576
39. Wu Z, Zhang S, Cai L, Meng HM (2006) Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar. In INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing (INTERSPEECH 2006-ICSLP). Pittsburgh, PA, United states, DUMMY PUBID, pp. 1802–1805
40. Xie L, Liu Z-Q (2007) Realistic mouth-synching for speech-driven talking face using articulatory modelling. IEEE Trans Multimedia 9(3):500–510
41. Zhang L, Ai H, Xin S, Huang C, Tsukiji S, Lao S (2005) robust face alignment based on local texture classifiers. In IEEE International Conference on Image Processing (ICIP 2005). Institute of Electrical and Electronics Engineers Computer Society, pp. 354–357
42. Zhang Y, Ji Q, Zhu Z, Yi B (2008) Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters. IEEE Trans Circ Syst Video Technol 18:1383–1396
43. Zhang S, Wu Z, Meng HM, Cai L (2007) Facial expression synthesis using PAD emotional parameters for a Chinese expressive avatar. In 2nd International conference on Affective Computing and Intelligent Interaction (Lisbon, Portugal), pp. 24–35
44. Zhang S, Wu Z, Meng HM, Cai L (2007) Head movement synthesis based on semantic and prosodic features for a Chinese expressive avatar. In International conference on acoustics, speech and signal processing, vol. IV. Hawai'i Convention Center, Honolulu, Hawaii, USA, pp. 837–840
45. Zhou C, Lin X (2005) Facial expressional image synthesis controlled by emotional parameters. Pattern Recog Lett 26:2611–2627

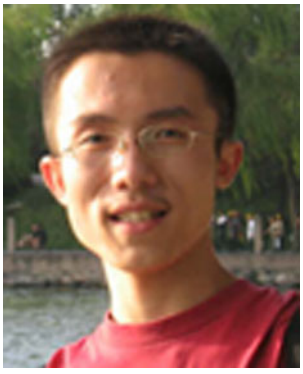
Jia Jia received the Ph.D. degree from Tsinghua University, Beijing, China, in 2008. She is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. Her current research interests include affective computing and computational speech perception.

Dr. Jia is 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.

Zhiyong Wu received the B.S. and Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 1999 and 2005, respectively.

He has been Postdoctoral Fellow in the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK) from 2005 to 2007. He joined the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, in 2007, where he is currently an Associate Professor. He is also with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. His research interests are in the areas of multimodal multimedia processing and communication, more specifically, audiovisual bimodal modeling, text-to-audio-visual-speech synthesis, and natural language understanding and generation.

Dr. Wu is a member of the Technical Committee of Intelligent Systems Application under the IEEE Computational Intelligence Society and the International Speech Communication Association.



Shen Zhang received the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. His research interests include affective computing and image processing.

Helen M. Meng received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology (MIT), Cambridge.

She has been Research Scientist with the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She joined

The Chinese University of Hong Kong (CUHK) in 1998, where she is currently a Professor in the Department of Systems Engineering and Engineering Management and Associate Dean of Research of the Faculty of Engineering. In 1999, she established the Human–Computer Communications Laboratory at CUHK and serves as Director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies, which was upgraded to MoE Key Laboratory in 2008, and serves as Co-Director. She is also Co-Director of the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. Her research interest is in the area of human–computer interaction via multimodal and multilingual spoken language systems, as well as trans-lingual speech retrieval technologies.

Lianhong Cai received the B.S. degree in computer science and technology from Tsinghua University, Beijing, China, in 1970.

She is currently a Professor with the Department of Computer Science and Technology, Tsinghua University. She was Director of the Institute of Human–Computer Interaction and Media Integration from 1999 to 2004. Her major research interests include human–computer speech interaction, speech synthesis, speech corpus development, and multimedia technology. She has undertaken 863 National High Technology Research and Development Program and National Natural Science Foundation of China projects.

Prof. Cai is a member of the Multimedia Committee of Chinese Graphics and Image Society and Chinese Acoustic Society.